A Finite-Sample Analysis of an Actor-Critic Algorithm for Mean-Variance Optimization in a Discounted MDP

Anonymous authors
Paper under double-blind review

Keywords: Risk-Sensitive RL, Mean-Variance Optimization, SPSA, Actor-Critic, TD Learning.

Summary

In many practical applications of reinforcement learning (RL), such as finance and mobility, safety considerations are paramount. Rather than solely maximizing expected rewards, one must also account for risk to ensure reliable decision-making. Traditional RL primarily focuses on expected reward maximization, a well-studied paradigm with both empirical and theoretical breakthroughs. In this paper, we adopt an alternative approach that integrates risk-awareness into policy optimization. Despite extensive research in risk-neutral RL, analyzing risk-sensitive RL algorithms remains challenging, as each risk metric requires a distinct analytical framework. We focus on variance—an intuitive and widely used risk measure—and analyze the **M**ean-**V**ariance **S**imultaneous **P**erturbation **S**tochastic **A**pproximation **A**ctor-**C**ritic (MV-SPSA-AC) algorithm, establishing finite-sample theoretical guarantees for the discounted reward Markov Decision Process (MDP) setting. Our analysis covers both policy evaluation and policy improvement within the actor-critic framework. We study a Temporal Difference (TD) learning algorithm with linear function approximation (LFA) for policy evaluation and derive finite-sample bounds that hold in both the mean-squared sense and with high probability under tail iterate averaging, with and without regularization. Additionally, we analyze the actor update using a simultaneous perturbation-based approach and establish convergence guarantees. These results contribute to the theoretical understanding of risk-sensitive actor-critic methods in RL, offering insights into variance-based risk-aware policy optimization.

Contribution(s)

1. We consider mean-variance optimization in a discounted MDP, and derive finite-sample guarantees for an actor-critic algorithm, with a critic based on linear function approximation, and an actor based on SPSA.
   Context: We consider a mean-variance MDP with the variance of the *return,* whose expectation is the usual risk-neutral objective. For this problem, existing work (L.A. & Ghavamzadeh, 2016) provides only asymptotic convergence guarantees.
2. For mean-variance policy evaluation, we employ TD learning with linear function approximation. We derive finite-sample bounds that hold (i) in the mean-squared sense and (ii) with high probability under tail iterate averaging, with and without regularization. Notably, our analysis for the regularized TD variant holds for a universal step size.
   Context: Non-asymptotic policy evaluation bounds are not available for variance of the return in a discounted MDP.
3. We employ an SPSA-based actor for policy optimization, and obtain an $O(n^{-\frac{1}{4}})$ bound in the number of actor iterations.
   Context: Notably, we resort to an SPSA-based actor, since the policy gradient theorem for variance is not amenable for direct use in an actor-critic algorithm; see L.A. & Ghavamzadeh (2016). Further, finite-sample bounds for a SPSA-based actor-critic algorithm are not available, even in the risk-neutral RL setting, to the best of our knowledge.

A Finite-Sample Analysis of an Actor-Critic Algorithm for Mean-Variance Optimization in a Discounted MDP

Anonymous authors
Paper under double-blind review

Abstract

Motivated by applications in risk-sensitive reinforcement learning, we study mean-variance optimization in a discounted reward Markov Decision Process (MDP). Specifically, we analyze a Temporal Difference (TD) learning algorithm with linear function approximation (LFA) for policy evaluation. We derive finite-sample bounds that hold (i) in the mean-squared sense and (ii) with high probability under tail iterate averaging, both with and without regularization. Our bounds exhibit an exponentially decaying dependence on the initial error and a convergence rate of $O(1/t)$ after t iterations. Moreover, for the regularized TD variant, our bound holds for a universal step size. Next, we integrate a Simultaneous Perturbation Stochastic Approximation (SPSA)-based actor update with an LFA critic and establish an $O(n^{-\frac{1}{4}})$ convergence guarantee, where n denotes the iterations of the SPSA-based actor-critic algorithm. These results establish finite-sample theoretical guarantees for risk-sensitive actor-critic methods in reinforcement learning, with a focus on variance as a risk measure.

1 Introduction

In the standard reinforcement learning (RL) setting, the objective is to learn a policy that maximizes the value function, which is the expected value of the cumulative reward obtained over a finite or infinite time horizon. However, in many practical scenarios such as finance, automated driving and drug testing, a risk sensitive learning paradigm is crucial, where the value function (an expectation) must be balanced with an appropriate risk metric associated with the reward distribution. One approach is to formulate a constrained optimization problem, using the risk metric as a constraint and the value function as the objective. Variance is a popular risk measure and is typically incorporated into risk-sensitive optimization as a constraint while optimizing for the expected value. This mean-variance formulation was introduced in the seminal work of Markowitz (1952). Mean-variance optimization in RL has been studied in several works; see, e.g., Mannor & Tsitsiklis (2013); Tamar et al. (2016); L.A. & Ghavamzadeh (2016). We study mean-variance optimization in a discounted reward Markov decision process (MDP). Our key contribution is the analysis of an actor-critic algorithm for mean-variance optimization, along with finite-sample guarantees in this setting.

Main Contributions. We study a discounted reward MDP with variance as the risk criterion and present two main contributions. Since one common approach to variance estimation is based on the difference between the second moment and the square of the first moment, estimating both moments is essential. Our first key contribution concerns the sub-problem of jointly evaluating the value function (first moment) and the second moment of the discounted cumulative reward. For simplicity, we refer to the second moment of the discounted cumulative reward as the square-value function. To address the curse of dimensionality in large state-action spaces, we analyze temporal difference (TD) learning with linear function approximation (LFA) for these estimates.

Table 1: Summary of the MSE bounds for a TD-critic.

Paper	**Iterate**	**Objective**	**Rate**	**Step size**	**Universal step size**
L.A. & Ghavamzadeh (2016)	Last iterate	Mean-variance	–[1]	$\frac{c_0 c}{c+t}$	✗
Dalal et al. (2018)	Last iterate	Mean	$O(1/t^\sigma)$	$1/t^\sigma$	✓
Bhandari et al. (2021)[2]	Full average	Mean	$O(1/t)$	$1/\sqrt{T}$	✓
Eldowa et al. (2022)	Full average	Mean-variance[3]	$O(1/t)$	constant	✗
Patil et al. (2023)	Tail average	Mean	$O(1/t)$	constant	✓
Agrawal et al. (2024)	Tail average	Mean-variance[4]	$O(1/t)$	constant	✗
Mitra (2025)	Weighted average[5]	Mean	$O(1/t)$	constant	✗
This work	Tail average	Mean-variance	$O(1/t)$	constant	✗
This work	Regularized tail average	Mean-variance	$O(1/t)$	constant	✓

[1] Asymptotic convergence of mean-variance TD shown. Here, c_0 and c are arbitrary constants depending on the minimum eigenvalue. [2] T = number of TD iterations. [3] Variance of per-step reward as the risk measure. [4] Asymptotic variance for average-reward MDP as the risk measure. [5] Weights are determined by $(1 - \alpha A)^{-(t+1)}$ with $A = 0.5\omega(1 - \gamma)$, which makes them indirectly dependent on the minimum eigenvalue ω and the discount factor γ. Here, α is step size dependent on the minimum eigenvalue ω.

We present finite-sample bounds that quantify the deviation of the iterates from the fixed point, both in expectation and with high probability. The fixed point is joint in the sense that it includes both the value function and the square-value function. We present bounds for a constant step-size with and without tail-averaging; see Table 1 for a summary. Next, we establish $O(1/t)$ finite-time convergence bounds for tail-averaged TD iterates, where t denotes the number of iterations of the TD algorithm. Furthermore, we present a finite-sample analysis of the regularized TD algorithm. From this analysis, we establish an $O(1/t)$ bound, similar to the unregularized case. An advantage of regularization is that the step-size choice is universal, i.e., it does not require knowledge of the eigenvalues of the underlying linear system, whereas the unregularized TD bounds depend on such eigenvalue information, which is typically unknown in practice.

While finite-sample analysis of TD with LFA has been studied in several recent works (cf. Prashanth et al., 2021; Dalal et al., 2018; Bhandari et al., 2021; Samsonov et al., 2024; Agrawal et al., 2024), to the best of our knowledge, no prior work has established finite-sample bounds for policy evaluation of variance in the discounted reward MDP setting. Our bounds explicitly characterize their dependence on the discount factor, feature bounds, and rewards. Compared to existing finite-sample bounds for TD learning, the analysis of mean-variance-style TD updates is more intricate, as it requires tracking the solution of an additional projected fixed point by solving a separate Bellman equation for the square-value function. Furthermore, the Bellman equation associated with the square-value function includes a cross-term involving the value function (see (25) in the supplementary material). Due to this cross-term, obtaining a standard $O(1/t)$ mean-squared error bound is challenging when using a constant step size, unless the spectral properties of the underlying linear system are known. To overcome this dependence, we investigate a regularized version of the mean-variance TD updates. To the best of our knowledge, ours is the first work to obtain a $O(1/t)$ MSE bound with a universal step size for mean-variance TD. Prior works on TD-type algorithms for other notions of variance, cf. Agrawal et al. (2024); Eldowa et al. (2022), present $O(1/t)$ bounds with a step size choice that requires underlying eigenvalue information.

Our second key contribution lies in analyzing an actor-critic algorithm for mean-variance and deriving finite-sample guarantees. The critic part uses the aforementioned LFA-based policy evaluation for a fixed policy parameter. The actor uses an SPSA-based gradient estimator (Spall, 1992), departing from the more common risk-neutral approach of employing a likelihood ratio-based gradient estimator supported by the policy gradient theorem (see Section 4 for a discussion on SPSA's necessity). SPSA estimates policy gradients for the value and square-value functions using two policy trajectories: one generated using the current policy parameter and another using a randomly perturbed parameter.

We provide non-asymptotic convergence rates for an SPSA-based actor in the mean-variance framework. This result quantifies convergence to the stationary point in terms of the gradient norm of the Lagrangian, addressing a gap in prior work that focused exclusively on asymptotic guarantees. As an aside, mean-variance optimization has been shown to be NP-hard, even with model information available (Mannor & Tsitsiklis, 2013). Actor-critic methods present a viable alternative approach, and our analysis provides the rate of convergence for such an algorithm tailored to the mean-variance setting. Specifically, we show an $O(n^{-\frac{1}{4}})$ performance guarantee for the overall algorithm, where n is the number of actor loop iterations. To the best of our knowledge, there are no finite-sample guarantees for zeroth order actor-critic, even for the risk-neutral setting.

Our results are beneficial for three reasons. First, we exhibit $O(1/t)$ bounds for the regularized TD variant with a step size that is universal. In contrast, a universal step size for vanilla mean-variance TD is not feasible owing to certain cross-terms that are unique to the case of mean-variance policy evaluation. Our key observation is that regularization enables the use of a universal step size that is independent of the eigenvalues of the underlying system. Second, our proof is tailored to mean-variance TD, making the constants clear. In contrast, it is difficult to infer them from the general LSA bounds in (Durmus et al., 2024; Mou et al., 2020). Third, we provide high-probability bounds that exhibit better scaling w.r.t. the confidence parameter as compared to Samsonov et al. (2024).

Related Work. This paper performs a finite-sample analysis of a TD critic, and an SPSA actor for mean-variance optimization in a discounted RL setting. We briefly review relevant works on each of these topics.

Critic. TD learning, originally proposed by Sutton (1988), has been widely used for policy evaluation in RL. Tsitsiklis & Van Roy (1997) established asymptotic convergence guarantees for TD learning with LFA. Many recent works have focused on providing non-asymptotic convergence guarantees for TD learning (Bhandari et al., 2021; Dalal et al., 2018; Lakshminarayanan & Szepesvari, 2018; Srikant & Ying, 2019; Prashanth et al., 2021; Patil et al., 2023; Durmus et al., 2024). In a recent study by Samsonov et al. (2024), the authors derived refined error bounds for TD learning by combining proof techniques from (Mou et al., 2020; Durmus et al., 2024) with a stability result for the product of random matrices. In contrast, our results target a different system of linear equations. Moreover, as mentioned before, our bounds for regularized TD feature a universal step size. The reader is referred to Section 3 for a detailed comparison of our critic bounds to the current literature.

Actor-Critic. In (Lei et al., 2025), the authors propose a zeroth-order actor critic in a risk-neutral RL setting. However, they do not provide a finite-sample analysis. In (L.A. & Ghavamzadeh, 2016), which is the closest related work, the authors propose an SPSA-based actor-critic algorithm for mean-variance optimization, and establish asymptotic convergence. In contrast, we provide a finite-sample analysis of their algorithm with a few variations: (i) We incorporate tail-averaging in TD-critic and derive finite-sample bounds for a universal step size; (ii) We prove a smoothness result for the Lagrangian of the mean-variance problem and use this result to provide a non-asymptotic bound for the SPSA-based actor that employs mini-batching for the critic updates. In (Xu et al., 2020; Kumar et al., 2023), the authors analyze risk-neutral actor critic algorithms with a gradient estimate based on the likelihood ratio method. They provide a finite-sample analysis. However, the likelihood ratio method for gradient estimation does not work for the case of variance, and hence, our non-asymptotic analysis involves a significant departure in the proof for the SPSA-based actor that we consider.

