Implicit Updates for Average-Reward Temporal Difference Learning

Hwanwoo Kim*
Duke University

Dongkyu Derek Cho*
Duke University

Eric Laber
Duke University

Abstract

Temporal difference (TD) learning is a cornerstone of reinforcement learning. In the average-reward setting, standard TD(λ) is highly sensitive to the choice of step-size and thus requires careful tuning to maintain numerical stability. We introduce average-reward implicit TD(λ), which employs an implicit fixed point update to provide data-adaptive stabilization while preserving the per iteration computational complexity of standard average-reward TD(λ). In contrast to prior finite-time analyses of average-reward TD(λ), which impose restrictive step-size conditions, we establish finite-time error bounds for the implicit variant under substantially weaker step-size requirements. Empirically, average-reward implicit TD(λ) operates reliably over a much broader range of step-sizes and exhibits markedly improved numerical stability. This enables more efficient policy evaluation and policy learning, highlighting its effectiveness as a robust alternative to average-reward TD(λ).

1 INTRODUCTION

Temporal difference (TD) learning (Barto, 2021) is a core component of modern reinforcement learning (RL), combining the strengths of Monte Carlo sampling and dynamic programming thereby enabling efficient value estimation from state-action-reward trajectories exhibiting Markovian dependence. As a foundational method, TD learning underlies many RL algorithms and has been successfully applied across diverse domains, including robotics (Kober et al., 2013), financial decision-making (Nevmyvaka et al., 2006), and games (Tesauro, 1995). While originally developed in the discounted-reward setting, TD learning has since been adapted to the average-reward setting (Tsitsiklis and Van Roy, 1999), which can be more natural in many applications (Giannoccaro and Pontrandolfo, 2002; Dai and Gluzman, 2022; Tangamchit et al., 2002).

Despite its widespread use and practical relevance, standard average-reward TD(λ) (Tsitsiklis and Van Roy, 1999) is sensitive to step-size selection. From a theoretical standpoint, stability is certified by finite-time error bounds, and existing analyses establish such bounds only in small step-size regimes (Zhang et al., 2021a). In practice, larger step-sizes can accelerate learning but at the risk of numerical instability; conversely, smaller step sizes are more numerically stable but can also yield slower learning. This stability–efficiency trade-off motivates methods that preserve the simplicity of the average-reward TD(λ) while substantially expanding the range of step-sizes for which learning remains stable. We address this sensitivity by proposing an average-reward implicit TD(λ) with finite-time error guarantees under substantially less restrictive step-size conditions. In addition, the proposed algorithm retains the computational complexity of standard average-reward TD(λ).

1.1 Related Literature

Discounted-Reward Setting. Almost sure convergence of TD(λ) with linear function approximation was first established in Tsitsiklis and Van Roy (1997). Subsequent work derived finite-time error bounds under both i.i.d. data streams (Dalal et al., 2018) and Markovian samples, using projection-based mean-path analysis (Bhandari et al., 2018), Lyapunov-function arguments (Srikant and Ying, 2019), and induction-based proofs (Mitra, 2024). In addition, TD-type methods formulated as two-time scale stochastic approximation algorithms—used for off-policy evaluation in the discounted setting—have been analyzed in Sutton et al. (2008, 2009); Xu et al. (2019).

Despite the aforementioned theoretical developments

*Joint first authors. Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.
Copyright 2026 by the author(s).

in the discounted-reward setting, classical TD methods typically require restrictive step-size conditions (Bhandari et al., 2018; Srikant and Ying, 2019; Mitra, 2024) and display marked empirical sensitivity: larger steps may accelerate progress but risk divergence, whereas smaller steps improve stability at the cost of substantially slower convergence (Dabney and Barto, 2012; Tamar et al., 2014). A principled remedy is to use implicit stochastic updates that recast the recursion as a fixed-point equation, providing data-adaptive stabilization, as shown in the stochastic optimization literature (Toulis et al., 2014; Toulis and Airoldi, 2014; Chee et al., 2023). Building on this principle in reinforcement learning, recent work establishes asymptotic and finite-time error bounds for implicit variants of discounted TD in both on- and off-policy tasks without restrictive step-size requirements (Kim et al., 2025). Experiments further show improved numerical stability in both policy evaluation and control tasks.

Average-Reward Setting. For foundations, background, and developments in average-reward policy evaluation, we refer to Puterman (2014); Singh (1994); Mahadevan (1996); Abounadi et al. (2001); Tsitsiklis and Van Roy (2002); Dewanto et al. (2020). The first convergence analysis specific to average-reward TD(λ) with linear feature approximation is due to Tsitsiklis and Van Roy (1999), under the assumption that the span of the feature vector does not include the constant vector of all ones (see Section 4 for additional discussion). Under the same assumption, Yu and Bertsekas (2009) established asymptotic convergence of the average-reward LSPE(λ), a least-squares based alternative to the average-reward TD(λ). Relaxing the aforementioned feature space restriction, Zhang et al. (2021b) derived finite-time bounds for average-reward TD(λ) with both constant and linearly decaying step-sizes (i.e., t^{th} step-size $\propto 1/t$), while imposing a restrictive condition on the initial step-size. More recent progress on average-reward off-policy evaluation with function approximation includes an asymptotically convergent tabular off-policy TD algorithm (Wan et al., 2021) and extensions of gradient TD methods (Sutton et al., 2008, 2009) for average-reward off-policy evaluation tasks with linear function approximation (Zhang et al., 2021b). Furthermore, TD-style methods for estimating a policy's asymptotic variance of the cumulative reward in average-reward setting are developed in Agrawal et al. (2024).

1.2 Contributions

We show that the step-size sensitivity of TD(λ), which has been well documented in the discounted setting also arises in the average-reward setting. To mitigate this sensitivity, we adopt the implicit stochastic update framework to construct average-reward implicit TD(λ). We establish finite-time error bounds under markedly weaker step-size conditions than those in Zhang et al. (2021b), and we demonstrate that this relaxation enables computationally efficient policy evaluation and learning across a range of examples. The primary contributions of our work are summarized as follows.

- We propose average-reward implicit TD(λ), which is more robust to step-size choice than the standard average-reward TD(λ).
- We provide finite-time error bounds under both constant and diminishing step-sizes, substantially relaxing the step-size conditions required by existing bounds for standard average-reward TD(λ) (Zhang et al., 2021b), thereby explaining the improved numerical stability of the implicit variant.
- In the case of a diminishing step-size, we establish the first finite-time error bounds in the average-reward setting for step-size sequences of the form $\alpha_t \propto t^{-s}$ with $s \in (0,1)$, covering both square-summable ($s > 1/2$) and non-square-summable ($0 < s \leq 1/2$) regimes, thereby further broadening the admissible family of step-sizes.
- We empirically demonstrate the robustness and efficiency of the proposed method through comprehensive experiments in both policy evaluation and control tasks.

2 POLICY EVALUATION IN THE AVERAGE-REWARD SETTING

Problem Formulation. Consider an infinite-horizon Markov decision process (MDP) defined by a finite state space $\mathcal{S}$, a finite action space $\mathcal{A}$, a bounded reward function $r : \mathcal{S} \times \mathcal{A} \to [0,1]$, and a transition function $p : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$. Under a deterministic stationary policy $\mu : \mathcal{S} \to \mathcal{A}$, at time t with a current state S_t^μ, the agent will take an action $A_t^\mu = \mu(S_t^\mu)$, receive a reward $R_t^\mu = r(S_t^\mu, A_t^\mu)$, and transition to next state S_{t+1}^μ according to the probability distribution $p(\cdot|S_t^\mu, A_t^\mu)$. The resulting state sequence $\{S_t^\mu\}_{t\in\mathbb{N}}$ induced by the policy μ forms a Markov chain with one-step transition probabilities $p^\mu\left(S_{t+1}^\mu|S_t^\mu\right) = p\left\{S_{t+1}^\mu|S_t^\mu, A_t^\mu = \mu(S_t^\mu)\right\}$.[1] To simplify notation, let $\mathcal{S} = \{1, 2, \cdots, |\mathcal{S}|\}$ and define time-homogeneous transition probability matrix $\boldsymbol{P}^\mu = \left[P_{ij}^\mu\right]_{i,j=1}^{|\mathcal{S}|}$ with $P_{ij}^\mu = p^\mu\left\{S_{t+1}^\mu = j \mid S_t^\mu = i\right\}$.

[1] Since any Markov reward process arises from an MDP under a fixed policy, the general MDP setting covers the Markov reward process case.

Likewise, let $\boldsymbol{r}^\mu = [r\{1, \mu(1)\}, \cdots, r\{|\mathcal{S}|, \mu(|\mathcal{S}|)\}]^\top$ be the reward vector.

One way to characterize the long-term performance of a given policy μ is via its average-reward, defined for each initial state $s \in \mathcal{S}$ as

$$\omega^\mu(s) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\mu\left(\sum_{t=0}^{T-1} R_t^\mu \;\middle|\; S_0^\mu = s\right),$$

where the expectation is taken over the randomness associated with the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ induced by the policy μ. Although the average-reward provides a natural evaluation criterion for μ, the limit need not exist in general (see, e.g., Chapter 8 of Puterman (2014)). To guarantee the existence and uniqueness of the average-reward, it is common to make the following assumption.

Assumption 2.1. The Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ is irreducible and aperiodic.

Under Assumption 2.1, the chain has a unique stationary distribution $\boldsymbol{\pi}^\mu = (\pi_i^\mu)_{i=1}^{|\mathcal{S}|}$ satisfying $\boldsymbol{\pi}^{\mu\top}\boldsymbol{P}^\mu = \boldsymbol{\pi}^{\mu\top}$ with $\pi_i^\mu > 0$ for every $i \in \mathcal{S}$ (Levin and Peres, 2017). Under the same assumption, one can further show that the average-reward is independent of the initial state (Bertsekas and Tsitsiklis, 1996); that is, $\omega^\mu(s) = \boldsymbol{\pi}^{\mu\top}\boldsymbol{r}^\mu$, $\forall s \in \mathcal{S}$. Unlike its discounted counterpart, the average-reward criterion carries no information about the relative desirability of individual states. To quantify long-run, state-dependent performance under a stationary policy μ, we introduce the basic differential value function $v^\mu : \mathcal{S} \to \mathbb{R}$,

$$v^\mu(s) := \mathbb{E}^\mu\left\{\sum_{t=0}^{\infty}\left(R_t^\mu - \omega^\mu\right) \;\middle|\; S_0^\mu = s\right\},$$

which measures the relative advantage (or disadvantage) of starting in state $s \in \mathcal{S}$. Accordingly, the quantities of interest are the 1) average reward: ω^μ and 2) pairwise contrast: $v^\mu(s) - v^\mu(s')$ for any $s, s' \in \mathcal{S}$ which captures the comparative long-run performance of states $s, s' \in \mathcal{S}$.

In high-dimensional or continuous state spaces, a common strategy is to use linear function approximation, where we model the pairwise difference as

$$v^\mu(s) - v^\mu(s') \approx \{\boldsymbol{\phi}(s) - \boldsymbol{\phi}(s')\}^\top\boldsymbol{\theta}$$

with a user-chosen feature map $\boldsymbol{\phi}(s) \in \mathbb{R}^d$ and weights $\boldsymbol{\theta} \in \mathbb{R}^d$. Because adding a constant to $v^\mu(s)$ leaves all differences unchanged, it suffices to learn $v^\mu(s)$ up to an additive constant. Let $\boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}|\times d}$ be the feature matrix whose i^{th} row is $\boldsymbol{\phi}(i)^\top$ and $\boldsymbol{M} := \operatorname{diag}(\boldsymbol{\pi}^\mu)$. Writing $\boldsymbol{v}^\mu := [v^\mu(1), \ldots, v^\mu(|\mathcal{S}|)]^\top$, one has the series representation $\boldsymbol{v}^\mu = \sum_{t=0}^{\infty}(\boldsymbol{P}^\mu)^t\left(\boldsymbol{r}^\mu - \omega^\mu\boldsymbol{e}\right)$, where $\boldsymbol{e}$ is the all-ones vector. With the weighted norm $\|\boldsymbol{x}\|_{\boldsymbol{M}} := (\boldsymbol{x}^\top\boldsymbol{M}\boldsymbol{x})^{1/2}$, our second goal translates to finding $\boldsymbol{\theta}^*$ such that the weighted discrepancy

$$\inf_{c\in\mathbb{R}}\left\|\boldsymbol{\Phi}\boldsymbol{\theta}^* - (\boldsymbol{v}^\mu + c\boldsymbol{e})\right\|_{\boldsymbol{M}}$$

is small. This quantity is zero when the feature space contains a constant shift of the basic differential value function (i.e., the differential value function). When the weighted discrepancy is small, it indicates that $\boldsymbol{\phi}(s)^\top\boldsymbol{\theta}^*$ approximates $v^\mu(s)$ up to an additive constant, so the estimation of the contrasts $v^\mu(s) - v^\mu(s')$ is correspondingly accurate, improving as the discrepancy decreases.

Average-Reward TD(λ) with Linear Approximation. The average-reward TD(λ) algorithm (Tsitsiklis and Van Roy, 1999) is a widely used stochastic-approximation method to achieve the aforementioned goals. At the tth iteration, the average-reward TD(λ) algorithm maintains both $\widehat{\omega}_t$, an estimate of the average-reward ω^μ, and an estimate $\widehat{\boldsymbol{\theta}}_t$ of the optimal weight $\boldsymbol{\theta}^*$. With non-increasing positive step-sizes α_t, β_t and exponential weighting parameter $\lambda \in [0, 1)$, the update rules are given by

$$\begin{aligned}\widehat{\omega}_{t+1} &= \widehat{\omega}_t + \alpha_t\left(R_t^\mu - \widehat{\omega}_t\right),\\ \widehat{\boldsymbol{\theta}}_{t+1} &= \widehat{\boldsymbol{\theta}}_t + \beta_t\delta_t\boldsymbol{z}_t,\end{aligned} \tag{1}$$

where the eligibility trace $\boldsymbol{z}_t$ and TD error δ_t are

$$\begin{aligned}\boldsymbol{z}_t &= \sum_{i=0}^{t}\lambda^{t-i}\boldsymbol{\phi}\left(S_i^\mu\right),\\ \delta_t &= R_t^\mu - \widehat{\omega}_t + \widehat{\boldsymbol{\theta}}_t^\top\left\{\boldsymbol{\phi}\left(S_{t+1}^\mu\right) - \boldsymbol{\phi}\left(S_t^\mu\right)\right\},\end{aligned}$$

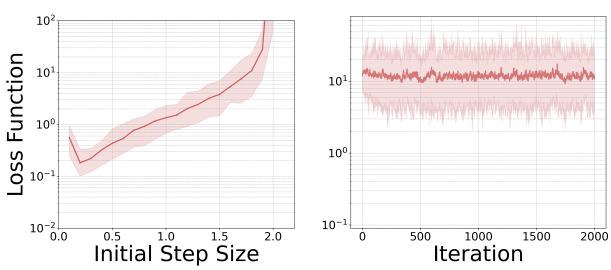
each respectively representing the geometrically weighted average of past feature vectors at visited states and the one-step TD error, which measures how the reward (after subtracting the current average-reward estimate $\widehat{\omega}_t$) plus the estimated value of the next state differs from the current value estimate. We restrict our attention to a single-time-scale average-reward TD(λ) algorithm by assuming $\alpha_t = c_\alpha\beta_t$ with fixed $c_\alpha > 0$ (Tsitsiklis and Van Roy, 1999; Zhang et al., 2021a). Exploring distinct decay rates, an instance of the two-time-scale stochastic approximation framework (Borkar, 2008), is interesting but outside our scope.

Step-Size Sensitivity. Despite its foundational role in RL, standard average-reward TD(λ) suffers from numerical sensitivity to step-size selection. To illustrate this issue, we present a simple numerical example. We consider a Markov reward process (MRP) with $|\mathcal{S}| = 100$ states and evaluate the performance

of average-reward TD(λ) learning with hyperparameter configuration $(c_\alpha, \lambda) = (1.0, 0.25)$ with a predetermined constant step-size $\beta_t = \beta_0 \in (0, 2)$, $\forall t \in \mathbb{N}$. The objective is to estimate both the optimal weight $\boldsymbol{\theta}^*$ and the average-reward ω^μ. Detailed descriptions of the evaluation criterion (loss function) and experimental setting are provided in Sections 4 and 6, respectively.

Figure 1 illustrates the instability induced by step-size choices. The left panel shows a non-monotonic trend in performance: overly small step-sizes (e.g., $\beta_0 < 0.20$) lead to slow convergence, while a modest increase in step-size causes the loss function values to grow rapidly. The right panel presents the result for a moderately large step-size ($\beta_0 = 1.8$), where the average-reward TD(λ) iterates exhibit oscillatory behavior. These empirical findings highlight the sensitivity of the TD learning to step-size selection and motivate the need for algorithms that are robust to such choices. In the following sections, we propose and analyze one such approach.

Figure 1: Sensitivity of average-reward TD(λ) to the choice of step-size with exponential weighting parameter $\lambda = 0.25$ and step-size ratio $c_\alpha = 1.0$. The solid line denotes the mean, and the shaded region indicates the 95% confidence interval.



(a) Performance of average-reward TD(λ) for step-sizes β_0 ranging from 0.1 to 2.0.

(b) Performance over iterations with step-size $\beta_0 = 1.8$, showing no improvement.

3 AVERAGE-REWARD IMPLICIT TD(λ)

As we have seen in the previous section, average-reward TD(λ) demands carefully tuned step-sizes for stability. Implicit stochastic recursions, developed for stochastic gradient descent (Toulis et al., 2014; Toulis and Airoldi, 2014; Toulis et al., 2021; Chee et al., 2023) and more recently extended to discounted-reward on- and off-policy TD (Kim et al., 2025), rewrite each update as a fixed-point equation by allowing the gradient or TD error to depend on the new iterate. This reformulation automatically induces an adaptive shrinkage in the effective step-size, vastly improving numerical stability without increasing computational complexity. Building on this idea, we introduce the average-reward implicit TD(λ), which retains the simplicity and efficiency of standard average-reward TD(λ) while supporting more flexible step-size choices with finite-time error guarantees. In this section, we propose a novel average-reward TD(λ) algorithm that incorporates implicit updates into its recursive structure.

To derive implicit average-reward TD(λ) updates, recall the update rule for $\widehat{\boldsymbol{\theta}}_{t+1}$ given in (1):

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{t+1} &= \widehat{\boldsymbol{\theta}}_t + \beta_t \left(R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\right) \boldsymbol{z}_t \\ &= \widehat{\boldsymbol{\theta}}_t + \beta_t \left\{R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - (\boldsymbol{z}_t - \lambda \boldsymbol{z}_{t-1})^\top \widehat{\boldsymbol{\theta}}_t\right\} \boldsymbol{z}_t \\ &= \widehat{\boldsymbol{\theta}}_t + \beta_t \left(R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t + \lambda \boldsymbol{z}_{t-1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\theta}}_t\right) \boldsymbol{z}_t,\end{aligned}$$

where we have used the identity $\boldsymbol{\phi}_t = \boldsymbol{z}_t - \lambda \boldsymbol{z}_{t-1}$ in the second equality. At time t, we update $\widehat{\omega}_t$ and $\widehat{\boldsymbol{\theta}}_t$ using a rule that depends on both the current iterate $(\widehat{\omega}_t, \widehat{\boldsymbol{\theta}}_t)$ and their updated values $(\widehat{\omega}_{t+1}, \widehat{\boldsymbol{\theta}}_{t+1})$:

$$\widehat{\omega}_{t+1} = \widehat{\omega}_t + c_\alpha \beta_t (R_t^\mu - \widehat{\omega}_{t+1}), \tag{2}$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t (R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t + \lambda \boldsymbol{z}_{t-1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{z}_t^\top \widehat{\boldsymbol{\theta}}_{t+1}) \boldsymbol{z}_t. \tag{3}$$

Solving the above recursions (2) and (3) admits the update rule for the average-reward implicit TD(λ) algorithm. Lemma 3.1 below characterizes the average-reward implicit TD(λ) algorithm, and its proof is given in the supplementary materials.

Lemma 3.1. *Average-reward implicit TD(λ) updates given in* (2) *and* (3) *can be expressed as*

$$\widehat{\omega}_{t+1} = \widehat{\omega}_t + \frac{c_\alpha \beta_t}{1 + c_\alpha \beta_t} (R_t^\mu - \widehat{\omega}_t)$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1 + \beta_t \|\boldsymbol{z}_t\|_2^2} \left(R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top \widehat{\boldsymbol{\theta}}_t\right) \boldsymbol{z}_t.$$

The update rule in Lemma 3.1 highlights a key mechanism underlying the robustness of the average-reward implicit TD(λ) learning. Unlike the standard average-reward TD(λ) methods, at each step $t \in \mathbb{N}$, the implicit updates dynamically rescale the step-size based on the magnitude of the eligibility trace as well as the step-size ratio parameter. Such shrinkage arises naturally from the implicit update mechanism, reducing the burden of laborious tuning. Importantly, the implicit algorithm has the same space and time complexity as the standard method in the average-reward setting, making it a practical replacement without additional computational burden or implementation difficulty. The benefits of this mechanism will be further clarified in the forthcoming theoretical analysis and subsequently illustrated through a suite of numerical examples.

To further enhance the numerical stability and facilitate the theoretical analysis of the average-reward implicit TD(λ) algorithm, we introduce a projection step

that forces each iterate $\widehat{\boldsymbol{\Theta}}_t := [\widehat{\omega}_t, \widehat{\boldsymbol{\theta}}_t]^\top$ to lie within the Euclidean ball of radius $R_{\boldsymbol{\Theta}}$ by enforcing the constraints $\|\boldsymbol{\Theta}\|_2 \leq R_{\boldsymbol{\Theta}}$. Such projection-based stabilization is well-studied in both the stochastic optimization and reinforcement learning literatures, and numerous theoretical results have been established (Nemirovski et al., 2009; Bubeck, 2015; Bhandari et al., 2018; Xu et al., 2019; Zou et al., 2019; Zhang et al., 2023). In practice, one can choose $R_{\boldsymbol{\Theta}}$ sufficiently large to ensure the limit point of $\widehat{\boldsymbol{\Theta}}_t$ is contained in the Euclidean ball of radius $R_{\boldsymbol{\Theta}}$. A complete algorithmic description of the average-reward implicit TD(λ) is provided in Algorithm 1.