2 Problem formulation

We consider an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$, both assumed to be finite. The reward function $r(s,a)$ maps state-action pairs (s,a) to a reward, with $s \in \mathcal{S}$ and $a \in \mathcal{A}$. In this work, we consider a stationary randomized policy π which maps each state to a probability distribution over the action space. We consider a discounted MDP setting, and use $\gamma \in (0,1)$ to denote the discount factor. We use $\mathbb{P}(s'|s,a)$ to denote the probability of transitioning from state s to next state s' given that action a is chosen following a policy π. The transition probability matrix $\mathbf{P}$ gives the probability of going from state s to s' given a policy π. The elements of this matrix of dimension $|\mathcal{S}| \times |\mathcal{S}|$ are given by $\mathbf{P}(s,s') = \sum_a \pi(a|s)\mathbb{P}(s'|s,a)$. The value function $V^\pi(s)$, which denotes the expected value of cumulative sum of discounted rewards when starting from state $s_0 = s$ and following the policy π, is defined as

$$V^\pi(s) \triangleq \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r(s_t,a_t) \,|\, s_0 = s\right]. \tag{1}$$

Furthermore, the variance of the infinite horizon discounted reward from state $s_0 = s$, denoted as $\Lambda^\pi(s)$, is defined as $\Lambda^\pi(s) \triangleq U^\pi(s) - V^\pi(s)^2$, where $U^\pi(s)$ represents the second moment of the cumulative sum of discounted rewards, and is defined as

$$U^\pi(s) \triangleq \mathbb{E}\left[\left(\sum_{t=0}^\infty \gamma^t r(s_t,a_t)\right)^2 \,\middle|\, s_0 = s\right]. \tag{2}$$

Henceforth, we shall refer to U^π as the square-value function. The well-known mean-variance optimization problem in a discounted MDP context is as follows: For a given state $s_0 = s$ and threshold $c > 0$, our goal is to solve the following constrained optimization problem:

$$\max_\pi V^\pi(s) \qquad \text{subject to} \qquad \Lambda^\pi(s) \le c. \tag{3}$$

The value function $V^\pi(s)$ satisfies the Bellman equation $T_1 V^\pi = V^\pi$, where $T_1 : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is the Bellman operator, defined by $T_1(V^\pi(s_0)) \triangleq \mathbb{E}^{\pi,\mathbf{P}}\left[r(s_0,a_0) + \gamma V^\pi(s')\right]$, where the actions are chosen according to the policy π. It is well known that T_1 is a contraction mapping. In Sobel (1982), the author derives a Bellman type equation for $\Lambda^\pi(s)$. However, the underlying operator of this equation is not monotone. To workaround this problem, Tamar et al. (2016); L.A. & Ghavamzadeh (2016) use the square-value function U^π, which satisfies a fixed point relation that is monotone. Given V^π, U^π, the variance can be calculated using Λ^π. Using Proposition 6.1 in (L.A & Fu, 2022), we expand the square-value function (2) as

$$U^\pi(s) = \sum_a \pi(a|s) r(s,a)^2 + \gamma^2 \sum_{a,s'} \pi(a|s)\mathbb{P}(s'|s,a)U^\pi(s') + 2\gamma \sum_{a,s'} \pi(a|s)\mathbb{P}(s'|s,a) r(s,a) V^\pi(s')$$

Similar to the value function, the square-value function also satisfies a Bellman equation $T_2 U^\pi = U^\pi$, where $T_2 : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is the Bellman operator, given by $T_2 U^\pi(s) \triangleq \mathbb{E}^{\pi,\mathbf{P}}[r(s,a)^2 + \gamma^2 U^\pi(s') + 2\gamma r(s,a) V^\pi(s')]$. For a given policy π, the Bellman operators T_1 and T_2 can be represented in a compact vector-matrix form as $T_1(V) = r + \gamma \mathbf{P} V,\ T_2(U) = \tilde{r} + 2\gamma \mathbf{R}\mathbf{P}V + \gamma^2 \mathbf{P} U$, where U, V, r and $\tilde{r}$ are $|\mathcal{S}| \times 1$ vectors with $r(s_i) = \sum_{a\in\mathcal{A}} \pi(a|s_i) r(s_i,a),\ \ \tilde{r}(s_i) = \sum_{a\in\mathcal{A}} \pi(a|s_i) r(s_i,a)^2$. Here, $\mathbf{R}$ is a $|\mathcal{S}| \times |\mathcal{S}|$ diagonal matrix with $r(s_i)$ as the diagonal elements for $i \in \{1,\ldots,|\mathcal{S}|\}$. Now, we construct an operator $T : \mathbb{R}^{2|\mathcal{S}|} \to \mathbb{R}^{2|\mathcal{S}|}$, which is given by $T(V,U) = (T_1(V), T_2(U))^\top$ A sub-problem of (3) is policy evaluation, i.e., estimation of $V^\pi(\cdot)$ and $\Lambda^\pi(\cdot)$ for a given policy π. L.A & Fu (2022); Tamar et al. (2016) establish that the operator T is a contraction mapping with respect to a weighted norm, ensuring a unique fixed point for T. In the next section, we describe a TD algorithm with LFA for policy evaluation, and this algorithm is based on (L.A. & Ghavamzadeh, 2016).

3 Mean-variance TD-critic

When the size of the underlying state space $|\mathcal{S}|$ is large, policy evaluation suffers the curse of dimensionality, necessitating the computation and storage of the value function for each state in the underlying MDP. A standard approach to overcome this difficulty is to use TD learning with *function approximation*, wherein the value function is approximated using a simple parametric class of functions. The most common example of this is TD learning with LFA (Tsitsiklis & Van Roy, 1997), where the value function for each state is approximated using a linear parameterized family, i.e., $V^\pi(s) \approx \omega^\top \phi(s)$, where $\omega \in \mathbb{R}^q$ is a tunable parameter common to all states, and $\phi : \mathcal{S} \to \mathbb{R}^q$ is a feature vector for each state $s \in \mathcal{S}$, and typically $q \ll |\mathcal{S}|$.

We approximate the value function $V^\pi(s)$ and the square-value function $U^\pi(s)$ using linear functions as follows: $V^\pi(s) \approx v^\top \phi_v(s)$, $U^\pi(s) \approx u^\top \phi_u(s)$, where the features $\phi_v(\cdot)$ and $\phi_u(\cdot)$ belong to low-dimensional subspaces in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Let $\mathbf{\Phi}_v$ and $\mathbf{\Phi}_u$ denote $|\mathcal{S}| \times d_1$ and $|\mathcal{S}| \times d_2$ dimensional matrices, with i-th and j-th column respectively as $\left(\phi_v^i(s_1), \ldots, \phi_v^i(s_{|\mathcal{S}|})\right)^\top$, $\left(\phi_u^j(s_1), \ldots, \phi_u^j(s_{|\mathcal{S}|})\right)^\top$ where $i \in \{1, \ldots, d_1\}$ and $j \in \{1, \ldots, d_2\}$. For analytical convenience, in our analysis we set $d_1 = d_2 = q$. We observe that owing to the function approximation, the actual fixed point remains inaccessible. Instead, the objective is to find the projected fixed points, denoted as $\bar{w} = (\bar{v}, \bar{u})^\top$ within the following subspaces: $S_v \coloneqq \left\{\mathbf{\Phi}_v v \,\middle|\, v \in \mathbb{R}^{d_1}\right\}$, $S_u \coloneqq \left\{\mathbf{\Phi}_u u \,\middle|\, u \in \mathbb{R}^{d_2}\right\}$. We approximate the value and square-value functions within the subspaces defined above. Accordingly, we construct projections onto S_v and S_u with respect to a weighted norm, using the stationary distribution as weights. For the analysis, we require the following assumptions that are standard for TD with LFA, (cf. Prashanth et al., 2021; Bhandari et al., 2021; Srikant & Ying, 2019; Patil et al., 2024).

Assumption 1. *The Markov chain underlying the policy π is irreducible.*

Assumption 2. *The matrices $\mathbf{\Phi}_v$ and $\mathbf{\Phi}_u$ have full column rank.*

With finite state and action spaces, Assumption 1 guarantees the existence of a unique stationary distribution χ_π for the Markov chain induced by policy π. Assumption 2, commonly made in the context of TD with LFA (cf. Bhatnagar et al. (2009); Bhandari et al. (2021); Prashanth et al. (2021)), mandates that the columns of the feature matrices $\mathbf{\Phi}_v$ and $\mathbf{\Phi}_u$ be linearly independent, guaranteeing the uniqueness of the fixed points. Additionally, it also ensures the existence of inverse of the feature covariance matrices ($\mathbf{\Phi}_v^\top \mathbf{D}^\pi \mathbf{\Phi}_v$ and $\mathbf{\Phi}_u^\top \mathbf{D}^\pi \mathbf{\Phi}_u$), to define the projection matrices in (4).

We denote $\mathbf{\Pi}_v$ and $\mathbf{\Pi}_u$ as the projection matrices which project from state space $\mathcal{S}$ onto the subspaces $\mathcal{S}_v$ and S_u, respectively. For a given policy π, projection matrices are defined as in (L.A. & Ghavamzadeh, 2016, Eq. (8)):

$$\mathbf{\Pi}_v = \mathbf{\Phi}_v(\mathbf{\Phi}_v^\top \mathbf{D}^\pi \mathbf{\Phi}_v)^{-1} \mathbf{\Phi}_v^\top \mathbf{D}^\pi \text{ and } \mathbf{\Pi}_u = \mathbf{\Phi}_u(\mathbf{\Phi}_u^\top \mathbf{D}^\pi \mathbf{\Phi}_u)^{-1} \mathbf{\Phi}_u^\top \mathbf{D}^\pi, \tag{4}$$

where $\mathbf{\Pi}_v$ and $\mathbf{\Pi}_u$ project the true value and square-value functions onto the linear spaces spanned by the columns of $\mathbf{\Phi}_v$ and $\mathbf{\Phi}_u$, respectively. In the above, $\mathbf{D}^\pi$ is a diagonal matrix with entries from the stationary distribution χ. In (L.A. & Ghavamzadeh, 2016), the authors established the following projected fixed point relations:

$$\mathbf{\Phi}_v \bar{v} = \mathbf{\Pi}_v T_v(\mathbf{\Phi}_v \bar{v}), \text{ and } \mathbf{\Phi}_u \bar{u} = \mathbf{\Pi}_u T_u(\mathbf{\Phi}_u \bar{u}). \tag{5}$$

(L.A & Fu, 2022, Proposition 6.2) establishes that the joint operator $T(U, V) = \begin{pmatrix} T_v \\ T_u \end{pmatrix}$ is a contraction with respect to a weighted norm. Since the operator $\mathbf{\Pi} = \begin{pmatrix} \mathbf{\Pi}_v & 0 \\ 0 & \mathbf{\Pi}_u \end{pmatrix}$ is non-expansive and the matrices $\mathbf{\Phi}_v$ and $\mathbf{\Phi}_u$ have full column rank, (Tamar et al., 2016, Proposition 8) ensures that the projected Bellman operator $\Pi T(U, V)$ is also a contraction with respect to a weighted norm. Consequently, the projected Bellman operator $\Pi T(U, V)$ admits a unique projected fixed point $\bar{w} = (\bar{v}, \bar{u})^\top$. The equations in (5) can therefore be equivalently expressed as the following linear system:

$$\begin{aligned} &-\mathbf{M}\bar{w} + \xi = 0, \text{ where } \mathbf{M} = \begin{pmatrix} \mathbf{\Phi}_v^\top \mathbf{D}(\mathbf{I}-\gamma\mathbf{P})\mathbf{\Phi}_v & 0 \\ -2\gamma \mathbf{\Phi}_u^\top \mathbf{DRP}\mathbf{\Phi}_v & \mathbf{\Phi}_u^\top \mathbf{D}(\mathbf{I}-\gamma^2\mathbf{P})\mathbf{\Phi}_u \end{pmatrix}, \quad \xi = \begin{pmatrix} \mathbf{\Phi}_v^\top \mathbf{DR} \\ \mathbf{\Phi}_u^\top \mathbf{D}\tilde{r} \end{pmatrix}, \\ &r = \begin{pmatrix} r(s_1) & \ldots & r(s_{|\mathcal{S}|}) \end{pmatrix}^\top, \text{and} \end{aligned} \tag{6}$$

the matrix $\mathbf{R}$ is diagonal, with its components given by $r(s_i) = \sum_{a \in \mathcal{A}} \pi(a|s_i) r(s_i, a)$ for $i \in \{1, \ldots, |\mathcal{S}|\}$. $\tilde{r}$ is a vector with its components given by $\tilde{r}(s_i) = \sum_{a \in \mathcal{A}} \pi(a|s_i) r(s_i, a)^2$.

Algorithm 1: TD with Tail Averaging (Critic)

Input: Initialize $w_0 = (v_0, u_0)$, step-size β, critic batch size m, tail index k
Output: Tail-averaged iterate $w_{k+1:m} = (\frac{1}{m-k}\sum_{t=k+1}^{m} v_t\ ,\ \frac{1}{m-k}\sum_{t=k+1}^{m} u_t)^\top$
for $t = 0$ **to** m **do**
  Sample action a_t using the policy $\pi(\cdot|s_t)$, observe the next state s_{t+1} and reward $r_t = r(s_t, a_t)$
  `/* Update the TD parameters as follows: */`

$$v_{t+1} = v_t + \beta\ \delta_t\ \phi_v(s_t), \quad u_{t+1} = u_t + \beta\ \epsilon_t\ \phi_u(s_t) \tag{7}$$

  where $\delta_t = r_t + \gamma v_t^\top \phi_v(s_{t+1}) - v_t^\top \phi_v(s_t)$,

$$\epsilon_t = {r_t}^2 + 2\gamma r_t v_t^\top \phi_v(s_{t+1}) + \gamma^2 u_t^\top \phi_u(s_{t+1}) - u_t^\top \phi_u(s_t).$$

end for

Basic algorithm. Letting $w_t = (v_t, u_t)^\top$, we rewrite (7) to obtain the following update iteration:

$$w_{t+1} = w_t + \beta(r_t\phi_t - \mathbf{M}_t w_t), \tag{8}$$

where $\phi_t = (\phi_v(s_t), r(s_t, a_t)\phi_u(s_t))^\top, \mathbf{M}_t \triangleq \begin{pmatrix} \mathbf{a}_t & \mathbf{o} \\ \mathbf{c}_t & \mathbf{b}_t \end{pmatrix}$ with $\mathbf{c}_t \triangleq -2\gamma r_t \phi_u(s_t)\phi_v(s_{t+1})^\top$, $\mathbf{a}_t \triangleq \phi_v(s_t)\phi_v(s_t)^\top - \gamma\phi_v(s_t)\phi_v(s_{t+1})^\top$ and $\mathbf{b}_t \triangleq \phi_u(s_t)\phi_u(s_t)^\top - \gamma^2\phi_u(s_t)\phi_u(s_{t+1})^\top$.

In (8), we have used r_t to denote $r(s_t, a_t)$, for notational convenience. We observe that the expected value of $\mathbf{M}_t$ is equal to $\mathbf{M}$, where $\mathbf{M}$ is defined in (6). An alternative view of the update rule is the following:

$$w_{t+1} = w_t + \beta(-\mathbf{M}w_t + \xi + \Delta M_t), \tag{9}$$

where $\Delta M_t = r_t\phi_t - \mathbf{M}_t w_t - \mathbb{E}\left[r_t\phi_t - \mathbf{M}_t w_t \mid \mathcal{F}_t\right]$, with ξ as defined in (6). Under an i.i.d. observation model (see Assumption 5), ΔM_t is a martingale difference w.r.t. the filtration $\{\mathcal{F}_t\}_{t\geq 0}$, where $\mathcal{F}_t$ is the sigma field generated by $\{w_0, \ldots, w_t\}$. We remark that we utilize the update iteration (8) instead of (9) to obtain finite-sample bounds in the next section. The rationale behind this choice is a technical advantage of not requiring a projection operator to keep the iterates w_t bounded. To elaborate, in the proof of finite-sample bounds, we unroll the iteration in (8) and bound the bias and variance terms. Specifically, letting $z_t = w_t - \bar{w}$ and $h_t(w_t) = r_t\phi_t - \mathbf{M}_t w_t$, we get $z_{t+1} = (\mathbf{I} - \beta\mathbf{M}_t)z_t + \beta h_t(\bar{w})$. The second term $h_t(\bar{w})$ does not depend on the iterate w_t and can be bounded directly. On the other hand, unrolling (9) would result in a term $\beta\Delta M_t$ in place of the $h_t(\bar{w})$, and bounding this term requires a projection since ΔM_t has the iterate w_t.

Tsitsiklis & Van Roy (1997) show asymptotic convergence of v_t to $\bar{v}$. They achieved this by verifying that the required conditions—on step-size, stability, and noise control—are satisfied with the TD update reinterpreted as as Linear Stochastic Approximation (LSA) iteration. Similarly, the convergence of w_t to $\bar{w}$ was established by L.A. & Ghavamzadeh (2016). Several recent works have analyzed the finite-sample behavior of TD learning with LFA, particularly focusing on deriving mean-squared error bounds (Bhandari et al., 2021). However, a direct finite-sample analysis of (8) is not available in the literature—a gap that we address next.

Bounds for the TD-critic. We make the following assumptions that are common in the finite-sample analysis of temporal difference (TD) learning, (cf. Prashanth et al., 2021; Bhandari et al., 2021; Patil et al., 2024).

Assumption 3. $\forall s \in \mathcal{S}$, $\|\phi_v(s)\|_2 \leq \phi^v_{\mathsf{max}} < \infty$, $\|\phi_u(s)\|_2 \leq \phi^u_{\mathsf{max}} < \infty$.

Assumption 4. $\forall s \in \mathcal{S}, a \in \mathcal{A}$, $|r(s,a)| \leq R_{\max} < \infty$.

Assumption 3 ensures the existence of the feature covariance matrices $\mathbf{\Phi}_v^\top \mathbf{D}^\pi \mathbf{\Phi}_v$ and $\mathbf{\Phi}_u^\top \mathbf{D}^\pi \mathbf{\Phi}_u$, as well as the projection matrices in (4). Assumption 4 bounds the rewards uniformly, ensuring the existence of the value function and the square-value function. We consider an i.i.d observation model, which is made precise in the assumption below.

Assumption 5. *The samples $\{s_t, r_t, s_{t+1}\}_{t\in\mathbb{N}}$ are formed as follows: For each t, (s_t, s_{t+1}) are drawn independently and identically from $\chi(s)\mathbf{P}(s, s')$, where χ is the stationary distribution un-*

derlying policy π, and $\mathbf{P}$ is the transition probability matrix of the Markov chain underlying the given policy π. Further, r_t is a function of s_t and a_t, which is chosen using the given policy π.

The i.i.d observation model is often considered as first step to analyse TD learning. Furthermore, the finite-time bounds obtained under the i.i.d. observation model can be directly extended to the Markovian setting using the constructions in (Patil et al., 2024, Remark 6) and (Samsonov et al., 2024, Section 5).

Mean-Squared Error Bounds. We first present a mean-squared error bound for the last iterate with a constant step size, with the proof in Section 7.

Theorem 3.1. *Suppose Assumptions 1 to 5 hold. Run TD Updates in (7) for t iterations with a step size β satisfying the following constraint: $\beta \leq \beta_{\max} = \frac{\mu}{c}$ where $\mu = \lambda_{\mathsf{min}}(\frac{\mathbf{M}^\top + \mathbf{M}}{2})$ and $c = \max\left\{4(\phi^v_{\mathsf{max}})^4 + 4\gamma^2 R^2_{\max}(\phi^u_{\mathsf{max}})^2(\phi^v_{\mathsf{max}})^2, 4(\phi^u_{\mathsf{max}})^4\right\} + 2\gamma R_{\max}((\phi^v_{\mathsf{max}})^2(\phi^u_{\mathsf{max}})^2 + (\phi^u_{\mathsf{max}})^4)$. Then, we have*

$$\mathbb{E}\left[\|w_{t+1} - \bar{w}\|_2^2\right] \leq 2\exp\left(-\beta\mu t\right)\mathbb{E}\left[\|z_0\|_2^2\right] + \frac{2\beta\sigma^2}{\mu}, \tag{10}$$

where w_0 is the initial parameter, $\bar{w}$ is the TD fixed point, $z_0 = w_0 - \bar{w}$ is initial error and $\sigma^2 = 2R^2_{\mathsf{max}}\left((\phi^v_{\mathsf{max}})^2 + R^2_{\max}(\phi^u_{\mathsf{max}})^2\right) + 2\left((\phi^v_{\mathsf{max}})^4(1+\gamma)^2 + (\phi^u_{\mathsf{max}})^4\left(1+\gamma^2\right)^2 + 4\gamma^2 R^2_{\max}(\phi^v_{\mathsf{max}})^2(\phi^u_{\mathsf{max}})^2\right)\|\bar{w}\|_2^2$.

Notice that the bound in (10) is for a constant stepsize that requires information about the minimum eigenvalue of the symmetric part of $\mathbf{M}$. In the context of regular TD, such a problematic eigenvalue dependence has been surmounted using tail-averaging, which we introduce next. We remark that tail-averaging for the case of mean-variance TD does not overcome the eigenvalue dependence. However, the benefit of tail averaging is that we obtain a bound that vanishes as as $t \to \infty$, while the bound in (10) does not vanish asymptotically.

Tail averaging. The tail-average is computed by averaging the iterates $\{w_{k+1}, \ldots, w_t\}$, given by $w_{k+1:t} = \frac{1}{t-k}\sum_{i=k+1}^{t} w_i$, where k is the tail index, and averaging starts at $k+1$. Polyak & Juditsky (1992); Fathi & Frikha (2013) investigated the advantages of iterate averaging, providing the asymptotic and non-asymptotic convergence guarantees in the stochastic approximation literature, respectively. Tail averaging preserves the advantages of iterate averaging, while also ensuring dependence on initial error is forgotten at a faster rate (Patil et al., 2023; Samsonov et al., 2024). Now, we present a mean-squared error bounds for the tail-averaged variant for the TD-critic, with the proof in Section 8.

Theorem 3.2. *Suppose Assumptions 1 to 5 hold. Run Algorithm 1 for t iterations with a step size β as specified in Theorem 3.1. Then, we have the following bound for the tail average iterate $w_{k+1:t} = \frac{1}{t-k}\sum_{i=k+1}^{t} w_i$:*

$$\mathbb{E}\left[\|w_{k+1:t} - \bar{w}\|_2^2\right] \leq \frac{10\exp\left(-k\beta\mu\right)}{\beta^2\mu(t-k)^2}\mathbb{E}[\|z_0\|_2^2] + \frac{10\sigma^2}{\mu^2(t-k)}, \tag{11}$$

where $z_0, \sigma, \bar{w}, \mu$ are as defined in Theorem 3.1.

As in the case of regular TD with tail averaging, it can be observed that the initial error (the first term in (11)) is forgotten exponentially. The second term, with $k = t/2$ (or any other fraction of t), decays as $O(1/t)$. Tail averaging is advantageous when compared to full iterate averaging (i.e., $k = 1$), as the latter would not result in an exponentially decaying initial error term. The bound for regular TD with tail averaging in Patil et al. (2024) uses a universal step-size, which does not require information about the eigenvalues of the underlying feature matrix. However, arriving at $O(1/t)$ bound for the case of variance is challenging owing to certain cross-terms that cannot be handled in a manner analogous to regular TD, see Section 6 for the details.

Regularization for universal step size. The results in Theorems 3.1–3.2 suffer from the disadvantage of a stepsize which requires knowledge of the spectral properties of the underlying matrix

$\mathbf{M}$. In practical RL settings, such information is seldom available. To circumvent this shortcoming, we propose a regularization-based TD algorithm that works with a universal step size, for a suitably chosen regularization parameter. Instead of (6), we solve the following regularized linear system for some $\zeta > 0$:

$$-(\mathbf{M} + \zeta \mathbf{I})\bar{w}_{\text{reg}} + \xi = 0, \tag{12}$$

The corresponding TD updates in (7) to solve (12) would become

$$\check{v}_{t+1} = (\mathbf{I} - \check{\beta}\zeta)\check{v}_t + \check{\beta}\ \check{\delta}_t\ \phi_v(s_t), \quad \check{u}_{t+1} = (\mathbf{I} - \check{\beta}\zeta)\check{u}_t + \check{\beta}\ \check{\epsilon}_t\ \phi_u(s_t), \tag{13}$$

where $\check{\delta}_t, \check{\epsilon}_t$ are the regularized variants of the corresponding quantities defined in (7), i.e., with v_t, u_t replaced by $\check{v}_t, \check{u}_t$ respectively. We combine the updates in (13) as

$$\check{w}_{t+1} = \check{w}_t + \check{\beta}(r_t \phi_t - (\zeta \mathbf{I} + \mathbf{M}_t)\check{w}_t), \tag{14}$$

where M_t, r_t, ϕ_t are defined in (8). We now present a result that shows the regularized tail-averaged variant (14) converges at the optimal rate of $O(1/t)$ in the mean-squared sense, for a step size that is universal.

Theorem 3.3. *Suppose Assumptions 1 to 5 hold. Let $\check{w}_{k+1:t} = \frac{1}{t-k}\sum_{i=k+1}^{t-k} \check{w}_i$ denote the tail-averaged regularized iterate. For $\zeta = \frac{1}{\sqrt{t-k}}$ and the step size $\check{\beta}$ satisfying $\check{\beta} \leq \check{\beta}_{\max} = \frac{\zeta}{\check{c}}$. Then we have*

$$\mathbb{E}\left[\|\check{w}_{k+1:t} - \bar{w}\|_2^2\right] \leq \frac{20\exp\left(-k\check{\beta}(2\mu + N^{-\frac{1}{2}})\right)}{\check{\beta}^2(2\mu + N^{-\frac{1}{2}})^2 N^2}\mathbb{E}\left[\|\check{w}_0 - \bar{w}_{\mathsf{reg}}\|_2^2\right] + \frac{20\check{\sigma}^2}{\mu^2 N} + \frac{2R_{\max}^2\big((\phi^v_{\max})^2 + R^2_{\max}(\phi^u_{\max})^2\big)}{\iota^2 N},$$

where $\check{c}$ and $\check{\sigma}$ are defined in Section 9, ι denotes the minimum singular value of $\mathbf{M}$, $N = t - k$, and $\mu = \lambda_{\min}(\frac{\mathbf{M}^\top + \mathbf{M}}{2})$

We first bound $\mathbb{E}\left[\|\check{w}_{k+1:t} - \bar{w}_{\mathsf{reg}}\|_2^2\right]$ in Theorem 9.1 in the supplementary material, specialize this bound for the case of $\zeta = \frac{1}{\sqrt{t-k}}$. Next, using the fact that $\|\bar{w}_{\mathsf{reg}} - \bar{w}\|_2^2$ is $O(\zeta^2)$, followed by a triangle inequality, we obtain the bound in the theorem above, see Section 9 for the proof.