Algorithm 1 Average-reward implicit TD(λ) (with projection)

Input: exponential weighting parameter $\lambda \in [0, 1)$, basis functions $\{\boldsymbol{\phi}_k\}_{k=1}^d$, step-size $\{\beta_t\}_{t\in\mathbb{N}}$, step-size ratio parameter c_α, projection radius $R_{\boldsymbol{\Theta}}$

Initialize: $\widehat{\omega}_0$, $\widehat{\boldsymbol{\theta}}_0$, S_0^μ and eligibility trace $\boldsymbol{z}_{-1} = 0$.

for $t = 0, 1, 2, \ldots$ **do**

Receive data: $(S_t^\mu, R_t^\mu, S_{t+1}^\mu)$

Get TD error:

$$\delta_t = R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}(S_{t+1}^\mu)^\top \widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}(S_t^\mu)^\top \widehat{\boldsymbol{\theta}}_t$$

Update eligibility trace: $\boldsymbol{z}_t = \lambda \boldsymbol{z}_{t-1} + \boldsymbol{\phi}(S_t^\mu)$

Update parameters:

$$\widehat{\omega}_{t+1} = \widehat{\omega}_t + \frac{c_\alpha \beta_t}{1 + c_\alpha \beta_t} \left(R_t^\mu - \widehat{\omega}_t\right),$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1 + \beta_t \|\boldsymbol{z}_t\|_2^2} \delta_t \boldsymbol{z}_t$$

For projected average-reward implicit TD(λ): if $(\widehat{\omega}_{t+1})^2 + \|\widehat{\boldsymbol{\theta}}_{t+1}\|^2 \geq R_{\boldsymbol{\Theta}}^2$,

$$\widehat{\omega}_{t+1} = \frac{R_{\boldsymbol{\Theta}}}{\sqrt{(\widehat{\omega}_{t+1})^2 + \|\widehat{\boldsymbol{\theta}}_{t+1}\|_2^2}} \widehat{\omega}_{t+1},$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \frac{R_{\boldsymbol{\Theta}}}{\sqrt{(\widehat{\omega}_{t+1})^2 + \|\widehat{\boldsymbol{\theta}}_{t+1}\|_2^2}} \widehat{\boldsymbol{\theta}}_{t+1}$$

end for

4 THEORETICAL ANALYSIS

In this section, we provide a theoretical analysis of the average-reward implicit TD(λ) algorithm incorporating a projection step. We first assume that the columns of $\boldsymbol{\Phi}$ are linearly independent, which implies that the basis functions span a d-dimensional feature space. Such an assumption ensures that redundant basis functions can be removed without loss of expressiveness. Hereafter, we use $\|\cdot\|$ to denote the Euclidean norm for vectors and operator norm for matrices. We assume that the feature vectors are normalized so that $\|\boldsymbol{\phi}(i)\| \leq 1$ for all $i \in \mathcal{S}$. Lastly, $\mathbb{E}^\mu$ denotes expectation with respect to the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ under policy μ with a fixed S_0^μ, and $\mathbb{E}^{\boldsymbol{\pi}^\mu}$ denotes expectation with respect to the stationary distribution of this chain.

To facilitate our analysis, we formulate the average-reward TD(λ) update into a matrix notation form, given by

$$\begin{bmatrix} \widehat{\omega}_{t+1} \\ \widehat{\boldsymbol{\theta}}_{t+1} \end{bmatrix} = \begin{bmatrix} \widehat{\omega}_t \\ \widehat{\boldsymbol{\theta}}_t \end{bmatrix} + \beta_t \begin{bmatrix} -c_\alpha & 0 \\ -\boldsymbol{z}_t & \boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top) \end{bmatrix} \begin{bmatrix} \widehat{\omega}_t \\ \widehat{\boldsymbol{\theta}}_t \end{bmatrix} + \begin{bmatrix} c_\alpha R_t^\mu \\ R_t^\mu \boldsymbol{z}_t \end{bmatrix},$$

which can be succinctly written as

$$\widehat{\boldsymbol{\Theta}}_{t+1} = \widehat{\boldsymbol{\Theta}}_t + \beta_t \{\boldsymbol{A}(\boldsymbol{X}_t)\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}(\boldsymbol{X}_t)\} \tag{4}$$

and its implicit version is given by

$$\widehat{\boldsymbol{\Theta}}_{t+1} = \widehat{\boldsymbol{\Theta}}_t + \boldsymbol{D}_t \{\boldsymbol{A}(\boldsymbol{X}_t)\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}(\boldsymbol{X}_t)\} \tag{5}$$

where

$$\widehat{\boldsymbol{\Theta}}_t := \begin{bmatrix} \widehat{\omega}_t \\ \widehat{\boldsymbol{\theta}}_t \end{bmatrix}, \quad \boldsymbol{A}(\boldsymbol{X}_t) := \begin{bmatrix} -c_\alpha & 0 \\ -\boldsymbol{z}_t & \boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top) \end{bmatrix},$$

$$\boldsymbol{b}(\boldsymbol{X}_t) := \begin{bmatrix} c_\alpha R_t^\mu \\ R_t^\mu \boldsymbol{z}_t \end{bmatrix}, \quad \boldsymbol{D}_t := \begin{bmatrix} \frac{1}{1+c_\alpha \beta_t} & 0 \\ 0 & \frac{1}{1+\beta_t ||\boldsymbol{z}_t||^2} \boldsymbol{I}_d \end{bmatrix}$$

with $\boldsymbol{X}_t := (S_t^\mu, S_{t+1}^\mu, \boldsymbol{z}_t)$.

Under suitable technical conditions, if $\boldsymbol{A} := \mathbb{E}^{\boldsymbol{\pi}^\mu}[\boldsymbol{A}(\boldsymbol{X}_t)]$ is negative definite, results from stochastic approximation (Benveniste et al., 2012) imply that the iterate $\widehat{\boldsymbol{\Theta}}_t$ converges almost surely to $\boldsymbol{\Theta}^* = [\omega^\mu, \boldsymbol{\theta}^*]^\top$, which solves $\boldsymbol{A}\boldsymbol{\Theta}^* + \boldsymbol{b} = 0$ with $\boldsymbol{b} := \mathbb{E}^{\boldsymbol{\pi}^\mu}[\boldsymbol{b}(\boldsymbol{X}_t)]$. Earlier work (Tsitsiklis and Van Roy, 1999) established almost sure convergence $\widehat{\boldsymbol{\Theta}}_t$ to $\boldsymbol{\Theta}^*$ by assuming $\boldsymbol{A}$ is negative definite (up to left multiplication by a diagonal matrix). However, such an assumption excludes cases where the feature matrix $\boldsymbol{\Phi}$ can yield value predictions that are constant across all states, i.e., when $\boldsymbol{e}$ lies in the column space of $\boldsymbol{\Phi}$.

To relax aforementioned assumption, (Zhang et al., 2021b) considered an auxiliary iterate $[\widehat{\omega}_t, \Pi_{\mathbb{O}}\widehat{\boldsymbol{\theta}}_t]^\top$, where $\Pi_{\mathbb{O}}$ denotes projection onto $\mathbb{O}$, defined as the orthogonal complement of $\mathbb{S}_{\boldsymbol{\Phi},\boldsymbol{e}} := \text{span}\{\boldsymbol{\theta} : \boldsymbol{\Phi}\boldsymbol{\theta} = \boldsymbol{e}\}$. Projecting onto $\mathbb{O}$ thus removes the constant-shift direction, i.e., the component of $\boldsymbol{\Phi}\boldsymbol{\theta}$ aligned with the all-one vector direction. Such component adds the same constant to every state's value prediction and does not affect estimates of value contrasts $v^\mu(s) - v^\mu(s')$. Accordingly, it is natural to assess performance using the projected iterate $\Pi_{\mathbb{O}}\widehat{\boldsymbol{\theta}}_t$ since any change in $\widehat{\boldsymbol{\theta}}_t$ along $\text{span}\{\boldsymbol{e}\}$ is not identifiable in the average-reward setting and can be ignored. Furthermore, on $\mathbb{R} \times \mathbb{O}$, one can restore strict negative definiteness of the matrix $\boldsymbol{A}$, formalized as Lemma 4.1 below.

Lemma 4.1 (Lemma 2 of (Zhang et al., 2021b)). *For $\lambda \in (0,1)$, let $\boldsymbol{M} = diag\left(\pi_1^\mu, \cdots, \pi_{|\mathcal{S}|}^\mu\right)$ and $\boldsymbol{P}^{(\lambda)} = (1-\lambda)\sum_{m=0}^{\infty} \lambda^m (\boldsymbol{P}^\mu)^{m+1}$. Under Assumption 2.1, we have $\Delta > 0$ where*

$$\Delta := \min_{\|\boldsymbol{\theta}\|=1, \boldsymbol{\theta}\in\mathbb{O}} \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top \boldsymbol{M}\left(\boldsymbol{I} - \boldsymbol{P}^{(\lambda)}\right)\boldsymbol{\Phi}\boldsymbol{\theta}.$$

In addition, for $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2(1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$,

$$\boldsymbol{\Theta}^\top \boldsymbol{A}\boldsymbol{\Theta} \leq -\frac{\Delta}{2}\|\boldsymbol{\Theta}\|^2, \quad \textit{for any} \quad \boldsymbol{\Theta} \in \mathbb{R}\times\mathbb{O}.$$

With the negative definiteness of the matrix $\boldsymbol{A}$, the limit point $\boldsymbol{\Theta}^*$ is assured to be unique and one can then ask how far the auxiliary iterates are from the limit point. Specifically non-asymptotic bounds on $(\widehat{\omega}_t - \omega^\mu)^2 + \left\|\Pi_{\mathbb{O}}\left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right)\right\|^2$ were established both for constant and decreasing step-size schedules (Zhang et al., 2021b). In addition to the finite-time error bounds, the approximation quality of $\boldsymbol{\theta}^*$ within the chosen feature class is captured by the $\boldsymbol{M}$-weighted discrepancy

$$\inf_{c\in\mathbb{R}} \|\boldsymbol{\Phi}\boldsymbol{\theta}^* - (\boldsymbol{v}^\mu + c\boldsymbol{e})\|_{\boldsymbol{M}} \leq \frac{\inf_{\boldsymbol{\theta}\in\mathbb{R}^d,\, c\in\mathbb{R}} \|\boldsymbol{\Phi}\boldsymbol{\theta} - (\boldsymbol{v}^\mu + c\boldsymbol{e})\|_{\boldsymbol{M}}}{\sqrt{1-c_\lambda^2}}.$$

with $c_\lambda \in [0,1)$ and $c_\lambda \to 0$ as $\lambda \to 1$ (Zhang et al., 2021b). Note that the right-hand term involves the best error achievable within the feature class. Thus $\boldsymbol{\theta}^*$ is optimal up to a multiplicative factor, which approaches 1 as $\lambda \to 1$. In particular, if the feature class is rich enough to represent any one differential value function, the best achievable error is zero; that is, $\inf_{c\in\mathbb{R}} \left\|\boldsymbol{\Phi}\boldsymbol{\theta}^* - (\boldsymbol{v}^\mu + c\boldsymbol{e})\right\|_{\boldsymbol{M}} = 0$.

Non-asymptotic Analysis of Average-Reward Implicit TD(λ). We are now ready to present finite-time error bounds for average-reward implicit TD(λ) with the projection step, formally stated in Theorems 4.4 and 4.5. Results are provided for both constant and decreasing step-sizes. The bounds are expressed in terms of the negative-definiteness margin Δ from Lemma 4.1, the step-size parameters c_α (ratio: α_t/β_t), β_0 (initial step-size), and s (decay rate) and the mixing time of the underlying Markov process $\{S_t^\mu\}_{t\in\mathbb{N}}$ whose formal definition is given below.

Definition 4.2 (Mixing Time). Let $\{S_t\}_{t\in\mathbb{N}} \subset \mathcal{S}$ be a Markov chain with stationary distribution π. For $\epsilon \in (0,1)$, its mixing time is the smallest positive integer $\tau_\epsilon \in \mathbb{N}$ such that for all $t \geq \tau_\epsilon$,

$$\sup_{s\in\mathcal{S}} d_{\mathrm{TV}}\left\{\mathbb{P}(S_t = \cdot \mid S_0 = s), \pi(\cdot)\right\} \leq \epsilon,$$

where d_{TV} denotes the total variation distance.

Remark 4.3. Under Assumption 2.1, the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic: there exist $m > 0$ and $\rho \in (0,1)$ such that $\sup_{s\in\mathcal{S}} d_{\mathrm{TV}}\left\{p^\mu(S_t^\mu = \cdot \mid S_0^\mu = s), \pi^\mu\right\} \leq m\rho^t$. Consequently, its mixing time τ_ϵ is of order $\mathcal{O}\left(\log\frac{1}{\epsilon}\right)$.

Theorem 4.4. *Suppose the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic with a rate parameter $\rho \in (0,1)$, the step-size ratio parameter is chosen to satisfy $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2(1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$ and the optimal parameter $\|\boldsymbol{\Theta}^*\| \leq R_{\boldsymbol{\Theta}}$. With $\lambda \in [0,1)$ and constant step-size $\beta_t = \beta$, the iterates of the projected average-reward implicit TD(λ) algorithm satisfy the following finite-time error bound*

$$\begin{aligned}
&\mathbb{E}^\mu\left\{(\widehat{\omega}_{t+1} - \omega^\mu)^2 + \left\|\Pi_{\mathbb{O}}\left(\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^*\right)\right\|^2\right\} \\
&\lesssim (1-\beta\gamma\Delta)^{t+1}\left\{(\widehat{\omega}_0^\mu - \omega^\mu)^2 + \left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\right\|^2\right\} \\
&\quad + \mathcal{O}\left(\beta\tau_\beta + h^{\tau_\beta} + \beta\tau_\beta t h^t\right), \quad t \geq 0
\end{aligned}$$

where $\gamma = \min\left\{\frac{1}{1+c_\alpha\beta}, \frac{(1-\lambda)^2}{(1-\lambda)^2+\beta}\right\}$ and $h = \max\{1 - \beta\gamma\Delta, \rho, \lambda\}$.

Theorem 4.5. *Suppose the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic with a rate parameter $\rho \in (0,1)$, the step-size ratio parameter is chosen to satisfy $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2(1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$ and the optimal parameter $\|\boldsymbol{\Theta}^*\| \leq R_{\boldsymbol{\Theta}}$. With $\lambda \in [0,1)$ and decreasing step-sizes $\beta_t = \frac{\beta_0}{(t+1)^s}$, $s \in (0,1)$, the iterates of the projected average-reward implicit TD(λ) algorithm satisfy the following finite-time error bound*

$$\begin{aligned}
&\mathbb{E}^\mu\left\{(\widehat{\omega}_{t+1} - \omega^\mu)^2 + \left\|\Pi_{\mathbb{O}}\left(\widehat{\boldsymbol{\theta}}_{t+1} - \boldsymbol{\theta}^*\right)\right\|^2\right\} \\
&\lesssim \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s} - 1\}\right]\left\{(\widehat{\omega}_0^\mu - \omega^\mu)^2 + \left\|\widehat{\boldsymbol{\theta}}_0 - \boldsymbol{\theta}^*\right\|^2\right\} \\
&\quad + \mathcal{O}\left\{\tau_{\beta_t} t\exp\left(-ct^{1-s}\right) + \tau_{\beta_t} t^{-s} + q^{\tau_{\beta_t}}\right\}, \quad t \geq 0
\end{aligned}$$

for some constant $c > 0$, $q = \max\{\rho, \lambda\}$ and $\gamma_0 = \min\left\{\frac{1}{1+c_\alpha\beta_0}, \frac{(1-\lambda)^2}{(1-\lambda)^2+\beta_0}\right\}$.

Remark 4.6. Our two theorems substantially relax the restrictive conditions required in (Zhang et al., 2021b).

- For constant step-sizes, we establish a finite-time bound without any initial step-size requirements. By contrast, previous analysis ties the initial step-size to a problem dependent quantity (e.g., $\Delta\beta < 2$) and further imposes restrictive upper bounds on the step-size as well as the mixing time (Zhang et al., 2021b).
- For decaying step-sizes, our theorem accommodates the polynomial schedule of the form $\beta_t = \beta_0/(t+1)^s$ for any $s \in (0,1)$, not just the $1/t$ rate

covered in (Zhang et al., 2021b). In addition, the bounds in (Zhang et al., 2021b) hold only under extra restrictions, for example, there must exist an index $t^* \in \mathbb{N}$ with bounded cumulative step-size up to t^*, and all subsequent step-sizes must stay below a problem-dependent threshold.

These relaxations remove delicate initial step-size conditions and broaden the range of admissible step-size schedules, while still providing finite-time error guarantees.

5 RELATED WORKS

Our work is motivated by the literature on implicit stochastic gradient methods, which recast standard gradient based recursions as fixed-point equations to improve numerical stability. This includes the implicit SGD line of works (Toulis et al., 2014; Toulis and Airoldi, 2014; Chee et al., 2023) as well as Tamar et al. (2014), who first proposed applying implicit updates to TD(λ) under the restrictive zero-reward assumption. Kim et al. (2025) extended this direction to the discounted-reward setting, establishing asymptotic convergence for implicit TD(0), TD(λ), and off-policy TD along with finite-time error bounds for their projected variants. This idea has since been further extended to control settings, including implicit variants of Q-learning and SARSA (Kim and Laber, 2026). Building on these works, we develop an implicit TD(λ) algorithm for the more challenging average-reward setting, where TD lacks a unique limit point, as the value function is only determined up to an additive constant, causing the TD iterates to converge to an initialization-dependent limit and complicating both algorithm design and analysis (Tsitsiklis and Van Roy, 1999; Zhang et al., 2021b). By incorporating a projection step on the parameter iterates and employing information-theoretic techniques, we establish all-time finite-time error guarantees while eliminating problem-dependent constraints on step-size choices, and further demonstrate strong empirical performance of the proposed algorithm.

While building upon these foundational contributions, our work also addresses critical limitations in prior finite-time analyses of average-reward TD. Tsitsiklis and Van Roy (1999) introduced average-reward TD(λ) and established asymptotic convergence, but under the unrealistic assumption that the all-ones vector $\boldsymbol{e}$ does not lie in the span of the feature basis. Zhang et al. (2021b) relaxed this assumption and proposed evaluating performance via the projected distance onto the subspace orthogonal to the constant direction, a metric we adopt as it enables finite-time analysis without unrealistic assumptions. However, their bounds hold only for $t \geq T^\star$, where $T^\star$ is a problem-dependent threshold depending on mixing time and step-size, offering no insight into early-time behavior. Furthermore, their step-size must satisfy stringent problem-dependent conditions that make practical calibration difficult. By contrast, our approach yields bounds that hold for all $t \in \mathbb{N}$ and completely eliminates problem-dependent constraints on step-size selection. In particular, our results establish that implicit average-reward TD exhibits no divergent behavior for any step-size choice, a stark contrast to the vanilla formulation, resolving the practical calibration burden that hampers prior methods.

6 NUMERICAL EXPERIMENTS

In this section, we demonstrate the effectiveness of the proposed average-reward implicit TD(λ) relative to standard average-reward TD(λ) on both policy evaluation and control tasks. All experiments were carried out on Intel(R) Xeon(R) Gold 6152 CPUs at 2.10 GHz with 32 GB RAM.

6.1 Policy Evaluation

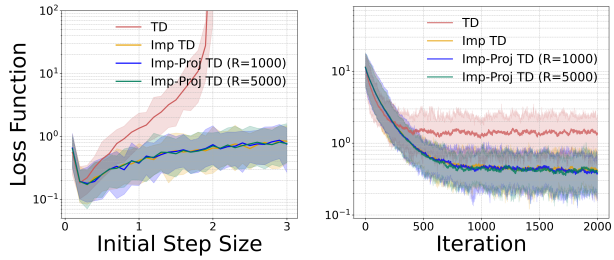
For policy evaluation, we use MRP and the Boyan chain examples; for policy learning, we consider the access-control and pendulum problems. Performance is quantified by the loss function given by $\left(\widehat{\omega}_t - \omega^\mu\right)^2 + \left\|\Pi_{\mathbb{O}}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\right\|^2$. A detailed description of how the loss value is computed is provided in the supplementary materials. We consider both constant and decaying step-size schedules. For decaying schedules, the step-size is held fixed for the first 150 iterations to promote exploration and then decreased thereafter. Each configuration is run for $T = 2000$ steps with $\widehat{\omega}_0^\mu = 0$ and $\widehat{\boldsymbol{\theta}}_0 \sim \mathrm{Unif}([-1,1]^d)$, and results are averaged over 50 independent trials. We fix the step-size ratio $c_\alpha = 1$ and the exponential weighting parameter $\lambda = 0.25$. We compare four methods: (i) average-reward standard TD(λ); (ii) average-reward implicit TD(λ) without projection; and (iii–iv) average-reward implicit TD(λ) with projection, using projection radius $R_{\boldsymbol{\Theta}} \in \{1000, 5000\}$. Full implementation details and additional experimental results are provided in the supplementary materials.

6.1.1 Markov Reward Process

Here we study an MRP with $|\mathcal{S}| = 100$ states; the transition matrix and reward vector are generated following Zhang et al. (2021b). We first consider constant step-sizes $\beta_t \equiv \beta_0$ with $\beta_0 \in \{0.1, \ldots, 3.0\}$. Figure 2 summarizes the results (solid line = mean, shaded band = 95% confidence interval). As shown

in the left panel, the average loss across 50 independent experiments increases for all methods as β_0 becomes larger, around $\beta_0 \approx 0.3$ or more. However, as $\beta_0 \to 2$, standard average-reward TD(λ) becomes unstable and its loss explodes, whereas average-reward implicit TD(λ) remains numerically stable with modest loss growth. To compare all four algorithms' behavior at a moderate step-size, the right panel fixes a step-size value $\beta_0 = 1.0$ and tracks performance over iterations: average-reward implicit TD(λ) maintains relatively low error throughout, while standard average-reward TD(λ) incurs substantially larger error over the horizon.