High-probability bounds. For the high probability bound, we consider the following update rule: $w_{t+1} = \Gamma(w_t + \gamma h_t(w_t))$, where Γ projects on to the set $\mathcal{C} \triangleq \{w \in \mathbb{R}^{2q} \mid \|w\|_2 \leq H\}$.

Assumption 6. *The projection radius H of the set $\mathcal{C}$ satisfies $H > \frac{\|\xi\|_2}{\mu}$, where $\mu = \lambda_{\min}(\frac{\mathbf{M}^\top + \mathbf{M}}{2})$ and ξ is as defined in (6).*

Under the additional projection-related assumption above, we state the high-probability bound for the tail-averaged variant of Algorithm 1. Subsequently, we introduce the regularized mean-variance TD variant to establish high-probability bounds. The following theorem provides a high-probability bound for the unregularized (vanilla) mean-variance TD.

Theorem 3.4. *Suppose Assumptions 1 to 6 hold. Run Algorithm 1 for t iterations with step size β as defined in Theorem 3.2. Then, for any $\delta \in (0, 1]$, we have the following bound for the projected tail-averaged iterate $w_{k+1:t}$:*

$$\mathbb{P}\left(\|w_{k+1:t} - \bar{w}\|_2 \leq \frac{2\tau}{\mu\sqrt{t-k}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{4\exp(-k\beta\mu)}{\beta\mu N}\mathbb{E}\left[\|w_0 - \bar{w}\|_2\right] + \frac{4\tau}{\mu\sqrt{t-k}}\right) \geq 1 - \delta,$$

where $w_0, \bar{w}, \beta$ are defined as in Theorem 3.1, and

$$\tau = \left(2R^2_{\max}\left((\phi^v_{\max})^2 + R^2_{\max}(\phi^u_{\max})^2\right) + 2\left((\phi^v_{\max})^4(1+\gamma)^2 + (\phi^u_{\max})^4\left(1+\gamma^2\right)^2 + 4\gamma^2 R^2_{\max}(\phi^v_{\max})^2(\phi^u_{\max})^2\right)H^2\right)^{\frac{1}{2}}.$$

The following theorem provides a high-probability bound for the regularized mean-variance TD.

Theorem 3.5. *Assume that the conditions in Assumptions 1 to 6 hold. Run Algorithm 1 for t iterations with a step size $\check{\beta} \leq \check{\beta}_{\mathsf{max}}$ as specified in Theorem 3.3. Then, for any $\delta \in (0,1]$, with probability at least $1-\delta$, the projected tail-averaged regularized TD iterate satisfies*

$$\left\|\check{w}_{k+1:t} - \bar{w}_{\mathsf{reg}}\right\|_2 \leq \frac{2\check{\tau}}{(2\mu+\zeta)\sqrt{N}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{4\exp(-k\check{\beta}(2\mu+\zeta))}{\check{\beta}(2\mu+\zeta)N}\mathbb{E}\left\|w_0 - \bar{w}_{\mathsf{reg}}\right\|_2 + \frac{4\check{\tau}}{(2\mu+\zeta)\sqrt{N}}.$$

where N, $\check{w}_0$, $\bar{w}_{\mathsf{reg}}$, and μ are defined as in Theorem 3.3. Moreover, $\check{\tau} = \left(2R_{\mathsf{max}}^2\left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^u)^2\right) + 4\left(\zeta^2 + (\phi_{\mathsf{max}}^v)^4(1+\gamma)^2 + (\phi_{\mathsf{max}}^u)^4(1+\gamma^2)^2 + 4\beta^2 R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^v)^2(\phi_{\mathsf{max}}^u)^2\right)H^2\right)^{\frac{1}{2}}$.

We use a martingale decomposition and Lipschitz concentration of sub-Gaussian random variables to establish the high-probability bounds. This technique has been employed for vanilla TD (Prashanth et al., 2021). Our contribution extends this technique to mean-variance TD and its regularized variant, enabling a universal step size. As in the MSE bound case, owing to the cross terms, a universal step size does not appear to be feasible sans regularization, and we believe this is a useful finding as it deviates from the corresponding result for vanilla TD. In contrast, the authors in (Samsonov et al., 2024) employ Berbee's coupling lemma to arrive at a sub-exponential tail bound.

Discussion: The update rule in (8) represents a Linear Stochastic Approximation (LSA), and mean-variance TD is indeed a special case of the general LSA framework. Several previous works, including Srikant & Ying (2019), provide a finite time analysis for LSA. Their bounds can be applied to (8). However, our analysis differs in the following ways: First, the step size ϵ in Srikant & Ying (2019) depends on the eigenvalues of the transition probability matrix P, which can be difficult to obtain. We alleviate this dependency by employing regularization to achieve a universal step size that is independent of spectral information. Second, we derive explicit constants for the matrix $\mathbf{M}$ (mean-variance TD) instead of the matrix $\mathbf{A}$ (vanilla TD). Third, our analysis focuses on the recursive structure of the error to the projected fixed point, whereas Srikant & Ying (2019) analyze the drift of a Lyapunov function. Finally, Srikant & Ying (2019) provide finite-time bounds for Mean Squared Error, while we additionally establish high-probability bounds.

The current literature on bounds for TD (or more generally, linear stochastic approximation) for Polyak-Ruppert averaging scheme does not achieve $O(1/t)$ bounds, to the best of our knowledge. Instead, with a Polyak-Ruppert stepsize $1/k^{\alpha}$, the bound is $O(1/t^{\alpha})$, with $\alpha < 1$, see (Prashanth et al., 2021). Tail-averaging with a "universal" step size was shown to close this gap for vanilla TD. Our contribution is to show that tail-averaging with universal step size may not be feasible to obtain an $O(1/t)$ for mean-variance TD, while regularization closes this gap.

In Samsonov et al. (2024), the authors provide high-probability bounds for a general linear stochastic approximation algorithm, and specialize them to obtain bounds for the regular TD algorithm. For mean-variance TD (8), we could, in principle, apply the bounds from the aforementioned reference. However, the bound that we derive in Theorem 3.4 enjoys a better dependence on the confidence parameter δ. Specifically, we obtain a $\sqrt{\log(1/\delta)}$ actor, corresponding to a sub-Gaussian tail, while the bounds in Samsonov et al. (2024) feature a $\log(1/\delta)$ factor, which is equivalent to a sub-exponential tail. Furthermore, our result makes all constants clear in the case of mean-variance TD.

4 SPSA-based Actor

In this section, we analyze an actor algorithm based on SPSA-based gradient estimates. Throughout, we consider a parametrized class of stationary randomized policies $\{\pi_\theta,\ \theta \in \mathbb{R}^d\}$. We denote the score function as $\psi_\theta(s,a) = \nabla_\theta \log \pi_\theta(a|s)$. We consider smoothly-parameterized polices, i.e., satisfying the following assumptions:

Assumption 7. $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$ *and* $\theta_1, \theta_2 \in \mathbb{R}^d$, $\exists$ *positive constants* L_ψ, C_ψ *and* C_π *such that (i)* $\left\|\psi_{\theta_1}(s,a) - \psi_{\theta_2}(s,a)\right\|_2 \leq L_\psi \left\|\theta_1 - \theta_2\right\|_2$*; (ii)* $\left\|\psi_\theta(s,a)\right\|_2 \leq C_\psi$*;*
(iii) $\left\|\pi_{\theta_1}(\cdot|s) - \pi_{\theta_2}(\cdot|s)\right\|_{TV} \leq C_\pi \left\|\theta_1 - \theta_2\right\|_2$*, where* $\|\cdot\|_{TV}$ *denotes the total-variation norm.*

In the above, (i) and (ii) imply that score function is smooth and bounded. This generally holds for most commonly used policy classes. Since we asssume finte action space, (iii) holds for any smooth

Algorithm 2: SPSA-based actor with TD critic for mean-variance optimization (MV-SPSA-AC)

Input: Initialize $\theta_0 \in \mathbb{R}^d$, perturbation constant $\{p_t\}$, critic batch size m, actor step size $\{\alpha_t\}$, critic step size $\{\beta_t\}$, number of iterations n, and tail-index k.

for $t \leftarrow 0$ **to** $n-1$ **do**

Generate $\Delta(t) \sim \{\pm 1\}^d$ (symmetric Bernoulli)

```
/* Critic: Obtaining tail-averaged TD iterates for policy evaluation */
```

Run Algorithm 1 for the unperturbed policy π_{θ_t} to compute $w_{k+1:m} = (v_{k+1:m},\, u_{k+1:m})^\top$

Run Algorithm 1 for the perturbed policy $\pi_{\theta_t + p_t\Delta(t)}$ to compute $w^+_{k+1:m} = (v^+_{k+1:m},\, u^+_{k+1:m})^\top$.

```
/* Actor: Estimating SPSA gradients for policy improvement */
```

$\nabla_i \hat{J}(\theta) = \dfrac{\phi_v(s_0)^\top (v^+_{k+1:m} - v_{k+1:m})}{p_t \Delta_i(t)}; \nabla_i \hat{U}(\theta) = \dfrac{\phi_u(s_0)^\top (u^+_{k+1:m} - u_{k+1:m})}{p_t \Delta_i(t)}$

$\theta_{t+1} = \theta_t + \alpha_t(\nabla \hat{J}(\theta_t) - \lambda(\nabla \hat{U}(\theta_t) - 2\hat{J}(\theta_t)\nabla \hat{J}(\theta_t)))$

end for

Output: Final policy θ_R chosen uniformly at random from $\{\theta_1, \ldots, \theta_n\}$

policy. A similar assumption has been made earlier for the analysis of actor-critic algorithms in a risk-neutral RL setting, cf. (Xu et al., 2021). By applying the Lagrangian relaxation procedure (Bertsekas, 1996) to (3), we get the following unconstrained optimization problem for a fixed $\lambda \geq 0$:

$$\min_{\theta} L(\theta) = -V^{\pi_\theta}(s_0) + \lambda(\Lambda^{\pi}_{\theta}(s_0) - c), \tag{15}$$

where $L(\theta)$ represents the Lagrangian function. In this paper, we treat λ as a fixed bias-variance tradeoff parameter, and find a 'good-enough' policy parameter for the problem (15) defined above. For the actor update, we require the gradient of the Lagrangian w.r.t. the policy parameter θ,

$$\nabla_\theta L(\theta) = -\nabla V_\theta(s_0) + \lambda(\nabla U_\theta(s_0) - 2V_\theta(s_0)\nabla V_\theta(s_0)). \tag{16}$$

For notational simplicity, we let $V_\theta(s_0) = J(\theta), U_\theta(s_0) = U(\theta),$ and $\nabla V_\theta(s_0) = \nabla J(\theta)$.

Basic algorithm. We describe the **M**ean **V**ariance **SPSA A**ctor **C**ritic (MV-SPSA-AC) algorithm for mean-variance optimization. Algorithm 2 presents the pseudocode of this algorithm. This algorithm is a variant of the actor-critic algorithm proposed in L.A. & Ghavamzadeh (2016), where the authors provide only asymptotic guarantees. MV-SPSA-AC algorithm deviates from their algorithm by incorporating tail averaging in the TD critic with LFA, and performing a mini-batch update for the SPSA-based actor. More importantly, we perform a finite-sample analysis.

Need for SPSA. The variance of the return we consider lacks a simple linear Bellman equation, unlike the value function in risk-neutral RL. To address this, variance is estimated as the difference between the second moment and the square of the first moment of the return. Since the second moment satisfies a simple linear Bellman equation, this approach makes variance estimation feasible. The policy gradient expression for the square-value function is as follows (see (L.A. & Ghavamzadeh, 2016) for the derivation):

$$\nabla U(\theta) = \tfrac{1}{1-\gamma^2}\Big(\underbrace{\textstyle\sum_{s,a}\tilde{\nu}_\theta(s,a)\nabla \log \pi_\theta(a|s) W_\theta(s,a)}_{T_1(\theta)} + 2\gamma \underbrace{\textstyle\sum_{s,a,s'}\tilde{\nu}_\theta(s,a)P(s'|s,a)\nabla V_\theta(s')}_{T_2(\theta)}\Big). \tag{17}$$

As seen from the expression above, the second term $T_2(\theta)$ requires the gradient $\nabla V_\theta(s')$ for every state $s' \in \mathcal{S}$. An actor-critic algorithm would require an estimate of the value gradient with every possible start state, making it impractical for implementations. SPSA-based gradient estimates offer a viable alternative to overcome this issue. $W_\theta(s,a)$ is equivalent of action-value function for $U(\theta)$.