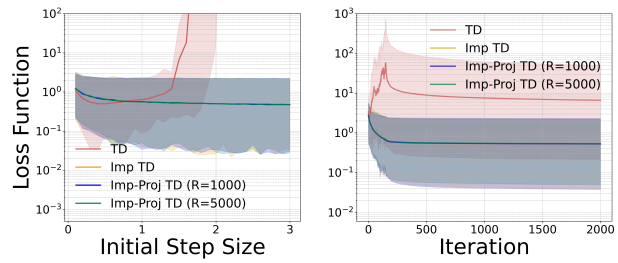
Figure 2: MRP experiment under constant step-size, with exponential weighting parameter and step-size ratio set to $(\lambda, c_\alpha) = (0.25, 1.0)$. The solid line represents the mean, and the shaded region denotes the 95% confidence interval. (Left) Loss value from step-size 0.1 to 3.0. (Right) Loss value over iterations with initial step-size $\beta_0 = 1.0$;



6.1.2 Average-Reward Boyan Chain

We next study the average-reward Boyan chain under deterministic policies. As in the MRP setting, the Boyan chain is a standard benchmark for TD learning (Boyan, 2002). Because the original formulation is not average-reward, we use the variant proposed by (Zhang et al., 2021a), which has 13 states and 2 actions. In each experiment, we construct a deterministic policy by assigning an action to each state via independent Bernoulli(0.5) draws. As with the MRP experiments, we assess performance using the average loss across 50 independent runs. Figure 3 shows results on the Boyan chain example under the decaying step-size schedule $\beta_t = \beta_0/(t+1)^{0.99}$. In the left panel, average-reward implicit TD(λ) methods remain stable across $\beta_0 \in [0.1, 3.0]$, whereas the standard average-reward TD(λ) becomes unstable and its loss grows rapidly as β_0 approaches 1.5. The right panel displays learning curves for $\beta_t = 1.5/(t+1)^{0.99}$ over 2000 iterations. The loss of standard average-reward TD(λ) method consistently exceeds that of the average-reward implicit TD(λ) methods, highlighting the latter's improved numerical stability and superior performance.

Figure 3: Boyan experiment with exponential weighting parameter and step-size ratio set to $(\lambda, c_\alpha) = (0.25, 1.0)$ under decaying step-size schedule $\beta_t = \beta_0/(t+1)^{0.99}$. Solid lines denote the mean, and shaded regions represent 95% confidence intervals. (Left) Loss value with initial step-sizes ranging from 0.1 to 3.0. (Right) Loss value over iterations with initial step-size $\beta_0 = 1.5$.



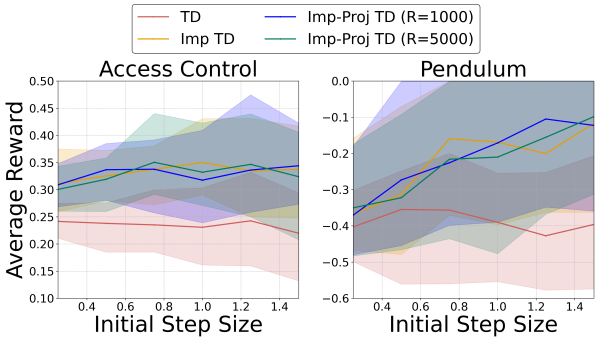
6.2 Control Experiments

In this section, we utilize the proposed average-reward implicit TD(λ) on control tasks. We use state–action–reward–state–action (SARSA) with linear function approximation to estimate the action-value function. Each experiment comprises $T = 15000$ time steps and is repeated over 30 independent runs. We employ a decaying step-size schedule $\beta_t = \beta_0/(t+400)^{0.99}$, holding β_t constant for the first 150 iterations to encourage early exploration before gradually reducing it thereafter.

6.2.1 Access-Control Queueing

We study the canonical access-control queuing problem (Barto, 2021) in the average-reward setting. At each decision epoch, an arriving customer belongs to one of four equiprobable classes, and the agent chooses whether to admit or reject the customer. There are ten identical servers; if admitted, the customer yields an immediate reward and occupies a server. Service completion occurs independently at each step with a fixed probability, inducing stochastic transitions in server availability. The goal is to learn an admission policy that optimally maps the current customer class and the number of available servers to an admit/reject decision. We illustrate average-reward learning results in the left panel of Figure 4. The average-reward implicit TD(λ) methods consistently outperform the average-reward TD(λ) method across varying initial step-sizes in terms of average-reward. As the initial step-size increases, the implicit methods show a mild performance gain and remain stable whereas standard version deteriorates and fails to benefit from larger steps.

Figure 4: Control experiment with exponential weighting parameter and step-size ratio parameter $(\lambda, c_\alpha) = (0.25, 1.0)$, under the decaying step-size schedule $\beta_t = \beta_0/(t+400)^{0.99}$. Initial step-size ranges from 0.25 to 1.5. Solid lines denote the mean, and shaded regions represent 95% confidence intervals.



6.2.2 Pendulum Environment

We also apply the proposed average-reward implicit TD(λ) to the `Pendulum-v1` environment. Because the environment is defined with episodic terminations (Towers et al., 2024), we modify it to match the infinite-horizon average-reward setting. The control objective is to keep the pendulum upright. The right panel of Figure 4 shows results for the pendulum environment with $(\lambda, c_\alpha) = (0.25, 1.0)$. Mirroring the access-control task, larger initial step-sizes benefit the average-reward implicit TD(λ) methods, which achieve higher average reward and remain stable, whereas standard average-reward TD(λ) fails to benefit from larger step-sizes and exhibits no improvement.

7 CONCLUSION

We introduced average-reward implicit TD(λ), a fixed-point variant of the average reward TD(λ) that preserves per iteration complexity while markedly improving stability to step-size choices. Our theoretical guarantees provide explicit finite-time error bounds under both constant and decaying step-sizes, established via a projection-based analysis. Across policy evaluation and control examples, the implicit updates deliver robust performance over a wide range of step-sizes, demonstrating strong practical performance consistent with the theory. Looking ahead, promising directions include a full theoretical analysis of average-reward implicit SARSA and rigorous two-time-scale extensions of both standard and implicit average-reward TD(λ). A related direction is to build implicit TD methods to estimate the asymptotic variance of cumulative reward in the average-reward regime.

References

Jinane Abounadi, Dimitri Bertsekas, and Vivek S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698, 2001.

Shubhada Agrawal, L. A. Prashanth, and Siva Theja Maguluri. Policy evaluation for variance in average reward reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Andrew G. Barto. Reinforcement learning: An introduction (book review). *SIAM Review*, 63(2):423–426, 2021.

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

Dimitri Bertsekas and John N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

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

Vivek S. Borkar. *Stochastic approximation: A dynamical systems viewpoint*, volume 100. Springer, 2008.

Justin A. Boyan. Technical update: Least-squares temporal difference learning. *Machine Learning*, 49 (2):233–246, 2002.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

Jerry Chee, Hwanwoo Kim, and Panos Toulis. "plus/minus the learning rate": Easy and scalable statistical inference with SGD. In *International Conference on Artificial Intelligence and Statistics*, pages 2285–2309. PMLR, 2023.

William Dabney and Andrew Barto. Adaptive step-size for online temporal difference learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 872–878, 2012.

Jim G. Dai and Mark Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022.

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

Vektor Dewanto, George Dunn, Ali Eshragh, Marcus Gallagher, and Fred Roosta. Average-reward model-free reinforcement learning: A systematic review and literature mapping. *arXiv preprint*, 2020.

Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: A reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.

Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, 2012.

Hwanwoo Kim and Eric Laber. Implicit q-learning and sarsa: Liberating policy control from step-size calibration. *arXiv preprint arXiv:2601.18907*, 2026.

Hwanwoo Kim, Panos Toulis, and Eric Laber. Stabilizing temporal difference learning via implicit stochastic approximation. *arXiv preprint*, 2025.

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

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

Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine Learning*, 22(1):159–195, 1996.

Aritra Mitra. A simple finite-time analysis of TD learning with linear function approximation. *IEEE Transactions on Automatic Control*, 2024.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680, 2006.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *Advances in neural information processing systems*, 21, 2008.

Satinder P. Singh. Reinforcement learning algorithms for average-payoff Markovian decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 94, pages 700–705, 1994.

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

Richard S. Sutton, Hamid Maei, and Csaba Szepesvári. A convergent O(n) temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in Neural Information Processing Systems*, 21, 2008.

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, pages 993–1000, 2009.

Aviv Tamar, Panos Toulis, Shie Mannor, and Edoardo M Airoldi. Implicit temporal differences. *arXiv preprint*, 2014.

Poj Tangamchit, John M. Dolan, and Pradeep K. Khosla. The necessity of average rewards in cooperative multirobot learning. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, volume 2, pages 1296–1301. IEEE, 2002.

Gerald Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3):58–68, 1995.

Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *arXiv preprint*, August 2014.

Panos Toulis, Edoardo M Airoldi, and Jason Rennie. Statistical analysis of stochastic gradient methods for generalized linear models. In *Proceedings of the International Conference on Machine Learning*, volume 32, pages 667–675. PMLR, June 2014.

Panos Toulis, Thibaut Horel, and Edoardo M Airoldi. The proximal Robbins–Monro method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1):188–212, February 2021.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun K. G., et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint*, 2024.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5), 1997.

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

John N. Tsitsiklis and Benjamin Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49:179–191, 2002.

Yi Wan, Abhishek Naik, and Richard S. Sutton. Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR, 2021.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. *Advances in Neural Information Processing Systems*, 32, 2019.

Huizhen Yu and Dimitri P. Bertsekas. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

Shangtong Zhang, Yi Wan, Richard S. Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *International Conference on Machine Learning*, pages 12578–12588. PMLR, 2021a.

Shangtong Zhang, Remi Tachet Des Combes, and Romain Laroche. On the convergence of SARSA with linear function approximation. In *International Conference on Machine Learning*, pages 41613–41646. PMLR, 2023.

Sheng Zhang, Zhe Zhang, and Siva Theja Maguluri. Finite sample analysis of average-reward TD learning and Q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242, 2021b.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for SARSA with linear function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

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. [Not Applicable]

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, provided in supplementary materials]
   (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). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (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). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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. [Yes]
   (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. [Yes]
   (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]

Implicit Updates for Average-Reward Temporal Difference Learning: Supplementary Materials

A List of notations

To provide rigorous details behind the established theoretical results, we provide a summary of the necessary assumptions, notations, and facts.

- $\mathcal{S} = \{1, \cdots, |\mathcal{S}|\}$: state space
- $\{S_t^\mu\}_{t\in\mathbb{N}}$: a sequence of states under policy μ
- $(A_t^\mu)_{t\in\mathbb{N}}$: a sequence of actions under policy μ with $A_t^\mu := \mu(S_t^\mu)$
- $(R_t^\mu)_{t\in\mathbb{N}}$: a sequence of rewards under policy μ with $R_t^\mu = r(S_t^\mu, A_t^\mu)$
- $p^\mu\left(S_{t+1}^\mu | S_t^\mu\right) := p\left\{S_{t+1}^\mu | S_t^\mu, A_t^\mu = \mu(S_t^\mu)\right\}$: transition probabilities under policy μ
- $\boldsymbol{P}^\mu = \left[P_{ij}^\mu\right]_{i,j=1}^{|\mathcal{S}|}$: time-homogeneous transition probability matrix with $P_{ij}^\mu = p^\mu\left\{S_{t+1}^\mu = j \mid S_t^\mu = i\right\}$
- $\boldsymbol{\pi}^\mu = (\pi_i^\mu)_{i=1}^{|\mathcal{S}|}$: a unique stationary distribution of $\{S_t^\mu\}_{t\in\mathbb{N}}$
- $\boldsymbol{r}^\mu = [r\{1, \mu(1)\}, \cdots, r\{|\mathcal{S}|, \mu(|\mathcal{S}|)\}]^\top$: a reward vector under policy μ
- $\mathbb{E}^\mu$: expectation with respect to the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ with a fixed initial state S_0^μ
- $\mathbb{E}^{\boldsymbol{\pi}^\mu}$: expectation with respect to the stationary distribution of the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$
- $\|\cdot\|$: Euclidean norm for vectors and the induced operator norm for matrices
- $\lesssim$: inequality up to a constant
- $\boldsymbol{X}_t := \left(S_t^\mu, S_{t+1}^\mu, \boldsymbol{z}_t\right), \quad \boldsymbol{X}_{t-\tau:t} = (S_t^\mu, S_{t+1}^\mu, \boldsymbol{z}_{t-\tau:t})$
- $\boldsymbol{\phi}_l = \boldsymbol{\phi}(S_l^\mu), \quad \|\boldsymbol{\phi}_l\| \leq 1, \quad \boldsymbol{\Phi} \in \mathbb{R}^{|\mathcal{S}|\times d}$: feature matrix whose i^{th} row is $\boldsymbol{\phi}(i)^\top$
- $\boldsymbol{z}_t = \sum_{l=0}^t \lambda^{t-l}\boldsymbol{\phi}_l, \quad \boldsymbol{z}_{t-\tau:t} = \sum_{l=t-\tau}^t \lambda^{t-l}\boldsymbol{\phi}_l, \quad \omega^\mu = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\mu\left(\sum_{t=0}^{T-1} R_t^\mu\right) = \boldsymbol{\pi}^{\mu\top}\boldsymbol{r}^\mu$
- $\Pi_{\mathbb{O}}$: projection operator onto $\mathbb{O} := \mathbb{S}_{\boldsymbol{\Phi},\boldsymbol{e}}^\perp$ where $\mathbb{S}_{\boldsymbol{\Phi},\boldsymbol{e}} := \text{span}\{\boldsymbol{\theta} : \boldsymbol{\Phi}\boldsymbol{\theta} = \boldsymbol{e}\}, \quad \boldsymbol{\Pi} := \begin{bmatrix} 1 & 0 \\ 0 & \Pi_{\mathbb{O}} \end{bmatrix}$
- $\boldsymbol{A}_t = \boldsymbol{A}(\boldsymbol{X}_t) := \begin{bmatrix} -c_\alpha & 0 \\ -\boldsymbol{z}_t & \boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top) \end{bmatrix}, \quad \boldsymbol{A} := \mathbb{E}^{\pi^\mu}\boldsymbol{A}_t$
- $\boldsymbol{A}_{t-\tau:t} = \boldsymbol{A}(\boldsymbol{X}_{t-\tau:t}) := \begin{bmatrix} -c_\alpha & 0 \\ -\boldsymbol{z}_{t-\tau:t} & \boldsymbol{z}_{t-\tau:t}(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top) \end{bmatrix}$
- $\boldsymbol{b}_t = \boldsymbol{b}(\boldsymbol{X_t}) := \begin{bmatrix} c_\alpha R_t^\mu \\ R_t^\mu \boldsymbol{z}_t \end{bmatrix}, \quad \boldsymbol{b} = \mathbb{E}^{\boldsymbol{\pi}^\mu}\boldsymbol{b}_t, \quad \boldsymbol{b}_{t-\tau:t} = \boldsymbol{b}(\boldsymbol{X}_{t-\tau:t}) := \begin{bmatrix} c_\alpha R_t^\mu \\ R_t^\mu \boldsymbol{z}_{t-\tau:t} \end{bmatrix}$
- $\widehat{\boldsymbol{\Theta}}_t := \begin{bmatrix} \widehat{\omega}_t \\ \widehat{\boldsymbol{\theta}}_t \end{bmatrix}, \quad \boldsymbol{\Theta}^* = \begin{bmatrix} \omega^\mu \\ \boldsymbol{\theta}^* \end{bmatrix} \in \mathbb{R}\times\mathbb{O}$ is the unique element satisfying $\boldsymbol{A}\boldsymbol{\Theta}^* = \boldsymbol{b}$.

- $\boldsymbol{D}_t := \begin{bmatrix} \frac{1}{1+c_\alpha\beta_t} & 0 \\ 0 & \frac{1}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{I}_d \end{bmatrix}, \quad \underline{\boldsymbol{D}_t} := \gamma_t \boldsymbol{I}_{d+1}, \quad \gamma_t = \min\left(\frac{1}{1+c_\alpha\beta_t}, \frac{(1-\lambda)^2}{(1-\lambda)^2+\beta_t}\right)$
- $\zeta_t(\boldsymbol{\Theta}, \boldsymbol{X}_t) := \left\langle (\boldsymbol{A}(\boldsymbol{X}_t) - \boldsymbol{A})\,\boldsymbol{\Theta}^*, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\right\rangle + \left\langle \boldsymbol{b}(\boldsymbol{X}_t) - \boldsymbol{b}, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\right\rangle$
- $\xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_t) := (\boldsymbol{\Theta}^* - \boldsymbol{\Theta})^\top \left(\boldsymbol{A}(\boldsymbol{X}_t)^\top - \boldsymbol{A}^\top\right) \underline{\boldsymbol{D}_t}\,(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})$
- $\boldsymbol{M} = \operatorname{diag}\left(\pi^\mu_1, \cdots, \pi^\mu_{|\mathcal{S}|}\right), \quad \Delta := \min_{\|\boldsymbol{\theta}\|=1, \boldsymbol{\theta}\in\mathbb{O}} \boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top \boldsymbol{M}\left\{\boldsymbol{I} - (1-\lambda)\sum_{m=0}^\infty \lambda^m (\boldsymbol{P}^\mu)^{m+1}\right\}\boldsymbol{\Phi}\boldsymbol{\theta}$

B Theoretical results

Lemma B.1. *The implicit update rule is*

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\left(R^\mu_t - \widehat{\omega}_t + \boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}^\top_t\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t$$
$$\widehat{\omega}_{t+1} = \widehat{\omega}_t + \frac{c_\alpha\beta_t}{1+c_\alpha\beta_t}(R^\mu_t - \widehat{\omega}_t)$$

Proof. We first revisit the recursion formula in (2) and (3):

$$\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t(R^\mu_t - \widehat{\omega}_t + \boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t + \lambda\boldsymbol{z}^\top_{t-1}\widehat{\boldsymbol{\theta}}_t - \boldsymbol{z}^\top_t\widehat{\boldsymbol{\theta}}_{t+1})\boldsymbol{z}_t$$
$$\widehat{\omega}_{t+1} = \widehat{\omega}_t + c_\alpha\beta_t(R^\mu_t - \widehat{\omega}_{t+1})$$

Our goal is to derive the update rule for $\widehat{\boldsymbol{\theta}}$ and $\widehat{\omega}$, which can be done by combining $\widehat{\boldsymbol{\theta}}_{t+1}, \widehat{\omega}_{t+1}$.

(1) Update Rule for $\widehat{\boldsymbol{\theta}}$: Let us first examine the update rule for the parameter $\widehat{\boldsymbol{\theta}}$. Combining the rightmost term to the left, we have

$$\left(\boldsymbol{I} + \beta_t\boldsymbol{z}_t\boldsymbol{z}^\top_t\right)\widehat{\boldsymbol{\theta}}_{t+1} = \widehat{\boldsymbol{\theta}}_t + \beta_t\left(R^\mu_t - \widehat{\omega}_t + \boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t + \lambda\boldsymbol{z}^\top_{t-1}\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t. \tag{6}$$

From the Woodbury matrix identity $\left(\boldsymbol{I} + \beta_t\boldsymbol{z}_t\boldsymbol{z}^\top_t\right)^{-1} = \boldsymbol{I} - \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t\boldsymbol{z}^\top_t$, we have

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_{t+1} &= \widehat{\boldsymbol{\theta}}_t + \beta_t\left(R^\mu_t - \widehat{\omega}_t + \boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t + \lambda\boldsymbol{z}^\top_{t-1}\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t - \frac{\beta_t\boldsymbol{z}^\top_t\widehat{\boldsymbol{\theta}}_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t - \beta_t\frac{\beta_t R^\mu_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t + \beta_t\frac{\beta_t\widehat{\omega}_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t \\
&\quad - \beta_t\frac{\beta_t\boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t - \beta_t\frac{\beta_t\lambda\boldsymbol{z}^\top_{t-1}\widehat{\boldsymbol{\theta}}_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \beta_t R^\mu_t\left(1 - \frac{\beta_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\right)\boldsymbol{z}_t - \beta_t\widehat{\omega}_t\left(1 - \frac{\beta_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\right)\boldsymbol{z}_t \\
&\quad + \beta_t\boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t\left(1 - \frac{\beta_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\right)\boldsymbol{z}_t + \beta_t\lambda\boldsymbol{z}^\top_{t-1}\widehat{\boldsymbol{\theta}}_t\left(1 - \frac{\beta_t\|\boldsymbol{z}_t\|^2}{1+\beta_t\|\boldsymbol{z}_t\|^2}\right)\boldsymbol{z}_t - \frac{\beta_t\boldsymbol{z}^\top_t\widehat{\boldsymbol{\theta}}_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\boldsymbol{z}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\left(R^\mu_t - \widehat{\omega}_t + \boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t + \lambda\boldsymbol{z}^\top_{t-1}\widehat{\boldsymbol{\theta}}_t - \boldsymbol{z}^\top_t\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t \\
&= \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\left(R^\mu_t - \widehat{\omega}_t + \boldsymbol{\phi}^\top_{t+1}\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}^\top_t\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t
\end{aligned}$$

(2) Update Rule for $\widehat{\omega}$: Similarly, for the update rule for the $\widehat{\omega}$, we combine the term to the left hand side, which yields

$$(1 + c_\alpha\beta_t)\widehat{\omega}_{t+1} = \widehat{\omega}_t + c_\alpha\beta_t R^\mu_t.$$

Now, dividing with $(1 + c_\alpha\beta_t)$ we have

$$\widehat{\omega}_{t+1} = \frac{1}{1+c_\alpha\beta_t}\widehat{\omega}_t + \frac{c_\alpha\beta_t}{1+c_\alpha\beta_t}R_t^\mu = \widehat{\omega}_t + \frac{c_\alpha\beta_t}{1+c_\alpha\beta_t}(R_t^\mu - \widehat{\omega}_t).$$

This completes the whole update rules. □

B.1 Proof of Main Theorems

In this section, we provide a proof of the finite-time error bounds of the projected average-reward TD(λ) algorithm. To this end, recall that the implicit average-reward TD(λ) update rule is given by

$$\begin{aligned}
\widehat{\omega}_{t+1} &= \widehat{\omega}_t + \frac{c_\alpha\beta_t}{1+c_\alpha\beta_t}\left(R_t^\mu - \widehat{\omega}_t\right)\\
\widehat{\boldsymbol{\theta}}_{t+1} &= \widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\left(R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t.
\end{aligned}$$

To gain numerical stability as well as to facilitate theoretical analysis, we impose additional projection steps to both the primary iterate $\widehat{\boldsymbol{\theta}}_t$ and the reward iterate $\widehat{\omega}_t$. Recall that the projection operator with a radius $R > 0$ is given by

$$\Pi_R(\boldsymbol{u}) = \begin{cases} \boldsymbol{u}, & \text{if } \|\boldsymbol{u}\| \le R \\ \frac{R}{\|\boldsymbol{u}\|}\boldsymbol{u} & \text{otherwise.} \end{cases}$$

Then the projected average-reward TD(λ) update is as follows

$$\begin{aligned}
\widehat{\omega}_{t+1} &= \Pi_{R_\omega}\left\{\widehat{\omega}_t + \frac{c_\alpha\beta_t}{1+c_\alpha\beta_t}\left(R_t^\mu - \widehat{\omega}_t\right)\right\},\\
\widehat{\boldsymbol{\theta}}_{t+1} &= \Pi_{R_{\boldsymbol{\theta}}}\left\{\widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\left(R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t\right\},
\end{aligned}$$

where $R_\omega > 0$ and $R_{\boldsymbol{\theta}} > 0$ are large enough such that $R_\omega \ge \|\omega^\mu\|$ and $R_{\boldsymbol{\theta}} \ge \|\boldsymbol{\theta}^*\|$. Following the analysis of the average-reward TD(λ) in (Zhang et al., 2021b), we consider the following auxiliary iterates

$$\widehat{\omega}_{t+1} = \Pi_{R_\omega}\left\{\widehat{\omega}_t + \frac{c_\alpha\beta_t}{1+c_\alpha\beta_t}\left(R_t^\mu - \widehat{\omega}_t\right)\right\}, \tag{7}$$

$$\widehat{\boldsymbol{\theta}}_{t+1} = \Pi_{\mathbb{O}}\left[\Pi_{R_{\boldsymbol{\theta}}}\left\{\widehat{\boldsymbol{\theta}}_t + \frac{\beta_t}{1+\beta_t\|\boldsymbol{z}_t\|^2}\left(R_t^\mu - \widehat{\omega}_t + \boldsymbol{\phi}_{t+1}^\top\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\phi}_t^\top\widehat{\boldsymbol{\theta}}_t\right)\boldsymbol{z}_t\right\}\right], \tag{8}$$

to facilitate finite-time error analysis. Here, the space $\mathbb{O}$ is the orthogonal complement of the space generated by $\boldsymbol{\theta}_{\boldsymbol{e}}$ where $\boldsymbol{\Phi}\boldsymbol{\theta}_{\boldsymbol{e}} = \boldsymbol{e}$. In short, when considering the primary iterate, any deviance in the direction of $\boldsymbol{\theta}_{\boldsymbol{e}}$ will be ignored under $\Pi_{\mathbb{O}}$. Using the matrix notations we introduced, we can now succinctly write both (7) and (8) as

$$\widehat{\boldsymbol{\Theta}}_{t+1} = \boldsymbol{\Pi}\left[\Pi_{R_{\boldsymbol{\Theta}}}\left\{\widehat{\boldsymbol{\Theta}}_t + \beta_t\boldsymbol{D}_t\left(\boldsymbol{A}_{\boldsymbol{t}}\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}_t\right)\right\}\right],$$

where $\Pi_{R_{\boldsymbol{\Theta}}}(\widehat{\boldsymbol{\Theta}}) := \left[\Pi_{R_\omega}(\widehat{\omega}), \Pi_{R_{\boldsymbol{\theta}}}(\widehat{\boldsymbol{\theta}})\right]^\top$ for $\widehat{\boldsymbol{\Theta}} = [\widehat{\omega}, \widehat{\boldsymbol{\theta}}]^\top$, $R_{\boldsymbol{\Theta}} = \sqrt{R_\omega^2 + R_{\boldsymbol{\theta}}^2}$. Since $\boldsymbol{\Theta}^* = \boldsymbol{\Pi}\left(\Pi_{R_{\boldsymbol{\Theta}}}\boldsymbol{\Theta}^*\right)$, we have

$$\begin{aligned}
\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 &= \left\|\boldsymbol{\Pi}\left(\Pi_{R_{\boldsymbol{\Theta}}}\boldsymbol{\Theta}^*\right) - \boldsymbol{\Pi}\left[\Pi_{R_{\boldsymbol{\Theta}}}\left\{\widehat{\boldsymbol{\Theta}}_t + \beta_t\boldsymbol{D}_t\left(\boldsymbol{A}_t\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}_t\right)\right\}\right]\right\|^2\\
&\le \left\|\Pi_{R_{\boldsymbol{\Theta}}}\boldsymbol{\Theta}^* - \Pi_{R_{\boldsymbol{\Theta}}}\left\{\widehat{\boldsymbol{\Theta}}_t + \beta_t\boldsymbol{D}_t\left(\boldsymbol{A}_{\boldsymbol{t}}\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}_t\right)\right\}\right\|^2\\
&\le \left\|\boldsymbol{\Theta}^* - \left[\widehat{\boldsymbol{\Theta}}_t + \beta_t\boldsymbol{D}_t\left(\boldsymbol{A}_t\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}_t\right)\right]\right\|^2\\
&\le \left\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\right\|^2 \underbrace{-2\beta_t\left[\boldsymbol{D}_t\left(\boldsymbol{A}_t\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}_t\right)\right]^\top\left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\right)}_{(*)} + \underbrace{\beta_t^2\left\|\boldsymbol{D}_t\left(\boldsymbol{A}_t\widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b}_t\right)\right\|^2}_{(**)},
\end{aligned} \tag{9}$$

where the first inequality is due to non-expansiveness of the operator $\mathbf{\Pi}$ and the second inequality is due to non-expansiveness of the projection operator $\Pi_{R_{\boldsymbol{\Theta}}}$. We first obtain an upper bound of the expression in $(*)$. To this end, note that

$$\begin{aligned}
(*) &= -2\beta_t \left\{ \boldsymbol{D}_t \left(\boldsymbol{A}_t \widehat{\boldsymbol{\Theta}}_t - \boldsymbol{A}_t \boldsymbol{\Theta}^* + \boldsymbol{A}_t \boldsymbol{\Theta}^* - \boldsymbol{A} \boldsymbol{\Theta}^* + \boldsymbol{A} \boldsymbol{\Theta}^* + \boldsymbol{b}_t \right) \right\}^\top \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) \\
&= 2\beta_t (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}_t^\top \boldsymbol{D}_t \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) - 2\beta_t \left\{ (\boldsymbol{A}_t - \boldsymbol{A}) \boldsymbol{\Theta}^* + (\boldsymbol{b}_t - \boldsymbol{b}) \right\}^\top \boldsymbol{D}_t (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t) \\
&= 2\beta_t (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}_t^\top \underline{\boldsymbol{D}_t} \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) - 2\beta_t \left\{ (\boldsymbol{A}_t - \boldsymbol{A}) \boldsymbol{\Theta}^* + (\boldsymbol{b}_t - \boldsymbol{b}) \right\}^\top \underline{\boldsymbol{D}_t} (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t) \\
&\quad + 2\beta_t (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}_t^\top (\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}) \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) - 2\beta_t \left\{ (\boldsymbol{A}_t - \boldsymbol{A}) \boldsymbol{\Theta}^* + (\boldsymbol{b}_t - \boldsymbol{b}) \right\}^\top (\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}) (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)
\end{aligned}$$

where the second equality follows from $\boldsymbol{A}\boldsymbol{\Theta}^* = \boldsymbol{b}$. We bound each term in the last expression separately. For the first term, note that

$$\begin{aligned}
(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}_t^\top \underline{\boldsymbol{D}_t} \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) &= (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \left(\boldsymbol{A}_t^\top - \boldsymbol{A}^\top \right) \underline{\boldsymbol{D}_t} \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) + (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}^\top \underline{\boldsymbol{D}_t} \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) \\
&\leq \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right)^\top \left(\boldsymbol{A}_t^\top - \boldsymbol{A}^\top \right) \underline{\boldsymbol{D}_t} \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) - \frac{\Delta \gamma_t}{2} \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2,
\end{aligned}$$

where the first inequality is the direct consequence of Lemma 4.1, i.e.,

$$(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}^\top \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) \leq -\frac{\Delta}{2} \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2, \quad (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t) \in \mathbb{R} \times \mathbb{O},$$

as long as $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2 (1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$. Therefore, we obtain the following bound for the first term

$$2\beta_t (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}_t^\top \underline{\boldsymbol{D}_t} \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) \leq 2\beta_t \xi_t (\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) - \beta_t \gamma_t \Delta \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2 \tag{10}$$

which holds almost surely. For the second term, notice that

$$-2\beta_t \left\{ (\boldsymbol{A}_t - \boldsymbol{A}) \boldsymbol{\Theta}^* + (\boldsymbol{b}_t - \boldsymbol{b}) \right\}^\top \underline{\boldsymbol{D}_t} (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t) = -2\beta_t \zeta_t (\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t). \tag{11}$$

For the last two terms, applying Cauchy-Schwarz inequality with $\|\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}\| \leq \frac{(1+c_\alpha)\beta_t}{(1-\lambda)^2}$ (see Lemma C.5) gives us

$$2\beta_t (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t)^\top \boldsymbol{A}_t^\top (\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}) \left(\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t \right) \leq \frac{8 R_{\boldsymbol{\Theta}}^2 A_{\max} (1 + c_\alpha) \beta_t^2}{(1-\lambda)^2} \tag{12}$$

$$-2\beta_t \left\{ (\boldsymbol{A}_t - \boldsymbol{A}) \boldsymbol{\Theta}^* + (\boldsymbol{b}_t - \boldsymbol{b}) \right\}^\top (\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}) (\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t) \leq \frac{8 R_{\boldsymbol{\Theta}} (A_{\max} R_{\boldsymbol{\Theta}} + b_{\max}) (1 + c_\alpha) \beta_t^2}{(1-\lambda)^2} \tag{13}$$

where $A_{\max} := \sqrt{c_\alpha^2 + \frac{5}{(1-\lambda)^2}}$ and $b_{\max} := \sqrt{c_\alpha^2 + \frac{1}{(1-\lambda)^2}}$, which respectively serves as a uniform bound on $\|\boldsymbol{A}_t\|$ and $\|\boldsymbol{b}_t\|$. Combining (10), (11), (12) and (13), we get

$$(*) \leq -\beta_t \gamma_t \Delta \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2 + 2\beta_t \xi_t (\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) - 2\beta_t \zeta_t (\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) + G_1 \beta_t^2, \tag{14}$$

where $G_1 = \{8 R_{\boldsymbol{\Theta}} (2 A_{\max} R_{\boldsymbol{\Theta}} + b_{\max})(1 + c_\alpha)\}/(1-\lambda)^2$.

Next, we obtain an upper bound of the expression in $(**)$. Thanks to the fact $\|\boldsymbol{D}_t\| \leq 1$, we get

$$(**) \leq \beta_t^2 \|\boldsymbol{A_t} \widehat{\boldsymbol{\Theta}}_t + \boldsymbol{b_t}\|^2 \leq 2\beta_t^2 \left(A_{\max}^2 R_{\boldsymbol{\Theta}}^2 + b_{\max}^2 \right) =: G_2 \beta_t^2. \tag{15}$$

Combining (14) and (15), we have

$$\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 \leq (1 - \beta_t \gamma_t \Delta) \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2 + 2\beta_t \xi_t (\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) - 2\beta_t \zeta_t (\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) + G \beta_t^2,$$

with $G = G_1 + G_2$. Taking expectations of both sides, we have

$$\begin{aligned}
\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 &\leq (1-\beta_t\gamma_t\Delta)\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2 + 2\beta_t\mathbb{E}^\mu\xi_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) - 2\beta_t\mathbb{E}^\mu\zeta_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t) + G\beta_t^2 \\
&\leq (1-\beta_t\gamma_t\Delta)\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|^2 + 2\beta_t\left|\mathbb{E}^\mu\xi_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t)\right| + 2\beta_t\left|\mathbb{E}^\mu\zeta_t(\widehat{\boldsymbol{\Theta}}_t, \boldsymbol{X}_t)\right| + G\beta_t^2 \\
&\leq \left\{\prod_{i=0}^{t}(1-\beta_i\gamma_i\Delta)\right\}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \\
&\quad + 2\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1-\beta_k\gamma_k\Delta)\right\}\beta_i\left\{\left|\mathbb{E}^\mu\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| + \left|\mathbb{E}^\mu\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right|\right\} \\
&\quad + G\sum_{i=0}^{t}\left\{\prod_{k=i+1}^{t}(1-\beta_k\gamma_k\Delta)\right\}\beta_i^2 \qquad (16)
\end{aligned}$$

where we have used the identity $1 - \beta_i\gamma_i\Delta \in (0,1)$ for all $i \in \mathbb{N}$, which is true by the assumption $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2(1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$.

Theorem B.2. *Suppose the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic with a rate parameter $\rho \in (0,1)$ and the step-size ratio parameter is chosen to satisfy $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2(1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$. With $\lambda \in [0,1)$ and constant step-size $\beta_t = \beta$, the iterates of the projected average-reward implicit TD(λ) algorithm satisfy the following finite-time error bound*

$$\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 \lesssim (1-\beta\gamma\Delta)^{t+1}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 + \mathcal{O}\left(\beta\tau_\beta + h^{\tau_\beta} + \beta t\tau_\beta h^t\right), \quad t \geq 0$$

where $\gamma = \min\left\{\frac{1}{1+c_\alpha\beta}, \frac{(1-\lambda)^2}{(1-\lambda)^2+\beta}\right\}$ and $h = \max\{1-\beta\gamma\Delta,\ \rho,\ \lambda\}$.

Proof. Starting from (16), we have

$$\begin{aligned}
\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 &\leq (1-\beta\gamma\Delta)^{t+1}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \\
&\quad + 2\beta\sum_{i=0}^{t}(1-\beta\gamma\Delta)^{t-i}\left\{\left|\mathbb{E}^\mu\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| + \left|\mathbb{E}^\mu\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right|\right\} \\
&\quad + G\beta^2\sum_{i=0}^{t}(1-\beta\gamma\Delta)^{t-i} \\
&\lesssim (1-\beta\gamma\Delta)^{t+1}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \\
&\quad + \beta^2\tau_\beta\sum_{i=0}^{2\tau_\beta}(1-\beta\gamma\Delta)^{t-i} + \beta\tau_\beta\sum_{i=0}^{2\tau_\beta}(1-\beta\gamma\Delta)^{t-i}q^i \\
&\quad + \beta(\beta\tau_\beta + q^{\tau_\beta})\sum_{i=2\tau_\beta+1}^{t}(1-\beta\gamma\Delta)^{t-i} + \beta\tau_\beta\sum_{i=2\tau_\beta+1}^{t}(1-\beta\gamma\Delta)^{t-i}q^i \\
&\quad + \beta^2\sum_{i=0}^{t}(1-\beta\gamma\Delta)^{t-i} \\
&\lesssim (1-\beta\gamma\Delta)^{t+1}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 + \frac{\beta\tau_\beta}{\gamma\Delta} + \beta\tau_\beta^2 h^t + \frac{\beta\tau_\beta + q^{\tau_\beta}}{\gamma\Delta} + \beta\tau_\beta t h^t + \frac{\beta}{\gamma\Delta}
\end{aligned}$$

where in the second inequality, we used Lemma C.14 and Lemma C.15 with $q = \max\{\rho, \lambda\}$. In the last inequality, we used $h = \max\{1-\beta\gamma\Delta,\ \rho,\ \lambda\}$. Summarizing the terms, we get

$$\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 \lesssim (1-\beta\gamma\Delta)^{t+1}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 + \mathcal{O}\left(\beta\tau_\beta + h^{\tau_\beta} + \beta\tau_\beta t h^t\right).$$

□

Theorem B.3. *Suppose the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic with a rate parameter $\rho \in (0,1)$ and the step-size ratio parameter is chosen to satisfy $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2(1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$. With $\lambda \in [0,1)$ and decreasing step-sizes $\beta_t = \frac{\beta_0}{(t+1)^s}$, $s \in (0,1)$, the iterates of the projected average-reward implicit TD(λ) algorithm satisfy the following finite-time error bound, for $t \geq 0$,*

$$\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 \lesssim \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s}-1\}\right]\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 + \mathcal{O}\left\{\tau_{\beta_t} t\exp\left(-ct^{1-s}\right) + \tau_{\beta_t} t^{-s} + q^{\tau_{\beta_t}}\right\},$$

for some constant $c > 0$, $\gamma_0 = \min\left\{\frac{1}{1+c_\alpha\beta_0}, \frac{(1-\lambda)^2}{(1-\lambda)^2+\beta_0}\right\}$ and $q = \max\{\rho, \lambda\}$.

Proof. Starting from (16) with the identity $1 - \beta_i\gamma_i\Delta \leq \exp(-\beta_i\gamma_i\Delta)$ for all $i \in \mathbb{N}$, we have

$$\begin{aligned}
\mathbb{E}^\mu\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 &\leq \exp\left(-\Delta\sum_{i=0}^t \beta_i\gamma_i\right)\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \\
&\quad + 2\sum_{i=0}^t \exp\left(-\Delta\sum_{k=i+1}^t \beta_k\gamma_k\right)\beta_i\left|\mathbb{E}^\mu\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) + \mathbb{E}^\mu\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \\
&\quad + G\sum_{i=0}^t \exp\left(-\Delta\sum_{k=i+1}^t \beta_k\gamma_k\right)\beta_i^2 \\
&\leq \exp\left(-\Delta\gamma_0\sum_{i=0}^t \beta_i\right)\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \\
&\quad + 2\sum_{i=0}^t \exp\left(-\Delta\gamma_0\sum_{k=i+1}^t \beta_k\right)\beta_i\left|\mathbb{E}^\mu\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) + \mathbb{E}^\mu\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \\
&\quad + G\sum_{i=0}^t \exp\left(-\Delta\gamma_0\sum_{k=i+1}^t \beta_k\right)\beta_i^2
\end{aligned}$$

where the second inequality follows from the fact that γ_i is increasing. The final expression can be re-expressed as

$$\exp\left\{-\Delta\gamma_0\beta_0\sum_{i=0}^t\frac{1}{(i+1)^s}\right\}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \tag{17}$$

$$+ 2\sum_{i=0}^t \exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^t\frac{1}{(k+1)^s}\right\}\beta_i\left|\mathbb{E}^\mu\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) + \mathbb{E}^\mu\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \tag{18}$$

$$+ G\sum_{i=0}^t \exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^t\frac{1}{(k+1)^s}\right\}\beta_i^2. \tag{19}$$

The first term in (17) admits the following bound

$$\exp\left\{-\Delta\gamma_0\beta_0\sum_{i=0}^t\frac{1}{(i+1)^s}\right\}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \leq \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s}-1\}\right]\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2. \tag{20}$$

For the second term in (18), we first note that

$$
\begin{aligned}
&2\sum_{i=0}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)+\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)\right| \\
&\lesssim\sum_{i=0}^{2\tau_{\beta_t}}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i\left(\tau_{\beta_t}\beta_0+iq^i\right) \\
&\quad+\sum_{i=2\tau_{\beta_t}+1}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i\left(\tau_{\beta_t}\beta_{i-2\tau_{\beta_t}}+\tau_{\beta_t}q^i+q^{\tau_{\beta_t}}\right) \\
&\lesssim\tau_{\beta_t}\beta_0\sum_{i=0}^{2\tau_{\beta_t}}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i && (21)\\
&\quad+\tau_{\beta_t}\sum_{i=0}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i q^i && (22)\\
&\quad+\tau_{\beta_t}\sum_{i=2\tau_{\beta_t}+1}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i\beta_{i-2\tau_{\beta_t}} && (23)\\
&\quad+q^{\tau_{\beta_t}}\sum_{i=2\tau_{\beta_t}+1}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i && (24)
\end{aligned}
$$

where we used Lemma C.14 and Lemma C.15 in the first inequality. We first establish an upper bound on (21) using Lemma C.17.

$$
\begin{aligned}
\tau_{\beta_t}\beta_0\sum_{i=0}^{2\tau_{\beta_t}}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i &\le\frac{\tau_{\beta_t}\beta_0 e^{\Delta\gamma_0\beta_0}}{\Delta\gamma_0}\exp\left[-\frac{\Delta\gamma_0\beta_0}{(1-s)}\left\{(1+t)^{1-s}-(1+2\tau_{\beta_t})^{1-s}\right\}\right] \\
&\lesssim\tau_{\beta_t}\exp\left[-\frac{\Delta\gamma_0\beta_0}{(1-s)}\left\{(1+t)^{1-s}-(1+2\tau_{\beta_t})^{1-s}\right\}\right]
\end{aligned}
$$

Next, we obtain an upper bound on (22). To this end, consider

$$
\begin{aligned}
\tau_{\beta_t}\sum_{i=0}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i q^i &= \tau_{\beta_t}\sum_{i=0}^{\lfloor t/2\rfloor}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i q^i && (25)\\
&\quad+\tau_{\beta_t}\sum_{i=\lfloor t/2\rfloor+1}^{t}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=i+1}^{t}\frac{1}{(k+1)^s}\right\}\beta_i q^i. && (26)
\end{aligned}
$$