Actor. The policy parameter θ is updated in the negative direction of gradient of the Lagrangian, with step size α_t as follows:

$$\theta_{t+1} = \theta_t + \alpha_t(\nabla \hat{J}(\theta_t) - \lambda(\nabla \hat{U}(\theta_t) - 2\hat{J}(\theta_t)\nabla \hat{J}(\theta_t))), \tag{18}$$

where (19) is used for computing $\nabla \hat{J}(\theta_t)$ and $\nabla \hat{U}(\theta_t)$ respectively. In a risk-neutral RL setting, the usual recipe for the actor part is to use the policy gradient theorem to form likelihood ratio-based gradient estimates. In L.A. & Ghavamzadeh (2016), it is shown that such an approach does not extend to cover the mean-variance case. The authors there proposed an alternative actor that uses SPSA for gradient estimation. This scheme uses two policy trajectories: one with parameter θ_t and another with a perturbed parameter $\theta_t + p_t\Delta(t)$, denoted by the superscript '+', where $\Delta(t)$ is a d-dimensional vector of independent Rademacher (± 1) random variables. Using these two trajectories, we form estimates of the gradient of the value and square-value functions as follows:

$$\nabla_i \hat{J}(\theta_t) = \frac{\phi_v(s_0)^\top (v^+_{k+1:m} - v_{k+1:m})}{p_t \Delta_i(t)},\ \nabla_i \hat{U}(\theta) = \frac{\phi_u(s_0)^\top (u^+_{k+1:m} - u_{k+1:m})}{p_t \Delta_i(t)}, \tag{19}$$

where $v_{k+1:m}$ and $v^+_{k+1:m}$ are the tail-averaged critic parameters for the value function under the unperturbed (θ_t) and perturbed ($\theta_t + p_t\Delta(t)$) policy parameters, respectively. Here, m is the critic batch size. Similarly, $u_{k+1:m}$ and $u^+_{k+1:m}$ are the tail-averaged critic parameters for the square-value function under the unperturbed and perturbed policy parameters, respectively. We describe next the policy evaluation components in the critic.

Critic. We perform m TD-critic updates to form the estimates for value function $\hat{J}(\theta) = \phi_v(s_0)^\top v_{k+1:m}$ and square-value function $\hat{U}(\theta) = \phi_u(s_0)^\top u_{k+1:m}$, respectively. Further, we perform m updates for the perturbed policy $\theta_t + p_t\Delta(t)$ to form the value and square-value function estimates as $\hat{J}(\theta + p_t\Delta(t)) = \phi_v(s_0)^\top v^+_{k+1:m}$ and $\hat{U}(\theta + p_t\Delta(t)) = \phi_u(s_0)^\top u^+_{k+1:m}$, respectively. We use tail-averaged critic variants for each policy evaluated above.

Main results. For every policy θ, we assume Assumption 1 holds, which implies the existence of the stationary distribution χ_{π_θ}, and scalars $\kappa > 0$ and $\rho \in (0,1)$ such that $\sup_{s\in S} \|\mathbb{P}(s_t \mid s_0 = s) - \chi_{\pi_\theta}\|_{\text{TV}} \leq \kappa\rho^t, \quad \forall t \geq 0$. For the analysis of MV-SPSA-AC algorithm, we need to establish that the Lagrangian $L(\cdot)$ is a smooth function of θ. Further, it can be seen from (16) that , the smoothness of $J(\cdot)$ and $U(\cdot)$ would imply to smoothness of $L(\cdot)$. In a risk-neutral setting, $J(\cdot)$ is the usual objective, and Xu et al. (2021, Proposition 1) established smoothness of $J(\cdot)$ in (20). On the other hand, smoothness of $U(\cdot)$ requires a new proof, and involves significant departures from the one for $J(\cdot)$. The result below states smoothness for $J(\cdot)$ and $U(\cdot)$, with the latter result being a technical contribution of this paper.

Lemma 4.1. *Suppose Assumptions 7 holds. Then, for any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\|_2 \leq L_J \|\theta_1 - \theta_2\|_2,\ \|\nabla U(\theta_1) - \nabla U(\theta_2)\|_2 \leq L_U \|\theta_1 - \theta_2\|, \tag{20}$$

where $L_J = \frac{R_{\max}}{(1-\gamma)}(4C_\nu C_\psi + L_\psi)$, $C_\nu = \frac{1}{2}C_\pi\left(1 + \lceil \log_\rho \kappa^{-1}\rceil + (1-\rho)^{-1}\right)$ *and* $L_U = \frac{1}{1-\gamma^2}\left(\frac{R^2_{\max}}{(1-\gamma)^2}(L_\psi + 4C_\psi C_\nu(1 + \frac{\gamma}{R_{\max}})) + 2L_J\right)$.

We remark that the smoothness result for the square-value function in Lemma 4.1, derived in the context of variance as a risk measure, holds independent significance, as it may prove useful in variants of actor-only or actor-critic methods for mean-variance optimization. Using smoothness of $J(\cdot)$ and $U(\cdot)$, we arrive at the following result.

Lemma 4.2. *Let* $L_o = L_J\left(1 + 2\lambda\frac{R_{\max}}{(1-\gamma)^2} + 2\lambda\left(\frac{R_{\max}C_\psi}{(1-\gamma)^2}\right)^2\right) + \lambda L_U$. *For any $\theta_1, \theta_2 \in \mathbb{R}^d$, we have*

$$\|\nabla L(\theta_1) - \nabla L(\theta_2)\|_2 \leq L_o \|\theta_1 - \theta_2\|_2. \tag{21}$$

The smoothness claim in the result above for the Lagrangian is a key technical contribution, as it serves as a building block for the analysis of the actor update. In particular, this smoothness result facilitates an SGD-type analysis for the actor update. For the analysis of Algorithm 2, we make the following assumption that ensures the value and square-value functions lie in a linear space.

Assumption 8. *For any given policy parameter θ, let $\bar{v}(\theta), \bar{u}(\theta)$ denote solutions to fixed point equations in (5). Then,* $\mathbb{E}[\phi(s_0)^\top \bar{v}(\theta)] = J(\theta), \mathbb{E}[\phi(s_0)^\top \bar{u}(\theta)] = U(\theta)$.

A similar assumption is made in (Kumar et al., 2023, Eq. (13)). Our analysis can be easily extended to include an approximation error term if Assumption 8 does not hold. The main result that establishes stationary convergence of the algorithm MV-SPSA-AC is given below (see Section 11 for a proof sketch and Section 12 for the detailed proof).

Theorem 4.3. *Suppose Assumptions 1 to 8 hold. Run MV-SPSA-AC*[1] *for n iterations with actor step size $\alpha_t \equiv \alpha = 1/n^{3/4}$, perturbation constant $p_t \equiv p = 1/n^{1/4}$, critic batch size $m = n$, and critic step size $\beta \leq \beta_{\mathsf{max}}$ as defined in Theorem 3.1. Let θ_R be chosen uniformly from $\{\theta_1, \ldots, \theta_n\}$. Then,*

$$\mathbb{E}\left[\|\nabla L(\theta_R)\|^2\right] \leq C/n^{1/4},$$

for some constant C that is specified in Section 12.

Remark 1. *We need to account for the biased nature of the SPSA gradient estimators in our analysis. This introduces the perturbation constant p_t, leading to the terms $\mathcal{O}(\frac{1}{p})$, $\mathcal{O}(\frac{1}{p_t^2})$, and $\mathcal{O}(p_t)$. Consequently, we face a trade-off that arises due to the bias in the SPSA gradient estimates, acting as a bottleneck.*

Remark 2. *Eldowa et al. (2022) study the variance of per-step rewards, analyzed as reward volatility (Bisi et al., 2020; Zhang et al., 2021), which is also equivalent to the discount-normalized variance in (Filar et al., 1989). Unlike the variance of the return, this objective lends itself to a REINFORCE-type policy gradient algorithm and does not require a zeroth-order gradient estimation scheme. This is because the gradient of the variance of per-step rewards does not feature a 'problematic' term like $T_2(\cdot)$; instead it only has a term analogous to $T_1(\cdot)$, which can be more easily handled similar to the risk-neutral case.*

The result above establishes the convergence to a stationary point of Lagrangian, and this is significant because $L(\theta)$ encapsulates both the mean and variance of returns. Optimizing $L(\theta)$ ensures a tradeoff between maximizing the value function and minimizing variance. This result is particularly notable as it establishes convergence guarantees for a non-convex function. Mean-variance optimization has been shown to be NP-hard even if the transition dynamics are available, see (Mannor & Tsitsiklis, 2013). Policy-gradient and actor-critic algorithms present a viable alternative where the usual convergence guarantees are to a stationary point. For instance, several policy gradient-type algorithms have been shown to converge to an approximate stationary point in the literature, cf. (Xu et al., 2021; Zhang et al., 2020).

We remark on the sample complexity required for ϵ-accurate convergence of the MV-SPSA-AC algorithm. Theorem 4.3 indicates that the actor loop must run $\Omega(\epsilon^{-4})$ times. However, in each iteration, the critic is executed twice—once for the perturbed and once for the unperturbed trajectories—using n samples per run to estimate the policy gradients. Thus, the total sample complexity for ϵ-accurate convergence is $O(\epsilon^{-4})$. While this represents slow convergence, the use of biased SPSA gradient estimates typically degrades the rate. To the best of our knowledge, finite-sample results for zeroth-order actor-critic methods remain unavailable, even in risk-neutral RL (Lei et al., 2025). Investigating whether sharper analyses or stronger assumptions could improve the convergence rate is an interesting direction for future work.

5 Concluding remarks

We considered a risk-aware discounted reward MDP through mean-variance optimization. Specifically, we analyzed an mean-variance actor-critic algorithm, and derived finite-sample performance guarantees. We first obtained an $O(1/t)$ bound on the convergence of the tail-averaged iterate of the mean-variance TD with LFA. We also obtained a high probability bound that effectively exhibits a sub-Gaussian tail. Next, we employed an SPSA-based actor in conjunction with the above critic, and obtained an $O(n^{-1/4})$ convergence guarantee in the number n of actor iterations.

[1] We employ the un-regularized variant of TD-critic for deriving the bound above. The modification to use the regularized critic for the analysis is straightforward, and we omit the details.

References

Shubhada Agrawal, Prashanth L A, and Siva Theja Maguluri. Policy evaluation for variance in average reward reinforcement learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 471–502. PMLR, 2024. URL https://proceedings.mlr.press/v235/agrawal24a.html.

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Number 4 in Athena Scientific Optimization and Computation Series. Athena Scientific, Belmont, Mass, 1996. ISBN 978-1-886529-04-5. URL https://www.mit.edu/~dimitrib/lagr_mult.html.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation. *Operations Research*, 69(3):950–973, 2021. doi:10.1287/opre.2020.2024.

Shalabh Bhatnagar, Richard S. Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009. ISSN 0005-1098. doi:10.1016/j.automatica.2009.07.008.

Lorenzo Bisi, Luca Sabbioni, Edoardo Vittori, Matteo Papini, and Marcello Restelli. Risk-averse trust region optimization for reward-volatility reduction. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 4583–4589. International Joint Conferences on Artificial Intelligence Organization, July 2020. doi:10.24963/ijcai.2020/632.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite Sample Analyses for TD(0) With Function Approximation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), April 2018. doi:10.1609/aaai.v32i1.12079.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-Time High-Probability Bounds for Polyak–Ruppert Averaged Iterates of Linear Stochastic Approximation. *Mathematics of Operations Research*, 2024. doi:10.1287/moor.2022.0179.

Khaled Eldowa, Lorenzo Bisi, and Marcello Restelli. Finite sample analysis of mean-volatility actor-critic for risk-averse reinforcement learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10028–10066. PMLR, 2022. URL https://proceedings.mlr.press/v151/eldowa22a.html.

M. Fathi and N. Frikha. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. *Electronic Journal of Probability*, 18:1–36, 2013. doi:10.1214/EJP.v18-2586.

J. Filar, L. Kallenberg, and H. Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989. doi:10.1287/moor.14.1.147.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013. doi:10.1137/120880811.

Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *Machine Learning*, 112(7): 2433–2467, July 2023. ISSN 1573-0565. doi:10.1007/s10994-023-06303-2.

Prashanth L.A and Michael C. Fu. Risk-Sensitive Reinforcement Learning via Policy Gradient Search. *Foundations and Trends® in Machine Learning*, 15(5):537–693, 2022. ISSN 1935-8237. doi:10.1561/2200000091.

Prashanth L.A. and Mohammad Ghavamzadeh. Variance-constrained actor-critic algorithms for discounted and average reward MDPs. *Machine Learning*, 105:367–417, 2016. doi:10.1007/s10994-016-5569-5.

C. Lakshminarayanan and C. Szepesvari. Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go? In *International Conference on Artificial Intelligence and Statistics*, volume 84, pp. 1347–1355, 2018. URL `https://proceedings.mlr.press/v84/lakshminarayanan18a.html`.

Yuheng Lei, Yao Lyu, Guojian Zhan, Tao Zhang, Jiangtao Li, Jianyu Chen, Shengbo Eben Li, and Sifa Zheng. Zeroth-Order Actor-Critic: An Evolutionary Framework for Sequential Decision Problems. *IEEE Transactions on Evolutionary Computation*, pp. 1–1, 2025. ISSN 1089-778X, 1089-778X, 1941-0026. doi:10.1109/TEVC.2025.3529503.

S. Mannor and J. N. Tsitsiklis. Algorithmic aspects of mean–variance optimization in Markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013. doi:10.1016/j.ejor.2013.06.019.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. doi:10.2307/2975974.

Aritra Mitra. A simple finite-time analysis of TD learning with linear function approximation. *IEEE Transactions on Automatic Control*, 70(2):1388–1394, 2025. doi:10.1109/TAC.2024.3469328.

W. Mou, C. J. Li, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pp. 2947–2997. PMLR, 2020. URL `https://proceedings.mlr.press/v125/mou20a.html`.

Gandharv Patil, Prashanth L.A., Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5438–5448. PMLR, 2023. URL `https://proceedings.mlr.press/v206/patil23a.html`.