The term in (25) admits the following bound

$$
(25)\le\tau_{\beta_t}\beta_0\sum_{i=0}^{\lfloor t/2\rfloor}\exp\left\{-\Delta\gamma_0\beta_0\sum_{k=\lfloor t/2\rfloor+1}^{t}\frac{1}{(k+1)^s}\right\}.
$$

Since

$$
\sum_{k=\lfloor t/2\rfloor+1}^{t}\frac{1}{(k+1)^s}\ge\int_{\lfloor t/2\rfloor+1}^{t}x^{-s}dx\ge\frac{t^{1-s}-(t/2+1)^{1-s}}{1-s}=\Omega\left(t^{1-s}\right),
$$

we have

$$
(25)\lesssim\tau_{\beta_t}t\exp\left(-ct^{1-s}\right),
$$

for some constant $c>0$. For the term in (26), we have

$$
(26)\le\tau_{\beta_t}\sum_{i=\lfloor t/2\rfloor+1}^{t}\beta_i q^i\le\tau_{\beta_t}\sum_{i=\lfloor t/2\rfloor+1}^{t}\frac{\beta_0}{(t/2)^s}q^i\lesssim\frac{\tau_{\beta_t}}{t^s}.
$$

And hence, we get the bound for (22), given by

$$\tau_{\beta_t} \sum_{i=0}^{t} \exp\left\{-\Delta\gamma_0\beta_0 \sum_{k=i+1}^{t} \frac{1}{(k+1)^s}\right\} \beta_i q^i \lesssim \tau_{\beta_t} t \exp\left(-ct^{1-s}\right) + \frac{\tau_{\beta_t}}{t^s}.$$

Next, we obtain an upper bound on (23). From Lemma C.17,

$$\tau_{\beta_t} \sum_{i=2\tau_{\beta_t}+1}^{t} \exp\left\{-\Delta\gamma_0\beta_0 \sum_{k=i+1}^{t} \frac{1}{(k+1)^s}\right\} \beta_i \beta_{i-2\tau_{\beta_t}} \lesssim \tau_{\beta_t} \left(\exp\left[-\frac{\Delta\gamma_0\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}\right] + \beta_{t-2\tau_{\beta_t}}\right).$$

The only remaining term is the one in (24), whose bound is given by

$$q^{\tau_{\beta_t}} \sum_{i=2\tau_{\beta_t}+1}^{t} \exp\left\{-\Delta\gamma_0\beta_0 \sum_{k=i+1}^{t} \frac{1}{(k+1)^s}\right\} \beta_i \lesssim q^{\tau_{\beta_t}},$$

where we have used (53). Combining altogether to obtain a bound of (18), we get

$$\begin{aligned}
(18) &\lesssim \tau_{\beta_t} \exp\left[-\frac{\Delta\gamma_0\beta_0}{(1-s)}\left\{(1+t)^{1-s} - (1+2\tau_{\beta_t})^{1-s}\right\}\right] \\
&\quad + \tau_{\beta_t} t \exp\left(-ct^{1-s}\right) + \frac{\tau_{\beta_t}}{t^s} \\
&\quad + \tau_{\beta_t}\left(\exp\left[-\frac{\Delta\gamma_0\beta_0}{2(1-s)}\left\{(1+t)^{1-s}-1\right\}\right] + \beta_{t-2\tau_{\beta_t}}\right) + q^{\tau_{\beta_t}} \\
&\lesssim \tau_{\beta_t} t \exp\left(-ct^{1-s}\right) + \tau_{\beta_t} t^{-s} + q^{\tau_{\beta_t}},
\end{aligned} \tag{27}$$

for some constant $c > 0$.

Lastly, from Lemma C.16, the last term in (19) is upper bounded by

$$\begin{aligned}
G \sum_{i=0}^{t} \exp\left\{-\Delta\gamma_0\beta_0 \sum_{k=i+1}^{t} \frac{1}{(k+1)^s}\right\} \beta_i^2 &\lesssim \exp\left\{-\frac{\Delta\gamma_0}{2}\beta_0 \sum_{k=0}^{t} \frac{1}{(k+1)^s}\right\} + \beta_t \\
&\lesssim \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s}-1\}\right] + \beta_t
\end{aligned} \tag{28}$$

Combining (20), (27) and (28), we have

$$\begin{aligned}
\mathbb{E}^{\mu}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 &\lesssim \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s}-1\}\right] \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 \\
&\quad + \tau_{\beta_t} t \exp\left(-ct^{1-s}\right) + \tau_{\beta_t} t^{-s} + q^{\tau_{\beta_t}} \\
&\quad + \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s}-1\}\right] + \beta_t,
\end{aligned}$$

which can be further succinctly written as

$$\mathbb{E}^{\mu}\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_{t+1}\|^2 \leq \exp\left[-\frac{\Delta\gamma_0\beta_0}{1-s}\{(1+t)^{1-s}-1\}\right] \|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_0\|^2 + \mathcal{O}\left\{\tau_{\beta_t} t \exp\left(-ct^{1-s}\right) + \tau_{\beta_t} t^{-s} + q^{\tau_{\beta_t}}\right\}$$

for some constant $c > 0$. □

C Supporting lemmas

Our goal is to establish a finite-time error bound on $\|\boldsymbol{\Theta}^* - \widehat{\boldsymbol{\Theta}}_t\|$. To this end, we state the preliminary lemmas and provide their proofs. The first lemma establishes a norm bound on the eligibility trace.

Lemma C.1. *Given an exponential weighting parameter $\lambda \in [0,1)$, $\|\boldsymbol{z}_t\| \leq \frac{1}{1-\lambda}$ for all $t \in \mathbb{N}$.*

Proof. From the definition of the eligibility trace, $\boldsymbol{z}_t = \sum_{i=0}^{t} \lambda^{t-i} \boldsymbol{\phi_i}$, we observe $\|\boldsymbol{z}_t\| = \left\|\sum_{i=0}^{t} \lambda^{t-i} \boldsymbol{\phi_i}\right\| \leq \sum_{i=0}^{t} \lambda^{t-i} \leq \frac{1}{1-\lambda}$. The first inequality comes from the triangle inequality, given $\|\boldsymbol{\phi_i}\| \leq 1$. The latter is true by the sum of a geometric series. □

Next lemma provides explicit bounds on the error incurred by replacing the true, steady-state eligibility trace with its τ-step truncated version.

Lemma C.2. *Given an initial state $s_0 \in \mathcal{S}$, suppose $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic with a rate parameter $\rho \in (0,1)$. For all $t \in \mathbb{N}$, let $\tau \in \{0, \cdots, t\}$, then*

1. $\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) - \mathbb{E}^{\mu}(\boldsymbol{z}_{t-\tau:t})\right\| \lesssim \tau q^t + \lambda^\tau$
2. $\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{z}_t \boldsymbol{\phi}_t^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{z}_{t-\tau:t} \boldsymbol{\phi}_t^\top\right)\right\| \lesssim \tau q^t + \lambda^\tau$
3. $\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{z}_t \boldsymbol{\phi}_{t+1}^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{z}_{t-\tau:t} \boldsymbol{\phi}_{t+1}^\top\right)\right\| \lesssim \tau q^t + \lambda^\tau$

where $q := \max\{\lambda, \rho\}$.

Proof. We leverage the steady-state expression of the eligibility trace $\boldsymbol{z}_t = \sum_{l=-\infty}^{t} \lambda^{t-l} \boldsymbol{\phi}_l$ whenever the expectation is with respect to the stationary distribution induced by the policy μ.

Proof of the first statement:

$$
\begin{aligned}
\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) - \mathbb{E}^{\mu}(\boldsymbol{z}_{t-\tau:t}) &= \mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\sum_{l=-\infty}^{t} \lambda^{t-l} \boldsymbol{\phi}_l\right) - \mathbb{E}^{\mu}\left(\sum_{l=t-\tau}^{t} \lambda^{t-l} \boldsymbol{\phi}_l\right) \\
&= \sum_{l=-\infty}^{t} \lambda^{t-l} \mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{\phi}_l) - \sum_{l=t-\tau}^{t} \lambda^{t-l} \mathbb{E}^{\mu}(\boldsymbol{\phi}_l) \\
&= \sum_{l=t-\tau}^{t} \lambda^{t-l} \left\{\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{\phi}_l) - \mathbb{E}^{\mu}(\boldsymbol{\phi}_l)\right\} + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l} \mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{\phi}_l) \\
&= \sum_{l=t-\tau}^{t} \lambda^{t-l} \left[\sum_{s\in\mathcal{S}} \{\pi_s^\mu \boldsymbol{\phi}(s) - p^\mu(S_l^\mu = s | S_0^\mu = s_0)\boldsymbol{\phi}(s)\}\right] + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l} \mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{\phi}_l)
\end{aligned}
$$

Note that $\sup_{s\in\mathcal{S}} |\pi_s^\mu - p^\mu(S_l^\mu = s | S_0^\mu = s_0)| \lesssim \rho^l$ for some $\rho \in (0,1)$ follows from the geometric ergodicity of the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$. From the finiteness of $\mathcal{S}$ and normalized features, we have

$$
\begin{aligned}
\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) - \mathbb{E}^{\mu}(\boldsymbol{z}_{t-\tau:t})\right\| &\lesssim \sum_{l=t-\tau}^{t} \lambda^{t-l} \rho^l + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l} \\
&\lesssim \sum_{l=t-\tau}^{t} q^t + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l} \quad \text{where } q = \max\{\lambda, \rho\} \in (0,1) \\
&\lesssim \tau q^t + \lambda^\tau.
\end{aligned}
$$

Proof of the second statement:

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{z}_t\boldsymbol{\phi}_t^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{z}_{t-\tau:t}\boldsymbol{\phi}_t^\top\right) = \mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\sum_{l=-\infty}^{t}\lambda^{t-l}\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right) - \mathbb{E}^{\mu}\left(\sum_{l=t-\tau}^{t}\lambda^{t-l}\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right)\\
&= \sum_{l=-\infty}^{t}\lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right) - \sum_{l=t-\tau}^{t}\lambda^{t-l}\mathbb{E}^{\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right)\\
&= \sum_{l=t-\tau}^{t}\lambda^{t-l}\left\{\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right)\right\} + \sum_{l=-\infty}^{t-\tau-1}\lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_t^\top\right)\\
&= \sum_{l=t-\tau}^{t}\lambda^{t-l}\left[\sum_{s\in\mathcal{S}}\left\{\pi_s^\mu\boldsymbol{\phi}(s)\mathbb{E}^{\mu}\left(\boldsymbol{\phi}_t^\top|S_l^\mu=s\right) - p^\mu(S_l^\mu=s|S_0^\mu=s_0)\boldsymbol{\phi}(s)\mathbb{E}^{\mu}\left(\boldsymbol{\phi}_t^\top|S_l^\mu=s\right)\right\}\right]\\
&\quad + \sum_{l=-\infty}^{t-\tau-1}\lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_l^\top\right)
\end{aligned}$$

By the same logic as in the first statement, we have

$$\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{z}_t\boldsymbol{\phi}_t^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{z}_{t-\tau:t}\boldsymbol{\phi}_t^\top\right)\right\| \lesssim \sum_{l=t-\tau}^{t}\lambda^{t-l}\rho^l + \sum_{l=-\infty}^{t-\tau-1}\lambda^{t-l} \lesssim \tau q^t + \lambda^\tau,$$

where $q = \max\{\lambda, \rho\} \in (0,1)$.

Proof of the third statement:

$$\begin{aligned}
&\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{z}_t\boldsymbol{\phi}_{t+1}^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{z}_{t-\tau:t}\boldsymbol{\phi}_{t+1}^\top\right) = \mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\sum_{l=-\infty}^{t}\lambda^{t-l}\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right) - \mathbb{E}^{\mu}\left(\sum_{l=t-\tau}^{t}\lambda^{t-l}\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right)\\
&= \sum_{l=-\infty}^{t}\lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right) - \sum_{l=t-\tau}^{t}\lambda^{t-l}\mathbb{E}^{\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right)\\
&= \sum_{l=t-\tau}^{t}\lambda^{t-l}\left\{\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right)\right\} + \sum_{l=-\infty}^{t-\tau-1}\lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_{t+1}^\top\right)\\
&= \sum_{l=t-\tau}^{t}\lambda^{t-l}\left[\sum_{s\in\mathcal{S}}\left\{\pi_s^\mu\boldsymbol{\phi}(s)\mathbb{E}^{\mu}\left(\boldsymbol{\phi}_{t+1}^\top|S_l^\mu=s\right) - p^\mu\left(S_l^\mu=s|S_0^\mu=s_0\right)\boldsymbol{\phi}(s)\mathbb{E}^{\mu}\left(\boldsymbol{\phi}_{t+1}^\top|S_l^\mu=s\right)\right\}\right]\\
&\quad + \sum_{l=-\infty}^{t-\tau-1}\lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{\phi}_l\boldsymbol{\phi}_{l+1}^\top\right),
\end{aligned}$$

where in the last equality, we use the law of iterated expectations. Following the proof of the first statement, we have

$$\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(\boldsymbol{z}_t\boldsymbol{\phi}_{t+1}^\top\right) - \mathbb{E}^{\mu}\left(\boldsymbol{z}_{t-\tau:t}\boldsymbol{\phi}_{t+1}^\top\right)\right\| \lesssim \sum_{l=t-\tau}^{t}\lambda^{t-l}\rho^l + \sum_{l=-\infty}^{t-\tau-1}\lambda^{t-l} \lesssim \tau q^t + \lambda^\tau,$$

where $q = \max\{\lambda, \rho\} \in (0,1)$. □

Subsequent lemma provides explicit bounds on the error incurred by replacing the steady-state reward expectation with its non steady-state version.

Lemma C.3. *Given an initial state $s_0 \in \mathcal{S}$, suppose $\{S_t^\mu\}_{t\in\mathbb{N}}$ is uniformly geometrically ergodic with a rate parameter $\rho \in (0,1)$. For $\tau \in \{0, \cdots, t\}$,*

1. $|\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu) - \mathbb{E}^{\mu}(R_t^\mu)| \lesssim \rho^t$

2. $\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu \boldsymbol{z}_t) - \mathbb{E}^\mu(R_t^\mu \boldsymbol{z}_{t-\tau:t})\| \lesssim \tau q^t + \lambda^\tau$

where $q := \max\{\lambda, \rho\}$.

Proof. For the first statement, notice that

$$\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu) - \mathbb{E}^\mu(R_t^\mu) = \sum_{s\in\mathcal{S}} r\{s,\mu(s)\}\left\{\pi_s^\mu - p^\mu(S_t^\mu = s|S_0^\mu = s_0)\right\} \Longrightarrow |\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu) - \mathbb{E}^\mu(R_t^\mu)| \lesssim \rho^k,$$

where the last inequality follows from the uniform geometric ergodicity of the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$. For the second statement, we again leverage the steady-state expression of the eligibility trace $\boldsymbol{z}_t = \sum_{l=-\infty}^t \lambda^{t-l}\boldsymbol{\phi}_l$ whenever the expectation is with respect to the steady-state distribution induced by the policy μ.

$$\begin{aligned}
\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu \boldsymbol{z}_t) - \mathbb{E}^\mu(R_t^\mu \boldsymbol{z}_{t-\tau:t}) &= \mathbb{E}^{\boldsymbol{\pi}^\mu}\left(R_t^\mu \sum_{l=-\infty}^t \lambda^{t-l}\boldsymbol{\phi}_l\right) - \mathbb{E}^\mu\left(R_t^\mu \sum_{l=t-\tau}^t \lambda^{t-l}\boldsymbol{\phi}_l\right) \\
&= \sum_{l=-\infty}^t \lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(R_t^\mu \boldsymbol{\phi}_l\right) - \sum_{l=t-\tau}^t \lambda^{t-l}\mathbb{E}^\mu\left(R_t^\mu \boldsymbol{\phi}_l\right) \\
&= \sum_{l=t-\tau}^t \lambda^{t-l}\left\{\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(R_t^\mu \boldsymbol{\phi}_l\right) - \mathbb{E}^\mu\left(R_t^\mu \boldsymbol{\phi}_l\right)\right\} + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(R_t^\mu \boldsymbol{\phi}_l\right) \\
&= \sum_{l=t-\tau}^t \lambda^{t-l}\left[\sum_{s\in\mathcal{S}}\{\pi_s^\mu - p^\mu(S_l^\mu = s|S_0^\mu = s_0)\}\boldsymbol{\phi}(s)\mathbb{E}^\mu(R_t^\mu|S_l^\mu = s)\right] \\
&\quad + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l}\mathbb{E}^{\boldsymbol{\pi}^\mu}\left(R_t^\mu \boldsymbol{\phi}_l\right).
\end{aligned}$$

Taking the norm on both sides with the uniform geometric ergodicity of the Markov chain $\{S_t^\mu\}_{t\in\mathbb{N}}$, we have

$$\begin{aligned}
\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu \boldsymbol{z}_t) - \mathbb{E}^\mu(R_t^\mu \boldsymbol{z}_{t-\tau:t})\| &\lesssim \sum_{l=t-\tau}^t \lambda^{t-l}\rho^l + \sum_{l=-\infty}^{t-\tau-1} \lambda^{t-l} \\
&\lesssim \tau q^t + \lambda^\tau,
\end{aligned}$$

where the last inequality follows from $q = \max\{\lambda, \rho\}$. □

Lemma below establishes uniform bounds on the norm of $\boldsymbol{A}_t$ and $\boldsymbol{b}_t$, which will appear in the finite-time error bounds.

Lemma C.4. *For all* $t \in \mathbb{N}$, $\|\boldsymbol{A}_t\| \le A_{\max}$ *and* $\|\boldsymbol{b}_t\| \le b_{\max}$ *for some constants* $A_{\max}, b_{\max} \in \mathbb{R}_{>0}$

Proof.

$$\begin{aligned}
\|\boldsymbol{A}_t\|^2 &\le \|\boldsymbol{A}_t\|_F^2 \le c_\alpha^2 + \|\boldsymbol{z}_t\|^2 + \|\boldsymbol{z}_t\|^2\|\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top\|^2 \le c_\alpha^2 + \frac{5}{(1-\lambda)^2} =: A_{\max}^2 \\
\|\boldsymbol{b}_t\|^2 &\le c_\alpha^2 + \frac{1}{(1-\lambda)^2} =: b_{\max}^2,
\end{aligned}$$

where in the equality for $\|\boldsymbol{b}_t\|^2$, we used the fact $|R_t^\mu| \le 1$ for all $t \in \mathbb{N}$. □

Lemma C.5. *For all* $t \in \mathbb{N}$, $\|\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}\| \le \frac{(1+c_\alpha)\beta_t}{(1-\lambda)^2}$.

Proof. By the definition of the operator norm, we know

$$\|\boldsymbol{D}_t - \underline{\boldsymbol{D}_t}\| \le \max\left\{\left|\frac{1}{1+c_\alpha\beta_t} - \gamma_t\right|, \left|\frac{1}{1+\beta_t\|\boldsymbol{z}_t\|^2} - \gamma_t\right|\right\}.$$

Note that

$$\left|\frac{1}{1+c_\alpha\beta_t}-\gamma_t\right| \le \left|\frac{1}{1+c_\alpha\beta_t}-\frac{(1-\lambda)^2}{(1-\lambda)^2+\beta_t}\right| \le \left|\frac{\{1-(1-\lambda)^2c_\alpha\}\beta_t}{(1-\lambda)^2}\right| \le \frac{(1+c_\alpha)\beta_t}{(1-\lambda)^2}.$$

Since

$$\left|\frac{1}{1+\beta_t\|\boldsymbol{z}_t\|^2}-\frac{1}{1+c_\alpha\beta_t}\right| \le \left|\frac{c_\alpha\beta_t-\beta_t\|\boldsymbol{z}_t\|^2}{(1+\beta_t\|\boldsymbol{z}_t\|^2)(1+c_\alpha\beta_t)}\right| \le \left\{c_\alpha+\frac{1}{(1-\lambda)^2}\right\}\beta_t$$
$$\left|\frac{1}{1+\beta_t\|\boldsymbol{z}_t\|^2}-\frac{(1-\lambda)^2}{(1-\lambda)^2+\beta_t}\right| \le \left|\frac{\{1-(1-\lambda)^2\|\boldsymbol{z}_t\|^2\}\beta_t}{(1-\lambda)^2}\right| \le \frac{\beta_t}{(1-\lambda)^2}$$

we have

$$\left|\frac{1}{1+\beta_t\|\boldsymbol{z}_t\|^2}-\gamma_t\right| \le \frac{(1+c_\alpha)\beta_t}{(1-\lambda)^2},$$

which yields the desired bound. □

The following lemma establishes a uniform bound on the norm of the function ζ.

Lemma C.6. *For all $t\in\mathbb{N}$, $\boldsymbol{\Theta}\in\{\boldsymbol{\Xi}:\|\boldsymbol{\Xi}\|\le R_{\boldsymbol{\Theta}}\}$, $\|\zeta_t(\boldsymbol{\Theta},\boldsymbol{X}_t)\|\le C_\zeta$ for some constant $C_\zeta>0$.*

Proof. Note that $\|\underline{\boldsymbol{D}_t}\|\le 1$, $\|\boldsymbol{A}_t\|\le A_{\max}$ and $\|\boldsymbol{b}_t\|\le b_{\max}$,

$$\begin{aligned}\|\zeta_t(\boldsymbol{\Theta},\boldsymbol{X}_t)\| &\le \|\boldsymbol{A}_t-\boldsymbol{A}\|\,\|\boldsymbol{\Theta}^*\|\,\left\|\underline{\boldsymbol{D}_t}\right\|\,\|\boldsymbol{\Theta}^*-\boldsymbol{\Theta}\|+\|\boldsymbol{b}_t-\boldsymbol{b}\|\,\left\|\underline{\boldsymbol{D}_t}\right\|\,\|\boldsymbol{\Theta}^*-\boldsymbol{\Theta}\|\\ &\le 2\,\|\boldsymbol{A}_t-\boldsymbol{A}\|\,R_{\boldsymbol{\Theta}}^2+2\,\|\boldsymbol{b}_t-\boldsymbol{b}\|\,R_{\boldsymbol{\Theta}}\\ &\le 4A_{\max}R_{\boldsymbol{\Theta}}^2+4b_{\max}R_{\boldsymbol{\Theta}}=:C_\zeta\end{aligned}$$

□

Two lemmas below respectively establish Lipschitzness of the function ζ with respect to $\boldsymbol{\Theta}$ component and the deviance bound with respect to $\boldsymbol{X}$ component.