Gandharv Patil, Prashanth L. A., Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation, 2024. URL `https://arxiv.org/abs/2210.05918`.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi:10.1137/0330046.

L. A. Prashanth, N. Korda, and R. Munos. Concentration bounds for temporal difference learning with linear function approximation: The case of batch data and uniform sampling. *Mach. Learn.*, 110(3):559–618, 2021. doi:10.1007/s10994-020-05912-5.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved High-Probability Bounds for the Temporal Difference Learning Algorithm via Exponential Stability. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 4511–4547. PMLR, 2024. URL `https://proceedings.mlr.press/v247/samsonov24a.html`.

M. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pp. 794–802, 1982. doi:10.2307/3213832.

J.C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992. doi:10.1109/9.119632.

R. Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andTD learning. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2803–2830. PMLR, 2019. URL https://proceedings.mlr.press/v99/srikant19a.html.

R. S. Sutton. Learning to Predict by the Methods of Temporal Differences. *Mach. Learn.*, 3:9–44, 1988. doi:10.1007/bf00115009.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Aviv Tamar, Dotan Di Castro, and Shie Mannor. Learning the Variance of the Reward-To-Go. *Journal of Machine Learning Research*, 17(13):1–36, 2016. URL http://jmlr.org/papers/v17/14-335.html.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi:10.1109/9.580874.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4358–4369. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/2e1b24a664f5e9c18f407b2f9c73e821-Paper.pdf.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving Sample Complexity Bounds for (Natural) Actor-Critic Algorithms, 2021. URL https://arxiv.org/abs/2004.12956.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *SIAM Journal on Control and Optimization*, 58(6): 3586–3612, 2020. doi:10.1137/19M1288012.

Shangtong Zhang, Bo Liu, and Shimon Whiteson. Mean-Variance Policy Iteration for Risk-Averse Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12): 10905–10913, May 2021. doi:10.1609/aaai.v35i12.17302.

Supplementary Materials

The following content was not necessarily subject to peer review.

6 Outline of critic analysis

Below, we sketch the proof of Theorem 3.1 to highlight the main ideas and key differences from the standard TD proof. Full proofs of Theorem 3.1 and Theorems 3.2 to 3.5 are provided in Appendices 7–10.

As in proofs of standard TD bounds, we perform a bias-variance decomposition to obtain

$$\mathbb{E}\left[\|z_{t+1}\|^2\right] \leq 2\underbrace{\mathbb{E}\left[\left\|\mathbf{C}^{t:0}z_0\right\|^2\right]}_{z_t^{\text{bias}}} + 2\beta^2\underbrace{\mathbb{E}\left[\left\|\sum_{k=0}^{t}\mathbf{C}^{t:k+1}h_k(\bar{w})\right\|^2\right]}_{z_t^{\text{variance}}}, \tag{22}$$

where $\mathbf{C}^{i:j} = \begin{cases}(\mathbf{I}-\beta\mathbf{M}_i)(\mathbf{I}-\beta\mathbf{M}_{i-1})\ldots(\mathbf{I}-\beta\mathbf{M}_j) & \text{if } i \geq j\\ \mathbf{I} & \text{otherwise.}\end{cases}$

To bound the bias term, we expand the matrix product by one step, yielding

$$\begin{aligned} z_t^{\text{bias}} &= \mathbb{E}\left[\left\|\mathbf{C}^{t:0}z_0\right\|^2\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbf{C}^{t-1:0}z_{t-1}^{\text{bias}}\right)^\top(\mathbf{I}-\beta\mathbf{M}_t)^\top(\mathbf{I}-\beta\mathbf{M}_t)\left(\mathbf{C}^{t-1:0}z_{t-1}^{\text{bias}}\right)\,\middle|\,\mathcal{F}_t\right]\right]. \end{aligned}$$

Next, we establish a result for any $y \in \mathbb{R}^{2q}$ that aids in handling both the bias and variance terms.

$$\begin{aligned} \mathbb{E}\left[y^\top(\mathbf{I}-\beta\mathbf{M}_t)^\top(\mathbf{I}-\beta\mathbf{M}_t)\,y\,\middle|\,\mathcal{F}_t\right] = \|y\|_2^2 - \beta\underbrace{y^\top\mathbb{E}\left[\left(\mathbf{M}_t^\top+\mathbf{M}_t\right)|\mathcal{F}_t\right]y}_{\text{T1}}\\ +\beta^2\underbrace{y^\top\mathbb{E}\left[\mathbf{M}_t^\top\mathbf{M}_t \mid \mathcal{F}_t\right]y}_{\text{T2}} \end{aligned} \tag{23}$$

The term T1 is lower-bounded in a standard manner (as in regular TD), i.e.,

$$y^\top\mathbb{E}\left[\left(\mathbf{M}_t^\top+\mathbf{M}_t\right) \mid \mathcal{F}_t\right]y = y^\top\left(\mathbf{M}^\top+\mathbf{M}\right)y \geq 2\mu\,\|y\|_2^2, \tag{24}$$

where $\mu = \lambda_{\min}(\frac{\mathbf{M}^\top+\mathbf{M}}{2})$ is the minimum eigenvalue of the matrix $\frac{\mathbf{M}+\mathbf{M}^\top}{2}$.

On the other hand, bounding term T2 involves significant deviations. In particular,

$$\begin{aligned} y^\top\mathbb{E}\left[\mathbf{M}_t^\top\mathbf{M}_t \mid \mathcal{F}_t\right]y = \underbrace{v^\top\mathbb{E}\left[\mathbf{a}_t^\top\mathbf{a}_t+\mathbf{c}_t^\top\mathbf{c}_t \mid \mathcal{F}_t\right]v}_{\text{S1}} + \underbrace{u^\top\mathbb{E}\left[\mathbf{b}_t^\top\mathbf{b}_t \mid \mathcal{F}_t\right]u}_{\text{S2}}\\ +\underbrace{v^\top\mathbb{E}\left[\mathbf{c}_t^\top\mathbf{b}_t \mid \mathcal{F}_t\right]u}_{\text{S3}} + \underbrace{u^\top\mathbb{E}\left[\mathbf{b}_t^\top\mathbf{c}_t \mid \mathcal{F}_t\right]v}_{\text{S4}}. \end{aligned} \tag{25}$$

Here, S1 and S2 resemble terms that appear in the finite-sample analysis of regular TD, while S3 and S4 are cross-terms specific to the estimation of the square-value function.

We bound S1, S2 as follows:

$$S1 \leq \left((\phi_{\mathsf{max}}^v)^2(1+\gamma)^2 + 4\gamma^2R_{\max}^2{\phi_{\mathsf{max}}^u}^2\right)v^\top\mathbf{B}v, \tag{26}$$

$$S2 \leq (\phi^u_{\max})^2 \left(1 + 2\gamma^2 + \gamma^4\right) u^\top \mathbf{G} u.$$

In the above, $\mathbf{B}$ and $\mathbf{G}$ are expectations of the outer product of vectors $\phi_v(s_t)$ and $\phi_u(s_t)$ respectively. If the cross-terms were not present, then one could have related T2 to a constant multiple of $v^\top \mathbf{B} v + u^\top \mathbf{G} u$, leading to a universal step size choice, in the spirit of Patil et al. (2024). However, cross-terms present a challenge to this approach, and we bound the S3, S4 cross-terms as follows:

$$S3 + S4 \leq 2(\phi^u_{\max})^2 R_{\max} v^\top \left(\gamma(\mathbf{B}+\mathbf{G}) + \gamma^3(\mathbf{B}+\mathbf{G})\right) u. \tag{27}$$

We overcome the challenge of bounding the cross-terms (S3 and S4) through the following key observations: First, the cross-terms exhibit symmetry and are equal. Consequently, analyzing one term suffices, as the derived upper bound applies to the other term as well. Second, to bound the cross-term, we leverage the following inequality:

$$-v^\top \left(\frac{aa^\top + bb^\top}{2}\right) u \leq v^\top \left(ab^\top\right) u \leq v^\top \left(\frac{aa^\top + bb^\top}{2}\right) u.$$

A similar inequality, also employed in bounding S1 and S2, simplifies the bound in terms of the matrices $\mathbf{B}$ and $\mathbf{G}$, resulting in the expression in (27).

Combining the bounds on S1 to S4 in conjunction with the fact that $v^\top(\mathbf{B}+\mathbf{G})u \leq \frac{\lambda_{\max(\mathbf{B}+\mathbf{G})}}{2} \|y\|_2^2$ (see Lemma 7.2), we obtain the following bound for a step size $\beta \leq \beta_{\max}$ specified in Theorem 3.1 statement:

$$\mathbb{E}\left[y^\top \left(\mathbf{I} - \beta \mathbf{M}_t\right)^\top \left(\mathbf{I} - \beta \mathbf{M}_t\right) y \,\middle|\, \mathcal{F}_t\right] \leq (1 - \beta\mu) \|y\|_2^2. \tag{28}$$

Using the bound above, the bias term in (22) is handled as follows:

$$z_t^{bias} \leq \exp\left(-\beta\mu t\right) \mathbb{E}\left[\|z_0\|^2\right].$$

Using $\|h_k(\bar{w})\|^2 \leq \sigma^2$, we bound the variance term as follows:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{k=0}^{t} \mathbf{C}^{t:k+1} h_k(\bar{w})\right\|_2^2\right] &\leq \sigma^2 \sum_{k=0}^{t} \mathbb{E}\left[\mathbb{E}\left[\|(\mathbf{I} - \beta\mathbf{M}_t)\|^2 \,\middle|\, \mathcal{F}_t\right] \left\|\mathbf{C}^{t-1:k+1}\right\|_2^2\right] \\
&\leq \sigma^2 \sum_{k=0}^{t} (1-\beta\mu)\, \mathbb{E}\left[\left\|\mathbf{C}^{t-1:k+1}\right\|_2^2\right] \\
&\leq \sigma^2 \sum_{k=0}^{t} (1-\beta\mu)^{t-k} \leq \frac{\sigma^2}{\beta\mu}.
\end{aligned} \tag{29}$$

The main claim follows by combining the bounds on the bias and variance terms, followed by straightforward simplifications. The reader is referred to Section 7 for the full proof.

7 Proof of Theorem 3.1

Proof.

Step 1: Bias-variance decomposition

Recall the updates in Algorithm 1 can be rewritten as follows:

$$w_{t+1} = w_t + \beta(r_t \phi_t - \mathbf{M}_t w_t). \tag{30}$$

Defining the centered error as $z_{t+1} = w_{t+1} - \bar{w}$, we obtain

$$z_{t+1} \;\; = w_t - \bar{w} + \beta(r_t\phi_t - \mathbf{M}_t w_t) + \beta \mathbf{M}_t \bar{w} - \beta \mathbf{M}_t \bar{w}$$

$$= (\mathbf{I} - \beta \mathbf{M}_t)(w_t - \bar{w}) + \beta(r_t \phi_t - \mathbf{M}_t \bar{w})$$
$$= (\mathbf{I} - \beta \mathbf{M}_t) z_t + \beta(r_t \phi_t - \mathbf{M}_t \bar{w}).$$

Letting $h_t(w_t) = r_t \phi_t - \mathbf{M}_t w_t$, we have

$$z_{t+1} = (\mathbf{I} - \beta \mathbf{M}_t) z_t + \beta h_t(\bar{w}).$$

Unrolling the equation above, we obtain

$$\begin{aligned} z_{t+1} &= (\mathbf{I} - \beta \mathbf{M}_t)((\mathbf{I} - \beta \mathbf{M}_{t-1}) z_{t-1} + \beta h_{t-1}(\bar{w})) + \beta h_t(\bar{w}) \\ &= (\mathbf{I} - \beta \mathbf{M}_t)(\mathbf{I} - \beta \mathbf{M}_{t-1}) \dots (\mathbf{I} - \beta \mathbf{M}_0) z_0 \ + \beta h_t(\bar{w}) \\ &\quad + \beta (\mathbf{I} - \beta \mathbf{M}_t) h_{t-1}(\bar{w}) \\ &\quad + \beta (\mathbf{I} - \beta \mathbf{M}_t)(\mathbf{I} - \beta \mathbf{M}_{t-1}) h_{t-2}(\bar{w}) \\ &\quad \vdots \\ &\quad + \beta (\mathbf{I} - \beta \mathbf{M}_t)(\mathbf{I} - \beta \mathbf{M}_{t-1}) \dots (\mathbf{I} - \beta \mathbf{M}_1) h_0(\bar{w}). \end{aligned}$$

Define

$$\mathbf{C}^{i:j} = \begin{cases} (\mathbf{I} - \beta \mathbf{M}_i)(\mathbf{I} - \beta \mathbf{M}_{i-1}) \dots (\mathbf{I} - \beta \mathbf{M}_j) & \text{if } i \geq j \\ \mathbf{I} & \text{otherwise.} \end{cases}$$

Using the definition above, we obtain

$$\|z_{t+1}\|^2 = \left\| \mathbf{C}^{t:0} z_0 + \beta \sum_{k=0}^{t} \mathbf{C}^{t:k+1} h_k(\bar{w}) \right\|^2.$$

Taking expectations and using $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we obtain

$$\mathbb{E}[\|z_{t+1}\|^2] \leq 2 z_t^{\text{bias}} + 2\beta^2 z_t^{\text{variance}}, \tag{31}$$

where $z_t^{\text{bias}} = \mathbb{E}\left[\left\|\mathbf{C}^{t:0} z_0\right\|^2\right]$ and $z_t^{\text{variance}} = \mathbb{E}\left[\left\|\sum_{k=0}^{t} \mathbf{C}^{t:k+1} h_k(\bar{w})\right\|^2\right]$.