Lemma C.7. *For all $t\in\mathbb{N}$, $\boldsymbol{\Theta_1},\boldsymbol{\Theta_2}\in\{\boldsymbol{\Xi}:\|\boldsymbol{\Xi}\|\le R_{\boldsymbol{\Theta}}\}$,*

$$|\zeta_t(\boldsymbol{\Theta_1},\boldsymbol{X}_t)-\zeta_t(\boldsymbol{\Theta_2},\boldsymbol{X}_t)|\le L_\zeta\|\boldsymbol{\Theta_1}-\boldsymbol{\Theta_2}\|,$$

for some constant $L_\zeta>0$.

Proof.

$$\begin{aligned}|\zeta_t(\boldsymbol{\Theta_1},\boldsymbol{X}_t)-\zeta_t(\boldsymbol{\Theta_2},\boldsymbol{X}_t)| &= \left\langle(\boldsymbol{A}_t-\boldsymbol{A})\,\boldsymbol{\Theta}^*,\underline{\boldsymbol{D}_t}(\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1})\right\rangle+\left\langle\boldsymbol{b}_t-\boldsymbol{b},\underline{\boldsymbol{D}_t}(\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1})\right\rangle\\ &\le \|\boldsymbol{A}_t-\boldsymbol{A}\|\,\|\boldsymbol{\Theta}^*\|\|\underline{\boldsymbol{D}_t}\|\|\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1}\|+\|\boldsymbol{b}_t-\boldsymbol{b}\|\,\|\underline{\boldsymbol{D}_t}\|\|\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1}\|\\ &\le \|\boldsymbol{A}_t-\boldsymbol{A}\|\,R_{\boldsymbol{\Theta}}\|\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1}\|+\|\boldsymbol{b}_t-\boldsymbol{b}\|\,\|\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1}\|\\ &\le (2A_{\max}R_{\boldsymbol{\Theta}}+2b_{\max})\|\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1}\|\\ &=: L_\zeta\|\boldsymbol{\Theta_2}-\boldsymbol{\Theta_1}\|\end{aligned}$$

where in the first inequality, we used the Cauchy-Schwarz inequality. The second and third inequalities follow from $\|\underline{\boldsymbol{D}_t}\|\le 1$, $\|\boldsymbol{A}_t\|\le A_{\max}$ and $\|\boldsymbol{b}_t\|\le b_{\max}$. □

Lemma C.8. *For all $t\in\mathbb{N}$, $\boldsymbol{\Theta}\in\{\boldsymbol{\Xi}:\|\boldsymbol{\Xi}\|\le R_{\boldsymbol{\Theta}}\}$, let $\tau\in\{0,\cdots,t\}$, then we have*

$$|\zeta_t(\boldsymbol{\Theta},\boldsymbol{X}_t)-\zeta_t(\boldsymbol{\Theta},\boldsymbol{X}_{t-\tau:t})|\lesssim\lambda^\tau.$$

Proof. Note that

$$\begin{aligned}|\zeta_t(\boldsymbol{\Theta},\boldsymbol{X}_t)-\zeta_t(\boldsymbol{\Theta},\boldsymbol{X}_{t-\tau:t})| &= \left\langle(\boldsymbol{A}_t-\boldsymbol{A}_{t-\tau:t})\,\boldsymbol{\Theta}^*,\underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^*-\boldsymbol{\Theta})\right\rangle+\left\langle\boldsymbol{b}_t-\boldsymbol{b}_{t-\tau:t},\underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^*-\boldsymbol{\Theta})\right\rangle\\ &\le 2\,\|\boldsymbol{A}_t-\boldsymbol{A}_{t-\tau:t}\|\,R_{\boldsymbol{\Theta}}^2+2\,\|\boldsymbol{b}_t-\boldsymbol{b}_{t-\tau:t}\|\,R_{\boldsymbol{\Theta}}.\end{aligned}\tag{29}$$

where we used $\|\boldsymbol{\Theta}^*\| \leq R_{\boldsymbol{\Theta}}$, $\|\boldsymbol{\Theta}\| \leq R_{\boldsymbol{\Theta}}$, $\|\underline{\boldsymbol{D}_t}\| \leq 1$. To obtain a bound on $\|\boldsymbol{A}_t - \boldsymbol{A}_{t-\tau:t}\|$ and $\|\boldsymbol{b}_t - \boldsymbol{b}_{t-\tau:t}\|$, note that

$$\|\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t}\| \leq \sum_{l=0}^{t-\tau} \lambda^{t-l} \leq \lambda^\tau/(1-\lambda).$$

Now consider

$$\boldsymbol{A}_t - \boldsymbol{A}_{t-\tau:t} = \begin{bmatrix} 0 & 0 \\ -(\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t}) & (\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t})(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top), \end{bmatrix}$$

whose operator norm satisfies the following bound

$$\|\boldsymbol{A}_t - \boldsymbol{A}_{t-\tau:t}\|^2 \leq \|\boldsymbol{A}_t - \boldsymbol{A}_{t-\tau:t}\|_F^2 \leq \|\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t}\|^2 + \|\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t}\|^2 \left\|\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top\right\|^2 \leq 5\lambda^{2\tau}/(1-\lambda)^2. \tag{30}$$

Similarly, consider

$$\boldsymbol{b}_t - \boldsymbol{b}_{t-\tau:t} = \begin{bmatrix} 0 \\ R_t^\mu(\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t}) \end{bmatrix},$$

for which the Euclidean norm is bounded as follows

$$\|\boldsymbol{b}_t - \boldsymbol{b}_{t-\tau:t}\| \leq \|\boldsymbol{z}_t - \boldsymbol{z}_{t-\tau:t}\| \leq \lambda^\tau/(1-\lambda). \tag{31}$$

Plugging (30) and (31) into (29), we have

$$|\zeta_t(\boldsymbol{\Theta}, \boldsymbol{X}_t) - \zeta_t(\boldsymbol{\Theta}, \boldsymbol{X}_{t-\tau:t})| \lesssim \lambda^\tau.$$

□

As we did for the function ζ_t, we next establish boundedness, Lipschitzness of the function ξ_t with respect to $\boldsymbol{\Theta}$ as well as the deviance bound with respect to $\boldsymbol{X}$ component.

Lemma C.9. *For all $t \in \mathbb{N}$, $\boldsymbol{\Theta} \in \{\boldsymbol{\Xi} : \|\boldsymbol{\Xi}\| \leq R_{\boldsymbol{\Theta}}\}$, $|\xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_t)| \leq C_\xi$ for some constant $C_\xi > 0$.*

Proof.

$$\begin{aligned} |\xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_t)| &= \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\,(\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\right| \\ &\leq \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\|\, \|\boldsymbol{A}_t - \boldsymbol{A}\|\, \left\|\underline{\boldsymbol{D}_t}\right\|\, \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\| \\ &\leq 8R_{\boldsymbol{\Theta}}^2 A_{\max} =: C_\xi \end{aligned}$$

where in the first inequality, we used Cauchy-Schwarz inequality and in the second inequality, we used $\|\underline{\boldsymbol{D}_t}\| \leq 1$, $\|\boldsymbol{A}_t\| \leq A_{\max}$ and $\|\boldsymbol{A}\| \leq A_{\max}$. □

Two lemmas below respectively establish Lipschitzness of the function ξ_t with respect to $\boldsymbol{\Theta}$ component and the deviance bound with respect to $\boldsymbol{X}$ component.

Lemma C.10. *For all $t \in \mathbb{N}$, $\boldsymbol{\Theta_1}, \boldsymbol{\Theta_2} \in \{\boldsymbol{\Xi} : \|\boldsymbol{\Xi}\| \leq R_{\boldsymbol{\Theta}}\}$,*

$$|\xi_t(\boldsymbol{\Theta_1}, \boldsymbol{X}_t) - \xi_t(\boldsymbol{\Theta_2}, \boldsymbol{X}_t)| \leq L_\xi \|\boldsymbol{\Theta_1} - \boldsymbol{\Theta_2}\|,$$

for some constant $L_\xi > 0$.

Proof.

$$\begin{aligned} |\xi_t(\boldsymbol{\Theta_1}, \boldsymbol{X}_t) - \xi_t(\boldsymbol{\Theta_2}, \boldsymbol{X}_t)| &= \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1}) - (\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})\right| \\ &\leq \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1}) - (\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1})\right| \\ &\quad + \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1}) - (\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})\right| \\ &= \left|(\boldsymbol{\Theta_2} - \boldsymbol{\Theta_1})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1})\right| + \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top\, \underline{\boldsymbol{D}_t}(\boldsymbol{\Theta_2} - \boldsymbol{\Theta_1})\right| \\ &\leq \|\boldsymbol{\Theta_2} - \boldsymbol{\Theta_1}\|\, \|\boldsymbol{A}_t - \boldsymbol{A}\|\, \|\underline{\boldsymbol{D}_t}\|\, \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta_1}\| + \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta_2}\|\, \|\boldsymbol{A}_t - \boldsymbol{A}\|\, \|\underline{\boldsymbol{D}_t}\|\, \|\boldsymbol{\Theta_2} - \boldsymbol{\Theta_1}\| \\ &\leq 4R_{\boldsymbol{\Theta}} A_{\max}\, \|\boldsymbol{\Theta_2} - \boldsymbol{\Theta_1}\| \\ &=: L_\xi\, \|\boldsymbol{\Theta_1} - \boldsymbol{\Theta_2}\| \end{aligned}$$

where in the first inequality, we used the triangle inequality. The rest of the inequalities follow from Cauchy-Schwarz inequality, $\|\underline{\boldsymbol{D}_t}\| \leq 1$, $\|\boldsymbol{A}_t\| \leq A_{\max}$ and $\|\boldsymbol{A}\| \leq A_{\max}$. □

Lemma C.11. *For all $t \in \mathbb{N}$, $\boldsymbol{\Theta} \in \{\boldsymbol{\Xi} : \|\boldsymbol{\Xi}\| \leq R_{\boldsymbol{\Theta}}\}$, let $\tau \in \{0, \cdots, t\}$, then we have*

$$|\xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_t) - \xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_{t-\tau:t})| \lesssim \lambda^\tau.$$

Proof. Note that

$$\begin{aligned}
|\xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_t) - \xi_t(\boldsymbol{\Theta}, \boldsymbol{X}_{t-\tau:t})| &= \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})^\top (\boldsymbol{A}_t - \boldsymbol{A})^\top \underline{\boldsymbol{D}_t} (\boldsymbol{\Theta}^* - \boldsymbol{\Theta}) - (\boldsymbol{\Theta}^* - \boldsymbol{\Theta})^\top (\boldsymbol{A}_{t-\tau:t} - \boldsymbol{A})^\top \underline{\boldsymbol{D}_t} (\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\right| \\
&= \left|(\boldsymbol{\Theta}^* - \boldsymbol{\Theta})^\top (\boldsymbol{A}_t - \boldsymbol{A}_{t-\tau:t})^\top \underline{\boldsymbol{D}_t} (\boldsymbol{\Theta}^* - \boldsymbol{\Theta})\right| \\
&\leq \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\| \, \|\boldsymbol{A}_t - \boldsymbol{A}_{t-\tau:t}\| \, \|\underline{\boldsymbol{D}_t}\| \, \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\| \\
&\lesssim \lambda^\tau
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz. The last inequality follows from (30) with the fact $\|\underline{\boldsymbol{D}_t}\| \leq 1$. □

The next two lemmas will serve as key ingredients in establishing bounds on the non-steady state expectations of ζ_t and ξ_t terms.

Lemma C.12. *For all $t \in \mathbb{N}$, let $\tau \in \{0, \cdots, t\}$, then we have*

$$\|\mathbb{E}^\mu(\boldsymbol{A}_{t-\tau:t}) - \boldsymbol{A}\| \lesssim \tau q^t + \lambda^\tau$$

where $q := \max\{\lambda, \rho\}$.

Proof. Note that

$$\begin{aligned}
\mathbb{E}^\mu(\boldsymbol{A}_{t-\tau:t}) - \boldsymbol{A} &= \begin{bmatrix} -c_\alpha & 0 \\ -\mathbb{E}^\mu(\boldsymbol{z}_{t-\tau:t}) & \mathbb{E}^\mu\{\boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top)\} \end{bmatrix} - \begin{bmatrix} -c_\alpha & 0 \\ -\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) & \mathbb{E}^{\boldsymbol{\pi}^\mu}\{\boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top)\} \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0 \\ \mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) - \mathbb{E}^\mu(\boldsymbol{z}_{t-\tau:t}) & \mathbb{E}^\mu\{\boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top)\} - \mathbb{E}^{\boldsymbol{\pi}^\mu}\{\boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top)\} \end{bmatrix}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\mathbb{E}^\mu(\boldsymbol{A}_{t-\tau:t}) - \boldsymbol{A}\|^2 \\
&\leq \|\mathbb{E}^\mu(\boldsymbol{A}_{t-\tau:t}) - \boldsymbol{A}\|_F^2 \\
&\leq \left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) - \mathbb{E}^\mu(\boldsymbol{z}_{t-\tau:t})\right\|^2 + \left\|\mathbb{E}^\mu\{\boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top)\} - \mathbb{E}^{\boldsymbol{\pi}^\mu}\{\boldsymbol{z}_t(\boldsymbol{\phi}_{t+1}^\top - \boldsymbol{\phi}_t^\top)\}\right\|^2 \\
&\leq \left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t) - \mathbb{E}^\mu(\boldsymbol{z}_{t-\tau:t})\right\|^2 + 2\left\|\mathbb{E}^\mu(\boldsymbol{z}_t\boldsymbol{\phi}_{t+1}^\top) - \mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t\boldsymbol{\phi}_{t+1}^\top)\right\|^2 + 2\left\|\mathbb{E}^{\boldsymbol{\pi}^\mu}(\boldsymbol{z}_t\boldsymbol{\phi}_t^\top) - \mathbb{E}^\mu(\boldsymbol{z}_t\boldsymbol{\phi}_t^\top)\right\|^2 \\
&\lesssim \left(\tau q^t + \lambda^\tau\right)^2,
\end{aligned}$$

where the last line follows from Lemma C.2. □

Lemma C.13. *For all $t \in \mathbb{N}$, let $\tau \in \{0, \cdots, t\}$, then we have*

$$\|\mathbb{E}^\mu(\boldsymbol{b}_{t-\tau:t}) - \boldsymbol{b}\| \lesssim \tau q^t + \lambda^\tau,$$

where $q = \max\{\lambda, \rho\}$.

Proof. Note that

$$\mathbb{E}^\mu(\boldsymbol{b}_{t-\tau:t}) - \boldsymbol{b} = \begin{bmatrix} c_\alpha \mathbb{E}^\mu(R_t^\mu) - c_\alpha \mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu) \\ \mathbb{E}^\mu(R_t^\mu \boldsymbol{z}_{t-\tau:t}) - \mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu \boldsymbol{z}_t) \end{bmatrix}$$

Therefore,

$$\begin{aligned}\|\mathbb{E}^{\mu}\left(\boldsymbol{b}_{t-\tau:t}\right)-\boldsymbol{b}\|^2 &\leq c_\alpha^2\left|\mathbb{E}^{\mu}(R_t)-\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu)\right|^2+\left\|\mathbb{E}^{\mu}(R_t\boldsymbol{z}_{t-\tau:t})-\mathbb{E}^{\boldsymbol{\pi}^\mu}(R_t^\mu\boldsymbol{z}_t)\right\|^2\\ &\lesssim c_\alpha^2\rho^{2t}+\left(\tau q^t+\lambda^\tau\right)^2\\ &\lesssim\left(\tau q^t+\lambda^\tau\right)^2\end{aligned}$$

where the second inequality follows from Lemma C.3. □

We now establish bounds on the expectation of ζ_t.

Lemma C.14. *Suppose $(\beta_t)_{t\in\mathbb{N}}$ is a non-increasing sequence and $q=\max\{\lambda,\rho\}$. Given $t\in\mathbb{N}$, suppose $i>2\tau_{\beta_t}$ then,*

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)\right|\lesssim\tau_{\beta_t}\beta_{i-2\tau_{\beta_t}}+\tau_{\beta_t}q^i+q^{\tau_{\beta_t}}.$$

Otherwise,

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)\right|\lesssim\tau_{\beta_t}\beta_0+iq^i.$$

Proof. We begin by considering the case $i>2\tau_{\beta_t}$. From the triangle inequality, we have

$$\begin{aligned}\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)\right| &\leq\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)-\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_i)\right| &(32)\\ &+\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_i)-\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_{i-\tau_{\beta_t}:i})\right| &(33)\\ &+\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_{i-\tau_{\beta_t}:i})\right| &(34)\end{aligned}$$

To obtain an upper bound of (32), note that

$$\begin{aligned}&\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)-\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_i)\right|\\ &\leq\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)-\mathbb{E}^{\mu}\zeta_i(\boldsymbol{\Theta}_{i-1},\boldsymbol{X}_i)\right|+\cdots+\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}+1},\boldsymbol{X}_i)-\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_i)\right|\\ &\leq L_\zeta\sum_{l=i-2\tau_{\beta_t}}^{i-1}\left\|\widehat{\boldsymbol{\Theta}}_{l+1}-\widehat{\boldsymbol{\Theta}}_l\right\| &(35)\end{aligned}$$

where we used Lemma C.7 in the second inequality. For $\left\|\widehat{\boldsymbol{\Theta}}_{l+1}-\widehat{\boldsymbol{\Theta}}_l\right\|$, we observe the following inequality

$$\begin{aligned}\left\|\widehat{\boldsymbol{\Theta}}_{l+1}-\widehat{\boldsymbol{\Theta}}_l\right\| &=\left\|\boldsymbol{\Pi}\left[\Pi_{R_{\boldsymbol{\Theta}}}\left\{\widehat{\boldsymbol{\Theta}}_l+\beta_l\underline{\boldsymbol{D}_l}\left(\boldsymbol{A}_l\widehat{\boldsymbol{\Theta}}_l+\boldsymbol{b}_l\right)\right\}\right]-\boldsymbol{\Pi}\left(\Pi_{R_{\boldsymbol{\Theta}}}\widehat{\boldsymbol{\Theta}}_l\right)\right\|\\ &\leq\left\|\Pi_{R_{\boldsymbol{\Theta}}}\left\{\widehat{\boldsymbol{\Theta}}_l+\beta_l\underline{\boldsymbol{D}_l}\left(\boldsymbol{A}_l\widehat{\boldsymbol{\Theta}}_l+\boldsymbol{b}_l\right)\right\}-\Pi_{R_{\boldsymbol{\Theta}}}\widehat{\boldsymbol{\Theta}}_l\right\|\\ &\leq\left\|\widehat{\boldsymbol{\Theta}}_l+\beta_l\underline{\boldsymbol{D}_l}\left\{\boldsymbol{A}_l\widehat{\boldsymbol{\Theta}}_l+\boldsymbol{b}_l\right\}-\widehat{\boldsymbol{\Theta}}_l\right\|\\ &\leq\beta_l\left\|\boldsymbol{A}_l\widehat{\boldsymbol{\Theta}}_l+\boldsymbol{b}_l\right\|\\ &\leq\beta_l(A_{\max}R_{\boldsymbol{\Theta}}+b_{\max}). &(36)\end{aligned}$$

The first and second inequalities follow from the non-expansiveness of the projection operators, while the third inequality follows from the bound $\|\underline{\boldsymbol{D}_l}\|\leq 1$. Plugging (36) back to (35), we have

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i,\boldsymbol{X}_i)-\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}},\boldsymbol{X}_i)\right|\lesssim\sum_{l=i-2\tau_{\beta_t}}^{i-1}\beta_l \tag{37}$$

Next, by applying Lemma C.8 together with Jensen's inequality, we immediately observe the upper bound of the term in (33), namely,

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_{i-\tau_{\beta_t}:i})\right| \lesssim \lambda^{\tau_{\beta_t}}. \tag{38}$$

We now obtain an upper bound of the term in (34). To this end, let us set

$$f_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{Y}_{i-\tau_{\beta_t}:i}) := \zeta_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_{i-\tau_{\beta_t}:i}),$$

where $\boldsymbol{Y}_{i-\tau_{\beta_t}:i} = \left(S^{\mu}_{i-\tau_{\beta_t}}, S^{\mu}_{i-\tau_{\beta_t}+1}, \cdots, S^{\mu}_{i-1}, \boldsymbol{X}_i\right)$. We further define $\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y'}_{i-\tau_{\beta_t}:i}$ as random variables drawn independently from the marginal distributions of $\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y}_{i-\tau_{\beta_t}:i}$ respectively. Since

$$\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}} \to S^{\mu}_{i-2\tau_{\beta_t}} \to S^{\mu}_{i-\tau_{\beta_t}} \to S^{\mu}_{i} \to \boldsymbol{X}_i = (S^{\mu}_i, S^{\mu}_{i+1}, \boldsymbol{z}_i)$$

forms a Markov chain, an application of Lemma 9 in (Bhandari et al., 2018) results in

$$\left|\mathbb{E}^{\mu} f_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{Y}_{i-\tau_{\beta_t}:i})\right| \lesssim \left|\mathbb{E} f_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i})\right| + \rho^{\tau_{\beta_t}},$$

for all $i > 2\tau_{\beta_t}$. Since $\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y'}_{i-\tau_{\beta_t}:i}$ are independent to each other, we get

$$\begin{aligned}\mathbb{E} f_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i}) &= \boldsymbol{\Theta^*}^{\top}\mathbb{E}^{\mu}\left(\boldsymbol{A}_{i-\tau_{\beta_t}:i} - \boldsymbol{A}\right)^{\top}\underline{\boldsymbol{D}_i}\mathbb{E}^{\mu}\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right) \\ &\quad + \mathbb{E}^{\mu}\left(\boldsymbol{b}_{i-\tau_{\beta_t}:i} - \boldsymbol{b}\right)^{\top}\underline{\boldsymbol{D}_i}\mathbb{E}^{\mu}\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right).\end{aligned}$$

From the Cauchy-Schwarz inequality coupled with the Jensen's inequality, we get

$$\left|\mathbb{E} f_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i})\right| \le 2\left\|\mathbb{E}^{\mu}\left(\boldsymbol{A}_{i-\tau_{\beta_t}:i} - \boldsymbol{A}\right)\right\| R^2_{\boldsymbol{\Theta}} + 2\left\|\mathbb{E}^{\mu}\left(\boldsymbol{b}_{i-\tau_{\beta_t}:i} - \boldsymbol{b}\right)\right\| R_{\boldsymbol{\Theta}} \lesssim \tau_{\beta_t} q^i + \lambda^{\tau_{\beta_t}}$$

where the second inequality is due to Lemma C.12 and Lemma C.13. Therefore, we get

$$\left|\mathbb{E}^{\mu} f_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{Y}_{i-\tau_{\beta_t}:i})\right| \lesssim \tau_{\beta_t} q^i + \lambda^{\tau_{\beta_t}} + \rho^{\tau_{\beta_t}} \lesssim \tau_{\beta_t} q^i + q^{\tau_{\beta_t}}, \tag{39}$$

with $q = \max\{\lambda, \rho\}$. Combining (37), (38) and (39), we get

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \lesssim \sum_{l=i-2\tau_{\beta_t}}^{i-1} \beta_l + \tau_{\beta_t} q^i + q^{\tau_{\beta_t}} \lesssim \tau_{\beta_t}\beta_{i-2\tau_{\beta_t}} + \tau_{\beta_t} q^i + q^{\tau_{\beta_t}}.$$

Next we consider the case $i \le 2\tau_{\beta_t}$. From the triangle inequality, we have

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \le \left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right| + \left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right|.$$

From Lemma C.7 combined with (36), we have

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right| \lesssim \sum_{l=0}^{i-1}\beta_l \lesssim \tau_{\beta_t}\beta_0. \tag{40}$$

For the second term, since $\widehat{\boldsymbol{\Theta}}_0$ is deterministic,

$$\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i) = \left\langle \mathbb{E}^{\mu}\left(\boldsymbol{A}_i - \boldsymbol{A}\right)\boldsymbol{\Theta^*}, \underline{\boldsymbol{D}_i}(\boldsymbol{\Theta^*} - \widehat{\boldsymbol{\Theta}}_0)\right\rangle + \left\langle \mathbb{E}^{\mu}\left(\boldsymbol{b} - \boldsymbol{b}_i\right), \underline{\boldsymbol{D}_i}(\boldsymbol{\Theta^*} - \widehat{\boldsymbol{\Theta}}_0)\right\rangle,$$

and therefore

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right| \le 2\left\|\mathbb{E}^{\mu}\left(\boldsymbol{A}_i - \boldsymbol{A}\right)\right\| R^2_{\boldsymbol{\Theta}} + 2\left\|\mathbb{E}^{\mu}\left(\boldsymbol{b} - \boldsymbol{b}_i\right)\right\| R_{\boldsymbol{\Theta}} \lesssim iq^i + \lambda^i \lesssim iq^i \tag{41}$$

where the second inequality follows from Lemma C.12 and the last inequality is by the definition $q := \max\{\lambda, \rho\} \in (0, 1)$. Combining (40) and (41), we get

$$\left|\mathbb{E}^{\mu}\zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \lesssim \tau_{\beta_t}\beta_0 + iq^i.$$

□

Lemma C.15. *Suppose $(\beta_t)_{t\in\mathbb{N}}$ is a non-increasing sequence and $q = \max\{\lambda, \rho\}$. Given $t \in \mathbb{N}$, suppose $i > 2\tau_{\beta_t}$ then,*

$$\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \lesssim \tau_{\beta_t}\beta_{i-2\tau_{\beta_t}} + \tau_{\beta_t}q^i + q^{\tau_{\beta_t}}.$$

Otherwise,

$$\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \lesssim \tau_{\beta_t}\beta_0 + iq^i.$$

Proof. We begin by considering the case $i > 2\tau_{\beta_t}$. Again by the triangle inequality, we have

$$\begin{aligned}
\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| &\leq \left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i)\right| && (42)\\
&+ \left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_{i-\tau_{\beta_t}:i})\right| && (43)\\
&+ \left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_{i-\tau_{\beta_t}:i})\right| && (44)
\end{aligned}$$

To obtain an upper bound of (42), note that

$$\begin{aligned}
\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i)\right| &\leq \left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\boldsymbol{\Theta}_{i-1}, \boldsymbol{X}_i)\right| + \cdots\\
&+ \left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}+1}, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i)\right|\\
&\leq L_{\xi}\sum_{l=i-2\tau_{\beta_t}}^{i-1}\left\|\widehat{\boldsymbol{\Theta}}_{l+1} - \widehat{\boldsymbol{\Theta}}_l\right\| && (45)
\end{aligned}$$

where the second inequality is due to Lemma C.10. Recall from (36) that

$$\left\|\widehat{\boldsymbol{\Theta}}_{l+1} - \widehat{\boldsymbol{\Theta}}_l\right\| \leq \beta_l(A_{\max}R_{\boldsymbol{\Theta}} + b_{\max}).$$

Plugging (36) back to (45), we have

$$\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i)\right| \lesssim \sum_{l=i-2\tau_{\beta_t}}^{i-1}\beta_l \tag{46}$$

Next, by applying Lemma C.11 together with Jensen's inequality, we immediately obtain the upper bound of the term in (43), namely,

$$\left|\mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_i) - \mathbb{E}^{\mu}\xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_{i-\tau_{\beta_t}:i})\right| \lesssim \lambda^{\tau_{\beta_t}}. \tag{47}$$

We now obtain an upper bound of the term in (44). To this end, let us set

$$g_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{Y}_{i-\tau_{\beta_t}:i}) := \xi_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{X}_{i-\tau_{\beta_t}:i}),$$

where $\boldsymbol{Y}_{i-\tau_{\beta_t}:i} = \left(S^{\mu}_{i-\tau_{\beta_t}}, S^{\mu}_{i-\tau_{\beta_t}+1}, \cdots, S^{\mu}_{i-1}, \boldsymbol{X}_i\right)$. We further define $\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y'}_{i-\tau_{\beta_t}:i}$ as random variables drawn independently from the marginal distributions of $\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y}_{i-\tau_{\beta_t}:i}$ respectively. Since

$$\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}} \to S^{\mu}_{i-2\tau_{\beta_t}} \to S^{\mu}_{i-\tau_{\beta_t}} \to S^{\mu}_i \to \boldsymbol{X}_i = (S^{\mu}_i, S^{\mu}_{i+1}, \boldsymbol{z}_i)$$

forms a Markov chain, an application of Lemma 9 in (Bhandari et al., 2018) results in

$$\left|\mathbb{E}^{\mu}g_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{Y}_{i-\tau_{\beta_t}:i})\right| \lesssim \left|\mathbb{E}g_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i})\right| + \rho^{\tau_{\beta_t}},$$

for all $i > 2\tau_{\beta_t}$. Since $\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y'}_{i-\tau_{\beta_t}:i}$ are independent to each other, we get

$$\begin{aligned}
\mathbb{E}g_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i}) &= \mathbb{E}\left[\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top \left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i} - \boldsymbol{A}\right)^\top \underline{\boldsymbol{D}_i} \left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\right] \\
&= \mathbb{E}\left[\text{Trace}\left\{\left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i} - \boldsymbol{A}\right)^\top \underline{\boldsymbol{D}_i} \left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\}\right] \\
&= \text{Trace}\left[\mathbb{E}\left\{\left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i} - \boldsymbol{A}\right)^\top \underline{\boldsymbol{D}_i} \left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\}\right] \\
&= \text{Trace}\left[\mathbb{E}^\mu\left\{\left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i} - \boldsymbol{A}\right)^\top\right\} \underline{\boldsymbol{D}_i}\mathbb{E}^\mu\left\{\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\}\right].
\end{aligned}$$

where the last equality comes from the independence between $\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}$ and $\boldsymbol{Y'}_{i-\tau_{\beta_t}:i}$. By the Von-Neumann's trace inequality (see Theorem 7.4.1.1 of (Horn and Johnson, 2012)) with a nuclear norm notation $\|\cdot\|_*$, we have

$$\begin{aligned}
\left|\mathbb{E}g_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i})\right| &\le \left\|\mathbb{E}^\mu\left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i}\right) - \boldsymbol{A}\right\| \left\|\underline{\boldsymbol{D}_i}\mathbb{E}^\mu\left\{\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\}\right\|_* \\
&\le \left\|\mathbb{E}^\mu\left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i}\right) - \boldsymbol{A}\right\| \left\|\underline{\boldsymbol{D}_i}\right\| \left\|\mathbb{E}^\mu\left\{\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\}\right\|_*
\end{aligned}$$

where the second inequality is due to an identity $\|\boldsymbol{AB}\|_* \le \|\boldsymbol{A}\|\|\boldsymbol{B}\|_*$. Furthermore, notice that

$$\begin{aligned}
\left\|\mathbb{E}^\mu\left\{\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\}\right\|_* &\le \mathbb{E}^\mu\left\{\left\|\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)\left(\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right)^\top\right\|_*\right\} \\
&= \mathbb{E}^\mu\left\{\left\|\boldsymbol{\Theta^*} - \boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}\right\|^2\right\} \\
&\le 4R_{\boldsymbol{\Theta}}^2
\end{aligned}$$

where the first inequality is due to Jensen's inequality. Therefore, we arrive at

$$\left|\mathbb{E}^\mu g_i(\boldsymbol{\Theta'}_{i-2\tau_{\beta_t}}, \boldsymbol{Y'}_{i-\tau_{\beta_t}:i})\right| \lesssim \left\|\mathbb{E}^\mu\left(\boldsymbol{A'}_{i-\tau_{\beta_t}:i}\right) - \boldsymbol{A}\right\| \lesssim \tau_{\beta_t} q^i + \lambda^{\tau_{\beta_t}}$$

where the last inequality follows from Lemma C.12. This then gives us

$$\left|\mathbb{E}^\mu g_i(\widehat{\boldsymbol{\Theta}}_{i-2\tau_{\beta_t}}, \boldsymbol{Y}_{i-\tau_{\beta_t}:i})\right| \lesssim \tau_{\beta_t} q^i + \lambda^{\tau_{\beta_t}} + \rho^{\tau_{\beta_t}} \lesssim \tau_{\beta_t} q^i + q^{\tau_{\beta_t}}, \tag{48}$$

for $q = \max\{\lambda, \rho\}$. Combining (46), (47) and (48), we get

$$\left|\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \lesssim \sum_{l=i-2\tau_{\beta_t}}^{i-1} \beta_l + \tau_{\beta_t} q^i + q^{\tau_{\beta_t}} \lesssim \tau_{\beta_t}\beta_{i-2\tau_{\beta_t}} + \tau_{\beta_t} q^i + q^{\tau_{\beta_t}}.$$

Next we consider the case $i \le 2\tau_{\beta_t}$. From the triangle inequality, we have

$$\left|\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \le \left|\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right| + \left|\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right|.$$

From Lemma C.10 combined with (36), we have

$$\left|\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i) - \mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right| \lesssim \sum_{l=0}^{i-1} \beta_l \lesssim \tau_{\beta_t}\beta_0. \tag{49}$$

For the second term, since $\widehat{\boldsymbol{\Theta}}_0$ is deterministic,

$$\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i) = (\boldsymbol{\Theta^*} - \widehat{\boldsymbol{\Theta}}_0)^\top \mathbb{E}^\mu\left(\boldsymbol{A_i} - \boldsymbol{A}\right)^\top \underline{\boldsymbol{D}_i}(\boldsymbol{\Theta^*} - \widehat{\boldsymbol{\Theta}}_0)$$

and therefore

$$\left|\mathbb{E}^\mu \xi_i(\widehat{\boldsymbol{\Theta}}_0, \boldsymbol{X}_i)\right| \le 4R_{\boldsymbol{\Theta}}^2 \left\|\mathbb{E}\boldsymbol{A}_i - \boldsymbol{A}\right\| \lesssim iq^i + \lambda^i \lesssim iq^i \tag{50}$$

where the second inequality follows from Lemma C.12 and in the last inequality we used the definition $q := \max\{\lambda, \rho\} \in (0, 1)$. Combining (49) and (50), we get

$$\left|\mathbb{E}^\mu \zeta_i(\widehat{\boldsymbol{\Theta}}_i, \boldsymbol{X}_i)\right| \lesssim \tau_{\beta_t}\beta_0 + iq^i.$$

□

Lemma C.16. *For $t \in \mathbb{N}$, let $\beta_t = \frac{\beta_0}{(t+1)^s}$ and $s \in (0,1)$. With $\gamma > 0$,*

$$\sum_{i=0}^{t} \left(e^{-\gamma \sum_{k=i+1}^{t} \beta_k}\right) \beta_i^2 \leq 2\left(K_b e^{-\frac{\gamma}{2}\sum_{k=0}^{t}\beta_k} + \beta_t\right) \frac{e^{\frac{\gamma\beta_0}{2}}}{\gamma},$$

where $K_b = \beta_0 e^{\frac{\gamma}{2}\sum_{k=0}^{i_0}\beta_k}$ for some $i_0 \in \mathbb{N}$.

Proof. Let $T_t = \sum_{i=0}^{t-1}\beta_i$ and use the convention $\sum_{k=t+1}^{t}\beta_k = 0$ and $\sum_{k=t+1}^{t}\beta_k^2 = 0$. Notice that

$$\begin{aligned}
\sum_{i=0}^{t} \left(e^{-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k}\right)\beta_i &\leq \left(\sup_{i\geq 0} e^{\frac{\gamma}{2}\beta_i}\right)\left\{\sum_{i=0}^{t}\left(e^{-\frac{\gamma}{2}\sum_{k=i}^{t}\beta_k}\right)\beta_i\right\} \\
&= \left(\sup_{i\geq 0} e^{\frac{\gamma}{2}\beta_i}\right)\left\{\sum_{i=0}^{t}\left(e^{-\frac{\gamma}{2}(T_{t+1}-T_i)}\right)\beta_i\right\} \\
&\leq \left(\sup_{i\geq 0} e^{\frac{\gamma}{2}\beta_i}\right)\int_0^{T_{t+1}} e^{-\frac{\gamma}{2}(T_{t+1}-s)}ds \\
&\leq \left(\sup_{i\geq 0} e^{\frac{\gamma}{2}\beta_i}\right)\frac{2}{\gamma} \leq \frac{2e^{\frac{\gamma\beta_0}{2}}}{\gamma},
\end{aligned} \tag{51}$$

where we have used the definition of the left-Riemann sum in the first inequality. The last inequality is due to the fact that $\{\beta_t\}$ is a non-increasing sequence. Now consider

$$\begin{aligned}
\sum_{i=0}^{t}\left(e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\right)\beta_i^2 &\leq \sup_{0\leq i\leq t}\left(\beta_i e^{-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k}\right)\left\{\sum_{i=0}^{t}\left(e^{-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k}\right)\beta_i\right\} \\
&\leq \sup_{0\leq i\leq t}\left(\beta_i e^{-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k}\right)\frac{2e^{\frac{\gamma\beta_0}{2}}}{\gamma}
\end{aligned} \tag{52}$$

where the last inequality follows from (51). Note that $\beta_i e^{-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k}$ is eventually increasing, i.e., after some time $i_0 \in \mathbb{N}$, for all $t \geq i_0$, we have

$$\sup_{i_0\leq i\leq t}\left\{\beta_i \exp\left(-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k\right)\right\} \leq \beta_t,$$

where we used the convention $\sum_{k=t+1}^{t}\beta_k = 0$. Therefore, we have

$$\begin{aligned}
(52) &\leq \left\{\sup_{0\leq i\leq i_0}\left(\beta_i e^{-\frac{\gamma}{2}\sum_{k=i+1}^{t}\beta_k}\right) + \beta_t\right\}\frac{2e^{\frac{\gamma\beta_0}{2}}}{\gamma} \\
&\leq \left\{e^{-\frac{\gamma}{2}\sum_{k=0}^{t}\beta_k}\sup_{0\leq i\leq i_0}\left(\beta_i e^{\frac{\gamma}{2}\sum_{k=0}^{i}\beta_k}\right) + \beta_t\right\}\frac{2e^{\frac{\gamma\beta_0}{2}}}{\gamma} \\
&\leq \left(K_b e^{-\frac{\gamma}{2}\sum_{k=0}^{t}\beta_k} + \beta_t\right)\frac{2e^{\frac{\gamma\beta_0}{2}}}{\gamma},
\end{aligned}$$

where $K_b = \beta_0 e^{\frac{\gamma}{2}\sum_{k=0}^{i_0}\beta_k}$. □

Lemma C.17. *For $t \in \mathbb{N}$, let $\beta_t = \frac{\beta_0}{(t+1)^s}$ and $s \in (0,1)$. With $\gamma > 0$ and $\tau \in \{0, \cdots, t\}$,*

1. $\sum_{i=0}^{\tau} e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\beta_i \leq \frac{e^{\gamma\beta_0}}{\gamma}e^{-\frac{\gamma\beta_0}{1-s}\left\{(1+t)^{1-s}-(1+\tau)^{1-s}\right\}}$.
2. $\sum_{i=2\tau_{\beta_t}+1}^{t}\left(e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\beta_i\right) \leq \left[e^{\frac{-\gamma\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}}D_\beta \mathbb{I}_{\left(2\tau_{\beta_t}+1<i_\beta\right)} + \beta_{t-2\tau_{\beta_t}}\right]\frac{2e^{\gamma\beta_0/2}}{\gamma}$

where $D_\beta = \exp\{(\gamma/2)\sum_{k=0}^{i_\beta}\beta_k\}\beta_0$ *for some* $i_\beta \in \mathbb{N}$.

Proof. Let $T_t = \sum_{i=0}^{t-1}\beta_i$ and use the convention $\sum_{k=t+1}^{t}\beta_k = 0$. For the first statement,

$$\begin{aligned}\sum_{i=0}^{\tau} e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\beta_i &\le \max_{i\ge 0}\left(e^{\gamma\beta_i}\right)\sum_{i=0}^{\tau} e^{-\gamma\sum_{k=i}^{t}\beta_k}\beta_i = e^{\gamma\beta_0}\sum_{i=0}^{\tau} e^{-\gamma(T_{t+1}-T_i)}\beta_i \\ &\le e^{\gamma\beta_0}\int_0^{T_{\tau+1}} e^{-\gamma(T_{t+1}-s)}ds \le \frac{e^{\gamma\beta_0}}{\gamma}e^{-\gamma(T_{t+1}-T_{\tau+1})} \\ &= \frac{e^{\gamma\beta_0}}{\gamma}e^{-\gamma\beta_0\sum_{k=\tau+1}^{t}1/(1+k)^s} \le \frac{e^{\gamma\beta_0}}{\gamma}e^{-\frac{\gamma\beta_0}{1-s}\left\{(1+t)^{1-s}-(1+\tau)^{1-s}\right\}}.\end{aligned}$$

For the second statement, first notice that

$$\begin{aligned}\sum_{i=2\tau_{\beta_t}+1}^{t} e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\beta_i &\le \max_{i\ge 0}\left(e^{\gamma\beta_i}\right)\sum_{i=2\tau_{\beta_t}+1}^{t} e^{-\gamma\sum_{k=i}^{t}\beta_k}\beta_i = e^{\gamma\beta_0}\sum_{i=2\tau_{\beta_t}+1}^{t} e^{-\gamma(T_{t+1}-T_i)}\beta_i \\ &\le e^{\gamma\beta_0}\int_{T_{2\tau_{\beta_t}+1}}^{T_{t+1}} e^{-\gamma(T_{t+1}-s)}ds = \frac{e^{\gamma\beta_0}}{\gamma}\left\{1-e^{-\gamma\left(T_{t+1}-T_{2\tau_{\beta_t}+1}\right)}\right\} \le \frac{e^{\gamma\beta_0}}{\gamma}. \qquad (53)\end{aligned}$$

Then, we have

$$\begin{aligned}\sum_{i=2\tau_{\beta_t}+1}^{t} e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\beta_i &\le \max_{i\in\left[2\tau_{\beta_t}+1,t\right]}\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\}\sum_{i=2\tau_{\beta_t}+1}^{t} e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_i \\ &\le \max_{i\in\left[2\tau_{\beta_t}+1,t\right]}\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\}\frac{2e^{\gamma\beta_0/2}}{\gamma} \qquad (54)\end{aligned}$$

where the second inequality follows from (53). To bound the first term in (54), note that the sequence $\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\}_{i\in\mathbb{N}}$ is eventually increasing. In other words, there exists $i_\beta \in \mathbb{N}$ such that,

$$\max_{i\in\left[2\tau_{\beta_t}+1,t\right]}\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\} = \beta_{t-2\tau_{\beta_t}} \quad \text{if} \quad 2\tau_{\beta_t}+1 \ge i_\beta.$$

If $2\tau_{\beta_t}+1 < i_\beta$, then

$$\begin{aligned}&\max_{i\in\left[2\tau_{\beta_t}+1,t\right]}\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\} \\ &\le \max_{i\in[2\tau_\beta+1,i_\beta]}\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\} + \max_{i\in[i_\beta+1,t]}\left\{e^{(-\gamma/2)\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\} \\ &\le e^{(-\gamma/2)\sum_{k=0}^{t}\beta_k}\max_{i\in\left[2\tau_{\beta_t}+1,i_\beta\right]}\left\{e^{(\gamma/2)\sum_{k=0}^{i}\beta_k}\beta_{i-2\tau_{\beta_t}}\right\} + \beta_{t-2\tau_{\beta_t}} \\ &\le e^{-(\gamma/2)\sum_{k=0}^{t}\beta_k}e^{(\gamma/2)\sum_{k=0}^{i_\beta}\beta_k}\beta_0 + \beta_{t-2\tau_{\beta_t}} \\ &\le e^{\frac{-\gamma\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}}D_\beta + \beta_{t-2\tau_{\beta_t}}\end{aligned}$$

where $D_\beta = e^{(\gamma/2)\sum_{k=0}^{i_\beta}\beta_k}\beta_0$. Combining everything, we get

$$\sum_{i=2\tau_{\beta_t}+1}^{t} e^{-\gamma\sum_{k=i+1}^{t}\beta_k}\beta_{i-2\tau_{\beta_t}}\beta_i \le \left[e^{\frac{-\gamma\beta_0}{2(1-s)}\left\{(t+1)^{1-s}-1\right\}}D_\beta\mathbb{I}_{\left(2\tau_{\beta_t}+1<i_\beta\right)} + \beta_{t-2\tau_{\beta_t}}\right]\frac{2e^{\gamma\beta_0/2}}{\gamma}.$$

□

D Additional Experimental Details

We provide detailed explanations of the experimental setups for both evaluation and control experiments in this supplementary results section. For each run, we set the exponential weight parameter $\lambda = 0.25$ and the step-size ratio parameter $c_\alpha = 1.0$. Under this hyperparameter configuration, we evaluate four methods: (i) average-reward TD(λ), (ii) average-reward implicit TD(λ) without projection, and (iii–iv) average-reward implicit TD(λ) with projection, using parameter radius $R_{\boldsymbol{\Theta}} \in \{1000,\ 5000\}$. For the projection of the average-reward estimate, we fix the radius $R_\omega = 1$, which safely bounds the true average-reward since $\omega^\mu \in [-1,\ 1]$ by construction in all our settings.