Step 2: Bounding the bias term

Next, we state and prove a useful lemma that will assist in bounding the bias term in (31).

Lemma 7.1. *Consider a random vector $y \in \mathbb{R}^{2q}$ and let $\mathcal{F}_t$ be sigma-algebra generated by $\{w_0 \dots w_t\}$, For $\beta \leq \beta_{\mathsf{max}}$, we have*

$$\mathbb{E}\left[y^\top (\mathbf{I} - \beta \mathbf{M}_t)^\top (\mathbf{I} - \beta \mathbf{M}_t)\, y \,\middle|\, \mathcal{F}_t\right] \leq (1 - \beta\mu) \|y\|_2^2, \tag{32}$$

$$\mathbb{E}\left[\|(\mathbf{I} - \beta \mathbf{M}_t)\, y\| \mid \mathcal{F}_t\right] \leq \left(1 - \frac{\beta\mu}{2}\right) \|y\|_2, \tag{33}$$

where

$$\beta \leq \beta_{\mathsf{max}} = \frac{\mu}{k}. \tag{34}$$

$\mu = \lambda_{\mathsf{min}\left(\frac{\mathbf{M}^\top + \mathbf{M}}{2}\right)}$ *is the minimum eigenvalue of the matrix* $\frac{\mathbf{M}^\top + \mathbf{M}}{2}$ *and*

$$\begin{aligned} k = \max&\left\{4(\phi^v_{\mathsf{max}})^4 + 4\gamma^2 R^2_{\max} (\phi^u_{\mathsf{max}})^2 (\phi^v_{\mathsf{max}})^2, 4(\phi^u_{\mathsf{max}})^4\right\} \\ &+ 2\gamma R_{\max}((\phi^v_{\mathsf{max}})^2 (\phi^u_{\mathsf{max}})^2 + (\phi^u_{\mathsf{max}})^4). \end{aligned}$$

Proof. To prove the desired result, we split (32) as follows:

$$\mathbb{E}\left[y^\top \left(\mathbf{I}-\beta\mathbf{M}_t\right)^\top \left(\mathbf{I}-\beta\mathbf{M}_t\right) y \;\middle|\; \mathcal{F}_t\right] = \mathbb{E}\left[y^\top \left(\mathbf{I}-\beta\left(\mathbf{M}_t^\top+\mathbf{M}_t\right)+\beta^2\mathbf{M}_t^\top\mathbf{M}_t\right) y \mid \mathcal{F}_t\right]$$
$$= \|y\|_2^2 - \beta\underbrace{y^\top\mathbb{E}\left[\left(\mathbf{M}_t^\top+\mathbf{M}_t\right)\mid\mathcal{F}_t\right]y}_{\text{T1}} + \beta^2\underbrace{y^\top\mathbb{E}\left[\mathbf{M}_t^\top\mathbf{M}_t\mid\mathcal{F}_t\right]y}_{\text{T2}}. \tag{35}$$

We lower-bound the term T1 as follows:

$$y^\top\mathbb{E}\left[\left(\mathbf{M}_t^\top+\mathbf{M}_t\right)\mid\mathcal{F}_t\right]y = y^\top\left(\mathbf{M}^\top+\mathbf{M}\right)y \geq 2\mu\,\|y\|_2^2. \tag{36}$$

Next, we upper bound the term T2 as follows:

$$\mathbf{M}_t^\top\mathbf{M}_t = \begin{pmatrix}\mathbf{a}_t & \mathbf{o}\\ \mathbf{c}_t & \mathbf{b}_t\end{pmatrix}^\top \begin{pmatrix}\mathbf{a}_t & \mathbf{o}\\ \mathbf{c}_t & \mathbf{b}_t\end{pmatrix} = \begin{pmatrix}\mathbf{a}_t^\top\mathbf{a}_t+\mathbf{c}_t^\top\mathbf{c}_t & \mathbf{c}_t^\top\mathbf{b}_t\\ \mathbf{b}_t^\top\mathbf{c}_t & \mathbf{b}_t^\top\mathbf{b}_t\end{pmatrix},$$

Plugging the above in T2, we obtain

$$\begin{aligned}
y^\top\mathbb{E}\left[\mathbf{M}_t^\top\mathbf{M}_t\mid\mathcal{F}_t\right]y &= y^\top\mathbb{E}\left[\begin{pmatrix}\mathbf{a}_t^\top\mathbf{a}_t+\mathbf{c}_t^\top\mathbf{c}_t & \mathbf{c}_t^\top\mathbf{b}_t\\ \mathbf{b}_t^\top\mathbf{c}_t & \mathbf{b}_t^\top\mathbf{b}_t\end{pmatrix}\;\middle|\;\mathcal{F}_t\right]y\\
&= \begin{pmatrix}v^\top & u^\top\end{pmatrix}\mathbb{E}\left[\begin{pmatrix}\mathbf{a}_t^\top\mathbf{a}_t+\mathbf{c}_t^\top\mathbf{c}_t & \mathbf{c}_t^\top\mathbf{b}_t\\ \mathbf{b}_t^\top\mathbf{c}_t & \mathbf{b}_t^\top\mathbf{b}_t\end{pmatrix}\;\middle|\;\mathcal{F}_t\right]\begin{pmatrix}v\\ u\end{pmatrix}\\
&= \underbrace{v^\top\mathbb{E}\left[\mathbf{a}_t^\top\mathbf{a}_t+\mathbf{c}_t^\top\mathbf{c}_t\mid\mathcal{F}_t\right]v}_{\text{S1}} + \underbrace{u^\top\mathbb{E}\left[\mathbf{b}_t^\top\mathbf{b}_t\mid\mathcal{F}_t\right]u}_{\text{S2}}\\
&\quad + \underbrace{v^\top\mathbb{E}\left[\mathbf{c}_t^\top\mathbf{b}_t\mid\mathcal{F}_t\right]u}_{\text{S3}} + \underbrace{u^\top\mathbb{E}\left[\mathbf{b}_t^\top\mathbf{c}_t\mid\mathcal{F}_t\right]v}_{\text{S4}}.
\end{aligned} \tag{37}$$

To upper bound T2, we first establish upper bounds for the terms S1, S2, S3, and S4.

First, we consider the term S1.

$$v^\top\mathbb{E}\left[\mathbf{a}_t^\top\mathbf{a}_t+\mathbf{c}_t^\top\mathbf{c}_t\mid\mathcal{F}_t\right]v = \underbrace{v^\top\mathbb{E}\left[\mathbf{a}_t^\top\mathbf{a}_t\mid\mathcal{F}_t\right]v}_{\text{(a)}} + \underbrace{v^\top\mathbb{E}\left[\mathbf{c}_t^\top\mathbf{c}_t\mid\mathcal{F}_t\right]v}_{\text{(b)}}. \tag{38}$$

We bound (a) in (38) as:

$$\begin{aligned}
&v^\top\mathbb{E}\left[\mathbf{a}_t^\top\mathbf{a}_t\mid\mathcal{F}_t\right]v\\
&= v^\top\mathbb{E}[\left(\phi_v(s_t)\phi_v(s_t)^\top-\gamma\phi_v(s_t)\phi_v(s_{t+1})^\top\right)^\top\left(\phi_v(s_t)\phi_v(s_t)^\top\right.\\
&\qquad\left.-\gamma\phi_v(s_t)\phi_v(s_{t+1})^\top\right)\mid\mathcal{F}_t\,]v\\
&= v^\top\mathbb{E}\big[\phi_v(s_t)\phi_v(s_t)^\top\phi_v(s_t)\phi_v(s_t)^\top-\gamma\phi_v(s_t)\phi_v(s_t)^\top\phi_v(s_t)\phi_v(s_{t+1})^\top\\
&\qquad-\gamma\phi_v(s_{t+1})\phi_v(s_t)^\top\phi_v(s_t)\phi_v(s_t)^\top\\
&\qquad+\gamma^2\phi_v(s_{t+1})\phi_v(s_t)^\top\phi_v(s_t)\phi_v(s_{t+1})^\top\mid\mathcal{F}_t\big]v\\
&\overset{(i)}{=} v^\top\mathbb{E}\big[\,\|\phi_v(s_t)\|_2^2\left(\phi_v(s_t)\phi_v(s_t)^\top-\gamma\underbrace{\left(\phi_v(s_t)\phi_v(s_{t+1})^\top+\phi_v(s_{t+1})\phi_v(s_t)^\top\right)}_{(I)}\right.\\
&\quad\left.+\gamma^2\phi_v(s_{t+1})\phi_v(s_{t+1})^\top\right)\mid\mathcal{F}_t\big]v\\
&\overset{(ii)}{\leq} (\phi_{\max}^v)^2v^\top\mathbb{E}\big[\phi_v(s_t)\phi_v(s_t)^\top+\gamma\left(\phi_v(s_t)\phi_v(s_t)^\top+\phi_v(s_{t+1})\phi_v(s_{t+1})^\top\right)\\
&\quad+\gamma^2\phi_v(s_{t+1})\phi_v(s_{t+1})^\top\mid\mathcal{F}_t\big]v
\end{aligned}$$

$$\leq (\phi^v_{\mathsf{max}})^2 \left(1 + 2\gamma + \gamma^2\right) v^\top \mathbf{B} v, \tag{39}$$

where $\mathbf{B} = \mathbb{E}\left[\phi_v(s_t)\phi_v(s_t)^\top \mid \mathcal{F}_t\right]$. In the above, the inequality in (i) follows from $\|\phi_v(s_t)\|_2^2 = \phi_v(s_t)^\top \phi_v(s_t)$; (ii) follows by applying the bound on the features from Assumption 3 and using the following inequality for term (I) in (i):

$$-v^\top \left(\frac{aa^\top + bb^\top}{2}\right) v \leq v^\top \left(ab^\top\right) v \leq v^\top \left(\frac{aa^\top + bb^\top}{2}\right) v. \tag{40}$$

The final inequality in (39) follows by using the following equivalent forms for $\mathbf{B}$:

$$\begin{aligned}\mathbf{B} &= \mathbb{E}\left[\phi_v(s_t)\phi_v(s_t)^\top \mid \mathcal{F}_t\right] = \mathbb{E}\left[\phi_v(s_{t+1})\phi_v(s_{t+1})^\top \;\middle|\; \mathcal{F}_t\right] = \mathbb{E}^{\mathcal{X},\mathbf{P}}\left[\phi_v(s_t)\phi_v(s_t)^\top\right] \\ &= \mathbb{E}^{\mathcal{X},\mathbf{P}}\left[\phi_v(s_{t+1})\phi_v(s_{t+1})^\top\right].\end{aligned} \tag{41}$$

The equivalences above hold from the i.i.d observation model (Assumption 5).

Next, We bound (b) in (38) as:

$$\begin{aligned}v^\top \mathbb{E}\left[\mathbf{c}_t{}^\top \mathbf{c}_t \;\middle|\; \mathcal{F}_t\right] v &= v^\top \mathbb{E}\left[\left(-2\gamma r_t \phi_u(s_t)\phi_v(s_{t+1})^\top\right)^\top \left(-2\gamma r_t \phi_u(s_t)\phi_v(s_{t+1})^\top\right) \mid \mathcal{F}_t\right] v \\ &= 4\gamma^2 v^\top \mathbb{E}\left[r_t^2 \phi_v(s_{t+1})\phi_u(s_t)^\top \phi_u(s_t)\phi_v(s_{t+1})^\top \;\middle|\; \mathcal{F}_t\right] v \\ &\overset{(i)}{=} 4\gamma^2 v^\top \mathbb{E}\left[r_t^2 \left\|\phi_u(s_t)\right\|_2^2 \phi_v(s_{t+1})\phi_v(s_{t+1})^\top \;\middle|\; \mathcal{F}_t\right] v \\ &\overset{(ii)}{\leq} 4\gamma^2 R^2_{\mathsf{max}} (\phi^u_{\mathsf{max}})^2 v^\top \mathbf{B} v,\end{aligned} \tag{42}$$

where (i) follows from $\|\phi_u(s_t)\|_2^2 = \phi_u(s_t)^\top \phi_u(s_t)$ and (ii) follows from bound on rewards (Assumption 4) and value of $\mathbf{B}$ in (41).