D.1 Evaluation Experiments

Computing true parameters. We compute the oracle quantities used in the loss $(\widehat{\omega}_t - \omega^\mu)^2 + \|\Pi_{\mathbb{O}}(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*)\|^2$ as follows. Given a transition matrix $\boldsymbol{P}^\mu$, reward vector $\boldsymbol{r}^\mu$, and feature matrix $\boldsymbol{\Phi}$, we first obtain the stationary distribution $\boldsymbol{\pi}^\mu$ satisfying ${\boldsymbol{\pi}^\mu}^\top \boldsymbol{P}^\mu = {\boldsymbol{\pi}^\mu}^\top$ and ${\boldsymbol{\pi}^\mu}^\top \boldsymbol{e} = 1$. The average reward is then $\omega^\mu = {\boldsymbol{\pi}^\mu}^\top \boldsymbol{r}^\mu$. In addition, to compute the optimal weight vector $\boldsymbol{\theta}^*$, we first solve for the basic differential value $\boldsymbol{v}^\mu$, which is the unique solution to

$$(\boldsymbol{I} - \boldsymbol{P}^\mu)\boldsymbol{v}^\mu = \boldsymbol{r}^\mu - \omega^\mu \boldsymbol{e}, \qquad {\boldsymbol{\pi}^\mu}^\top \boldsymbol{v}^\mu = 0.$$

Because $\boldsymbol{e}$ and $\boldsymbol{v}^\mu$ are included as columns of $\boldsymbol{\Phi}$, which is of full rank, the optimal weight vector $\boldsymbol{\theta}^*$ can be obtained from solving the linear system $\boldsymbol{\Phi}\boldsymbol{\theta}^* = \boldsymbol{v}^\mu$. In an analogous fashion, to compute the projection to the space $\mathbb{O}$, we solve for $\boldsymbol{\theta_e}$ satisfying $\boldsymbol{\Phi}\boldsymbol{\theta_e} = \boldsymbol{e}$. Recall that $\boldsymbol{\theta_e}$ defines the projection direction that removes the constant component of the error. The projection operator to the space $\mathbb{O}$, can be expressed as $\boldsymbol{I} - \frac{\boldsymbol{\theta_e}{\boldsymbol{\theta_e}}^\top}{\|\boldsymbol{\theta_e}\|^2}$, i.e., $\Pi_{\mathbb{O}}(\boldsymbol{\theta}) = \left(\boldsymbol{I} - \frac{\boldsymbol{\theta_e}{\boldsymbol{\theta_e}}^\top}{\|\boldsymbol{\theta_e}\|^2}\right)\boldsymbol{\theta}$.

D.1.1 MRP

Experiment setup. We describe the construction of transition probabilities, rewards, and the feature matrix following (Zhang et al., 2021b); details are reproduced here for completeness.

- **Transition probabilities:** For each state s, we generated a probability distribution over the $|\mathcal{S}| = 100$ states by drawing $(|\mathcal{S}|-1)$ i.i.d samples from Unif$[0, 1]$. We then sorted them, and took successive differences. The final entry was set to ensure the components sum to one.
- **Rewards:** Each state s received a reward sampled independently from Unif$[0, 1]$.
- **Feature matrix:** Let d denote the feature dimension. We draw $\widetilde{\boldsymbol{\Phi}} \in \mathbb{R}^{|\mathcal{S}|\times(d-2)}$ with i.i.d. Bernoulli(0.5) entries. We then appended the all-ones vector $\boldsymbol{e}$ and the basic differential value $\boldsymbol{v}^\mu$ as columns to form

  $$\boldsymbol{\Phi} = \left[\widetilde{\boldsymbol{\Phi}}\ \ \boldsymbol{e}\ \ \boldsymbol{v}^\mu\right].$$

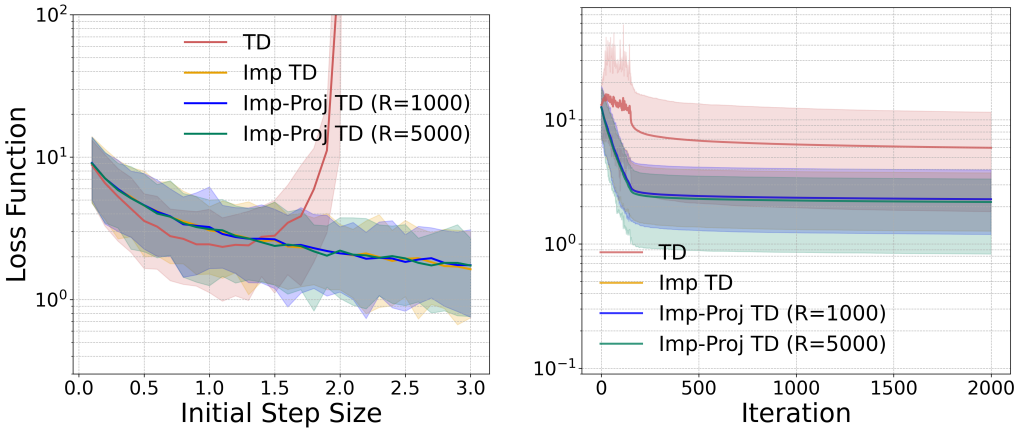
  If needed, we repeated the sampling until $\boldsymbol{\Phi}$ had full column rank, then was row-normalized so that $\|\boldsymbol{\phi}(s)\| \leq 1$ for all $s \in \mathcal{S}$.

Additional results. We report additional MRP results under a decaying step-size schedule. For each schedule, we run 50 independent trials, initialize $\boldsymbol{\theta}_0$ by sampling each coordinate from Unif$[-1, 1]$, and set the initial average-reward estimate $\widehat{\omega}_0 = 0$. Complementing the constant step-size results in the main text, Figure 5 shows performance under $\beta_t = \beta_0/(t+1)^{0.99}$ for initial step-sizes $\beta_0 \in \{0.1, 0.2, \ldots, 3.0\}$ using the same hyperparameters ($\lambda = 0.25$, $c_\alpha = 1.0$). Solid lines denote the mean loss across runs and shaded regions indicate 95% confidence intervals. As β_0 increases, average-reward implicit TD(λ) methods keep the loss controlled, whereas average-reward TD(λ) diverges for step-sizes larger than 2.0. For $\beta_t = 1.8/(t+1)^{0.99}$, the full loss trajectory (right panel) further underscores the gap: average-reward TD(λ) remains markedly worse than its implicit counterparts.

D.1.2 Boyan Chain

Experiment setup. We describe the construction of the transition probabilities, reward function, and feature matrix for the average-reward Boyan chain. The original Boyan chain was introduced by (Boyan, 2002) and later

Figure 5: MRP experiment results under decaying step-size schedule $\beta_t = \beta_0/(t+1)^{0.99}$, with exponential weighting parameter and step-size ratio set to $(\lambda, c_\alpha) = (0.25, 1.0)$. Solid lines denote the mean, and shaded regions represent 95% confidence intervals. (Left) Loss value for initial step-sizes from 0.1 to 3.0. (Right) Full trajectory of the loss value with initial step-size $\beta_0 = 1.8$.



adapted to the average-reward setting by (Zhang et al., 2021a). The chain consists of 13 states and two actions, denoted by $\{s_0, s_1, \ldots, s_{12}\}$ and $\{a_0, a_1\}$, respectively.

- **Transition probabilities:** The transition probabilities of the Boyan chain are defined as

$$\begin{aligned} &p(s_{i-2} \mid s_i, a_0) = 1, \quad p(s_{i-1} \mid s_i, a_1) = 1, \quad \forall i \in \{2, 3, \ldots, 12\}, \\ &p(s_0 \mid s_1, a_0) = p(s_0 \mid s_1, a_1) = 1, \\ &p(s_j \mid s_0, a_0) = p(s_j \mid s_0, a_1) = \tfrac{1}{13}, \quad \forall j \in \{0, 1, \ldots, 12\}. \end{aligned}$$

- **Reward:** The reward function is defined as $r(i, a_0) = 0.5$ and $r(i, a_1) = 1$ for all $i \in \{0, 1, \ldots, 12\}$.
- **Feature matrix:** Consistent with the MRP experiment, we first construct the matrix $\widetilde{\boldsymbol{\Phi}}$, and append both the all-ones vector $\boldsymbol{e}$ and the differential value function $\boldsymbol{v}^\mu$ as columns

$$\boldsymbol{\Phi} = \left[\widetilde{\boldsymbol{\Phi}} \;\; \boldsymbol{e} \;\; \boldsymbol{v}^\mu\right].$$

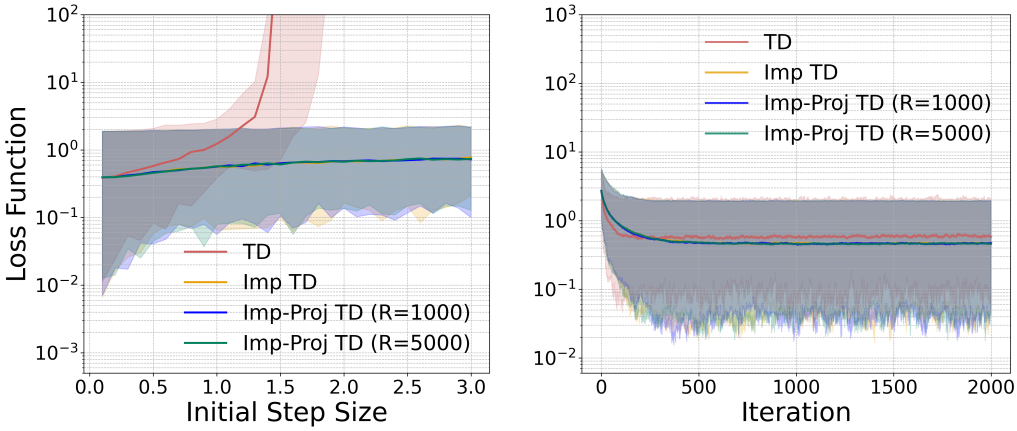
  Following Boyan (2002), the matrix $\widetilde{\boldsymbol{\Phi}}$ is constructed by linearly interpolating between the one-hot vectors $(1, 0, 0, 0)$ and $(0, 0, 0, 1)$ in increments of $1/4$. Specifically, the representation starts with state 0 as $(1, 0, 0, 0)$, then passes through intermediate states such as $(\frac{3}{4}, \frac{1}{4}, 0, 0)$ and $(\frac{1}{2}, \frac{1}{2}, 0, 0)$, eventually reaching $(0, 0, 0, 1)$. Finally, we normalize each row to ensure $\|\boldsymbol{\phi}(s)\| \leq 1$ for all $s \in \mathcal{S}$.

Additional results. We also report extended results for the Boyan experiments. For each step-size schedule, we conduct 50 independent runs with $T = 2000$. Each component of the parameter vector $\widehat{\theta}_0$ is initialized as $\mathcal{U}[-1, 1]$, and the initial average-reward estimate is set to $\widehat{\omega}_0 = 0$. In each experiment, we first sample actions in each state from a Binomial$(n = 13, p = 0.5)$ distribution. The sampled action will determine the deterministic policy to be evaluated. For each such sampled deterministic policy, an associated transition probability matrix $\boldsymbol{P}^\mu$ is induced. We provide results under the constant step-size schedule. The results as a function of the initial step-size, with hyperparameters $(\lambda, c_\alpha) = (0.25, 1.0)$, are shown in Figure 6. As shown in the left panel, the loss increases monotonically with the step-size across all methods. However, the growth is substantially reduced for the average-reward implicit TD(λ). In contrast, the average-reward TD(λ) diverges under similar conditions. The right panel further illustrates the trajectory of the loss value over training. At a moderately large initial step-size ($\beta_0 = 0.5$), the loss of average-reward TD(λ) exceeds that of its implicit counterpart.

D.2 Control Experiments

We provide details of the control experiment setup. For each action a, we form the joint feature $\boldsymbol{\phi}(s, a) = \boldsymbol{\phi}_{\text{RBF}}(s) \otimes e_a$, where e_a is the one-hot encoding of $a \in \mathcal{A}$. The state-action value function is approximated by

Figure 6: Boyan experiment results under the constant step-size, with exponential weighting parameter and step-size ratio set to $(\lambda, c_\alpha) = (0.25, 1.0)$. The solid line represents the mean, and the shaded region denotes the 95% confidence interval. (Left) Loss value with initial step-size β_0, from 0.1 to 3.0. (Right) Loss value over iterations with $\beta_0 = 0.5$.



$\widehat{Q}(s,a) = \boldsymbol{\phi}(s,a)^\top \widehat{\boldsymbol{\theta}}_t$, where $\widehat{\boldsymbol{\theta}}_t$ denotes the weight parameter at iteration t. We adopt SARSA with ϵ-greedy exploration: the agent selects the greedy action with probability $1-\epsilon$ and a random action with probability ϵ. The exploration parameter ϵ starts at 0.25, drops to 0.125 after 5000 iterations, and is set to 0 after 10000 iterations. Each experiment runs for $T = 15000$ steps, with 30 independent runs. Each component of the initial parameter estimate $\widehat{\theta}_0$ is initialized from $\mathcal{U}[-0.5, 0.5]$ and the initial average-reward estimate is $\widehat{\omega}_0 = 0$. The step-size schedule is $\beta_t = \beta_0/(t+400)^{0.99}$, with $\beta_0 \in \{400\times 0.25,\ 400\times 0.50,\ \ldots,\ 400\times 1.5\}$. Hence, the effective initial step-size $\beta_0/400^{0.99}$ ranges from 0.25 to 1.5.

D.2.1 Access-Control

Experiment setup We explain the state space, action space, and the reward function of the access-control experiment.

- **States and actions:** The state space is defined by the number of free servers and the class of arriving customer. Let $n \in \mathbb{N}$ be the total number of servers and $\mathcal{C} = \{1, 2, \ldots, C\}$ the set of customer classes. The state at time t is $S_t^\mu = (k_t, c_t) \in \{0, 1, \ldots, n\} \times \mathcal{C}$, where k_t is the number of free servers and c_t is the class of the arriving customer. Arrivals are equiprobable, that is: $\mathbb{P}(c_t = c) = \frac{1}{C}$ for each $c \in \mathcal{C}$, and this distribution is unknown to the decision maker. The decision maker either accepts the customer (action a_0) and assigns a free server, or rejects the customer (action a_1). Hence, the feasible action set is

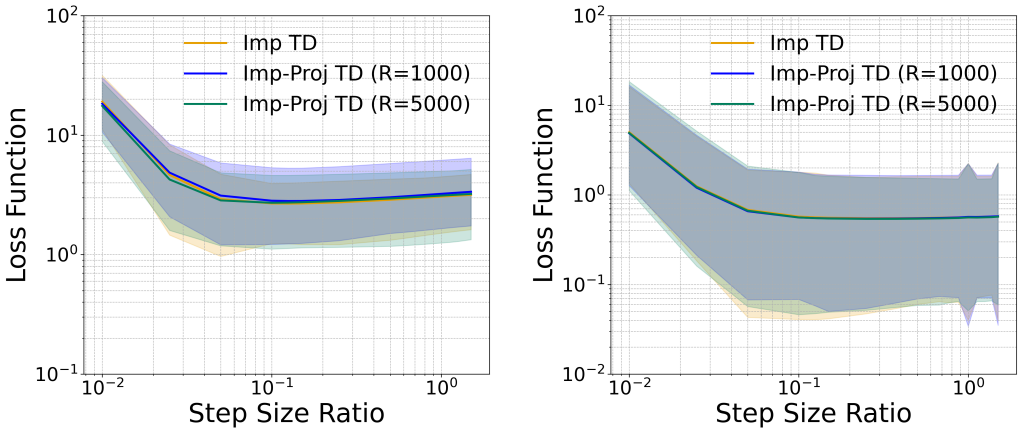
$$\mathcal{A}(S_t^\mu) = \begin{cases} \{a_0 \text{ (accept)}, a_1 \text{ (reject)}\}, & k_t > 0, \\ \{a_1 \text{ (reject)}\}, & k_t = 0. \end{cases}$$

- **Reward:** The one-step reward is $R_t^\mu = \frac{2^{c_t}}{2^C}\mathbb{I}\{a_t = a_0\}$, i.e., the decision maker receives $2^{c_t}/2^C$ if the customer is accepted and 0 otherwise.

- **State transitions:** If $a_t = a_0$ and $k_t > 0$, one available server is allocated. Let $\tilde{k}_t = k_t - \mathbb{I}\{a_t = a_0\}$ then $b_t = n - \tilde{k}_t$ is the number of busy servers immediately after the action. Each occupied server completes independently with probability $p \in (0,1)$ at the end of the period. Therefore, the number of free servers at the next step is $k_{t+1} = \min\{n,\ \tilde{k}_t + Y_t\}$, where $Y_t \sim \text{Binomial}(b_t, p)$. Lastly, the next arriving class c_{t+1} is drawn independently and uniformly from $\mathcal{C}$.

The goal is to find an optimal policy μ that maximizes the average-reward ω^μ. We follow the setup in Barto (2021), with $C = 4$ customer classes, $n = 10$ total servers, and a service completion rate $p = 0.06$. Accepted customers occupy a server until completion, at which point the server is freed. For feature representation, states

Figure 7: Effect of the step-size ratio c_α in both the MRP and Boyan experiments under a decaying step-size schedule, with exponential weighting parameter $\lambda = 0.25$ and initial step-size $\beta_0 = 1.0$. Solid lines denote mean values; shaded regions represent 95% confidence intervals. (Left) MRP (Right) Boyan



are rescaled to $[0, 1]$ and embedded via a single-scale random Fourier feature map (Pedregosa et al., 2011; Rahimi and Recht, 2008) implemented with scikit-learn's `RBFSampler`. The map uses twenty randomly drawn features and sets the inverse length-scale parameter to one, so that inner products in the resulting feature space provide a good approximation to the RBF kernel.

D.2.2 Pendulum

Experiment Setup At each time step t, the state is $S_t^\mu = (\cos\eta_t, \sin\eta_t, \dot{\eta}_t)$, where η_t is the pendulum angle and $\dot{\eta}_t$ its angular velocity. The action corresponds to applying a torque to the pendulum. We discretize the continuous action space $[-2, 2]$ into five actions ($\mathcal{A} = \{-2, -1, 0, 1, 2\}$). Unlike the episodic setting, this environment has no terminal state and runs indefinitely, so we optimize the long-run average-reward. The per step reward is $R_t^\mu = -\frac{\eta_t^2 + 0.1\dot{\eta}_t^2 + 0.001a_t^2}{16.27}$, where the normalization factor 16.27 scales the reward into $[-1, 0]$. We use the Gymnasium implementation of the pendulum environment (Towers et al., 2024). To approximate the state–action value, we use random Fourier features to approximate the RBF kernel. Concretely, we create two separate `RBFSampler` feature vectors—one using an inverse length-scale of 0.5 and the other 1.0, each with 150 features, and then concatenate them into a single 300-dimensional feature representation.

D.3 Effect of Step-Size Ratio

In this section, we study how the step-size ratio c_α affects stability and performance. We revisit the policy evaluation settings for the MRP and average-reward Boyan chain described in Sections D.1.1 and D.1.2, respectively. As in the main experiments, we report the loss value of the form $\left(\widehat{\omega} - \omega^\mu\right)^2 + \left\|\Pi_{\mathbb{O}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\right\|^2$ as the evaluation metric. Under the decaying step-size schedule $\beta_t = \beta_0/(t+1)^{0.99}$, we vary the step-size ratio $c_\alpha \in \{0.01, 0.05, 0.1, 0.125, 0.25, \ldots, 1.5\}$ while fixing the exponential weighting parameter $\lambda = 0.25$ and the initial step-size $\beta_0 = 1.0$. Figure 7 summarizes the result. For small values of c_α (e.g., $c_\alpha \leq 0.1$), the average-reward implicit TD(λ) method exhibits a modest increase in the loss value. However, beyond this threshold, the method remains stable across the entire range of c_α values.