Combining (39) and (42), we obtain the upper bound for S1 as follows:

$$v^\top \mathbb{E}\left[\mathbf{a}_t^\top \mathbf{a}_t + \mathbf{c}_t^\top \mathbf{c}_t \;\middle|\; \mathcal{F}_t\right] v \leq \left((\phi^v_{\mathsf{max}})^2 (1+\gamma)^2 + 4\gamma^2 R^2_{\mathsf{max}} (\phi^u_{\mathsf{max}})^2\right) v^\top \mathbf{B} v. \tag{43}$$

Next, we upper bound S2 in (37) as follows:

$$\begin{aligned}&u^\top \mathbb{E}\left[\mathbf{b}_t^\top \mathbf{b}_t \;\middle|\; \mathcal{F}_t\right] u \\ &= u^\top \mathbb{E}[\left(\phi_u(s_t)\phi_u(s_t)^\top - \gamma^2 \phi_u(s_t)\phi_u(s_{t+1})^\top\right)^\top (\phi_u(s_t)\phi_u(s_t)^\top \\ &\qquad - \gamma^2 \phi_u(s_t)\phi_u(s_{t+1})^\top) \mid \mathcal{F}_t] u \\ &= u^\top \mathbb{E}[\phi_u(s_t)\phi_u(s_t)^\top \phi_u(s_t)\phi_u(s_t)^\top - \gamma^2 \big(\phi_u(s_t)\phi_u(s_t)^\top \phi_u(s_t)\phi_u(s_{t+1})^\top \\ &\qquad + \phi_u(s_{t+1})\phi_u(s_t)^\top \phi_u(s_t)\phi_u(s_t)^\top\big) \\ &\qquad + \gamma^4 (\phi_u(s_{t+1})\phi_u(s_t)^\top \phi_u(s_t)\phi_u(s_{t+1})^\top) \mid \mathcal{F}_t] u \\ &\overset{(i)}{=} u^\top \mathbb{E}\big[\left\|\phi_u(s_t)\right\|_2^2 \big(\phi_u(s_t)\phi_u(s_t)^\top - \gamma^2 \underbrace{\left(\phi_u(s_t)\phi_u(s_{t+1})^\top + \phi_u(s_{t+1})\phi_u(s_t)^\top\right)}_{(II)} \\ &\qquad + \gamma^4 \phi_u(s_{t+1})\phi_u(s_{t+1})^\top\big) \mid \mathcal{F}_t\big] u \\ &\overset{(ii)}{\leq} (\phi^u_{\mathsf{max}})^2 u^\top \mathbb{E}\big[\phi_u(s_t)\phi_u(s_t)^\top + \gamma^2 \left(\phi_u(s_t)\phi_u(s_t)^\top + \phi_u(s_{t+1})\phi_u(s_{t+1})^\top\right) \\ &\qquad\qquad + \gamma^4 \phi_u(s_{t+1})\phi_u(s_{t+1})^\top \mid \mathcal{F}_t\big] u \\ &\overset{(iii)}{\leq} (\phi^u_{\mathsf{max}})^2 \left(1 + 2\gamma^2 + \gamma^4\right) u^\top \mathbf{G} u,\end{aligned} \tag{44}$$

where $\mathbf{G} = \mathbb{E}\left[\phi_u(s_t)\phi_u(s_t)^\top \;\middle|\; \mathcal{F}_t\right]$. In the above, the inequality in (i) follows from $\|\phi_u(s_t)\|_2^2 = \phi_u(s_t)^\top \phi_u(s_t)$; (ii) follows from bound on features (Assumption 3) and applying the inequality (40) to (II); and (44) follows by bound on features (Assumption 5).

The inequality in (44) follows by following equivalent forms of $\mathbf{G}$:

$$\mathbf{G} = \mathbb{E}\left[\phi_u(s_t)\phi_u(s_t)^\top \mid \mathcal{F}_t\right] = \mathbb{E}\left[\phi_u(s_{t+1})\phi_u(s_{t+1})^\top \mid \mathcal{F}_t\right] = \mathbb{E}^{\chi,\mathbf{P}}\left[\phi_u(s_t)\phi_u(s_t)^\top\right] \\ = \mathbb{E}^{\chi,\mathbf{P}}\left[\phi_u(s_{t+1})\phi_u(s_{t+1})^\top\right]. \tag{45}$$

The equivalences above hold from the i.i.d observation model (Assumption 5).

We observe that scalars S3 and S4 in (37) are equal, i.e.,

$$v^\top \mathbb{E}\left[\mathbf{c}_t^\top \mathbf{b}_t \mid \mathcal{F}_t\right] u = u^\top \mathbb{E}\left[\mathbf{b}_t^\top \mathbf{c}_t \mid \mathcal{F}_t\right] v.$$

We establish upper bound for S3 in (37) as follows:

$$\begin{aligned}
&v^\top \mathbb{E}\left[\mathbf{c}_t^\top \mathbf{b}_t\right] u \\
&= v^\top \mathbb{E}\big[-2\gamma r_t \phi_v(s_{t+1})\phi_u(s_t)^\top \phi_u(s_t)\phi_u(s_t)^\top \\
&\qquad + 2\gamma^3 r_t \phi_v(s_{t+1})\phi_u(s_t)^\top \phi_u(s_t)\phi_u(s_{t+1})^\top \mid \mathcal{F}_t\big] u \\
&\overset{(i)}{=} \|\phi_u(s_t)\|_2^2 \, v^\top \mathbb{E}[-2r_t\gamma \underbrace{\phi_v(s_{t+1})\phi_u(s_t)^\top}_{(III)} + 2r_t\gamma^3 \underbrace{\phi_v(s_{t+1})\phi_u(s_{t+1})^\top}_{(IV)} \mid \mathcal{F}_t] u \\
&\overset{(ii)}{\leq} (\phi^u_{\mathsf{max}})^2 R_{\mathsf{max}} v^\top \mathbb{E}\Big[\gamma\left(\phi_v(s_{t+1})\phi_v(s_{t+1})^\top + \phi_u(s_t)\phi_u(s_t)^\top\right) \\
&\qquad + \gamma^3(\phi_v(s_{t+1})\phi_v(s_{t+1})^\top + \phi_u(s_{t+1})\phi_u(s_{t+1})^\top) \mid \mathcal{F}_t\Big] u \\
&\leq (\phi^u_{\mathsf{max}})^2 R_{\mathsf{max}} v^\top \left(\gamma(\mathbf{B}+\mathbf{G}) + \gamma^3(\mathbf{B}+\mathbf{G})\right) u,
\end{aligned} \tag{46}$$

where (i) follows from $\|\phi_u(s_t)\|_2^2 = \phi_u(s_t)^\top \phi_u(s_t)$; (ii) follows from bounds on features and rewards (Assumptions 3 and 4) and applying the inequality below to the coefficients of γ (III) with $(a = \phi_v(s_{t+1}), b = \phi_u(s_t))$ and γ^3 (IV) with $(a = \phi_v(s_{t+1}), b = \phi_u(s_{t+1}))$ respectively.

$$-v^\top \left(\frac{aa^\top + bb^\top}{2}\right) u \leq v^\top \left(ab^\top\right) u \leq v^\top \left(\frac{aa^\top + bb^\top}{2}\right) u.$$

(46) follows by using values of matrices $\mathbf{B}$ (41) and $\mathbf{G}$ (45).

Substituting (43)–(46) in (37), we determine the upper bound for T2 as follows:

$$\begin{aligned}
y^\top \mathbb{E}\left[\mathbf{M}_t^\top \mathbf{M}_t \mid \mathcal{F}_t\right] y \leq &\left((\phi^v_{\mathsf{max}})^2 (1+\gamma)^2 + 4\gamma^2 R^2_{\mathsf{max}} (\phi^u_{\mathsf{max}})^2\right) v^\top \mathbf{B} v \\
&+ (\phi^u_{\mathsf{max}})^2 \left(1+\gamma^2\right)^2 u^\top \mathbf{G} u \\
&+ 2(\phi^u_{\mathsf{max}})^2 R_{\mathsf{max}} (\gamma(1+\gamma^2)) v^\top (\mathbf{B}+\mathbf{G}) u.
\end{aligned} \tag{47}$$

Next, we state and prove a useful result to simplify (47) further.

Lemma 7.2. *For any $y = (v, u)^\top \in \mathbb{R}^{2|S|}$ and matrix $\mathbf{B}+\mathbf{G}$ defined in (46), we have*

$$v^\top (\mathbf{B}+\mathbf{G}) u \leq \frac{\lambda_{\mathsf{max}(\mathbf{B}+\mathbf{G})}}{2} \|y\|_2^2.$$

Proof. We have

$$\begin{aligned}
v^\top (\mathbf{B}+\mathbf{G}) u &\overset{(a)}{\leq} \|v\|_{\mathbf{B}+\mathbf{G}} \|u\|_{\mathbf{B}+\mathbf{G}} \\
&\overset{(b)}{\leq} \sqrt{v^\top (\mathbf{B}+\mathbf{G}) v} \sqrt{u^\top (\mathbf{B}+\mathbf{G}) u} \\
&\overset{(c)}{\leq} \lambda_{\mathsf{max}(\mathbf{B}+\mathbf{G})} \sqrt{\|v\|_2^2 \|u\|_2^2}
\end{aligned}$$

$$\overset{(d)}{\leq} \lambda_{\max(\mathbf{B}+\mathbf{G})} \frac{\|v\|_2^2 + \|u\|_2^2}{2}$$
$$\overset{(e)}{\leq} \frac{\lambda_{\max(\mathbf{B}+\mathbf{G})}}{2} \|y\|_2^2 \,,$$

where (a) follows by Cauchy-Schwarz inequality; (b) follows by definition of the weighted norm; (c) follows by Rayleigh quotient theorem for a symmetric real matrix $\mathbf{Q}$, i.e., $x^\top \mathbf{Q} x \leq \lambda_{\max(\mathbf{Q})} \|x\|_2^2$; (d) follows by AM-GM inequality; and (e) follows by definition of $\|y\|_2^2 = \|v\|_2^2 + \|u\|_2^2$. □

Substituting the upper bounds obtained for T1 (36) and T2 (47) in (35), we get

$$\begin{aligned}
\mathbb{E}\left[y^\top (\mathbf{I} - \beta \mathbf{M}_t)^\top (\mathbf{I} - \beta \mathbf{M}_t)\, y \,\middle|\, \mathcal{F}_t\right] &= \|y\|_2^2 - \beta \underbrace{y^\top \mathbb{E}\left[\left(\mathbf{M}_t^\top + \mathbf{M}_t\right) | \mathcal{F}_t\right] y}_{\text{T1}} \\
&\quad + \beta^2 \underbrace{y^\top \mathbb{E}\left[\mathbf{M}_t^\top \mathbf{M}_t \,\middle|\, \mathcal{F}_t\right] y}_{\text{T2}} \\
&\leq \|y\|_2^2 - 2\beta\mu \|y\|_2^2 + \beta^2 \Bigg(\left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R_{\max}^2 (\phi^u_{\max})^2 \right) v^\top \mathbf{B} v \\
&\quad + (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 u^\top \mathbf{G} u + 2(\phi^u_{\max})^2 R_{\max}(\gamma(1+\gamma^2)) v^\top (\mathbf{B}+\mathbf{G})\, u \Bigg) \\
&\overset{(i)}{\leq} \|y\|_2^2 - 2\beta\mu \|y\|_2^2 + \beta^2 \Bigg(\left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R_{\max}^2 (\phi^u_{\max})^2 \right) \lambda_{\max(\mathbf{B})} \|v\|_2^2 \\
&\quad + (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \lambda_{\max(\mathbf{G})} \|u\|_2^2 + (\phi^u_{\max})^2 R_{\max}(\gamma(1+\gamma^2)) \lambda_{\max(\mathbf{B}+\mathbf{G})} \|y\|_2^2 \Bigg) \\
&\leq \|y\|_2^2 - 2\beta\mu \|y\|_2^2 + \beta^2 \Bigg(\max\Big\{ \left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R_{\max}^2 (\phi^u_{\max})^2 \right) \lambda_{\max(\mathbf{B})}, \\
&\quad (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \lambda_{\max(\mathbf{G})} \Big\} \|y\|_2^2 + (\phi^u_{\max})^2 R_{\max}(\gamma(1+\gamma^2)) \lambda_{\max(\mathbf{B}+\mathbf{G})} \|y\|_2^2 \Bigg) \\
&\leq \|y\|_2^2 - \beta \Bigg(2\mu - \beta \Bigg(\max\Big\{ \left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R_{\max}^2 (\phi^u_{\max})^2 \right) \lambda_{\max(\mathbf{B})}, \\
&\quad (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \lambda_{\max(\mathbf{G})} \Big\} + (\phi^u_{\max})^2 R_{\max}(\gamma(1+\gamma^2)) \lambda_{\max(\mathbf{B}+\mathbf{G})} \Bigg) \Bigg) \|y\|_2^2 \\
&\overset{(ii)}{\leq} \|y\|_2^2 - \beta \Bigg(2\mu - \beta \Bigg(\max\left\{ 4(\phi^v_{\max})^4 + 4\gamma^2 R_{\max}^2 (\phi^u_{\max})^2 (\phi^v_{\max})^2, 4(\phi^u_{\max})^4 \right\} \\
&\quad + 2\gamma R_{\max} \left((\phi^v_{\max})^2 (\phi^u_{\max})^2 + (\phi^u_{\max})^4 \right) \Bigg) \Bigg) \|y\|_2^2 \\
&\leq (1 - \beta\mu) \|y\|_2^2 \,,
\end{aligned} \tag{48}$$

where (i) follows from Lemma 7.2 and using $x^\top \mathbf{Q} x \leq \lambda_{\max(\mathbf{Q})} \|x\|_2^2$; (ii) follows using $\lambda_{\max(\mathbf{B})} \leq (\phi^v_{\max})^2, \lambda_{\max(\mathbf{G})} \leq (\phi^u_{\max})^2$, and $\lambda_{\max(\mathbf{B}+\mathbf{G})} \leq (\phi^v_{\max})^2 + (\phi^u_{\max})^2$ as $\mathbf{B}, \mathbf{G}$ are outer products of vectors $\phi_v(s_t)$ and $\phi_u(s_t)$ respectively; (48) follows by choosing $\beta \leq \beta_{\max}$.

Re-writing (48) in norm form gives:

$$\mathbb{E}\left[y^\top (\mathbf{I} - \beta \mathbf{M}_t)^\top (\mathbf{I} - \beta \mathbf{M}_t)\, y \,\middle|\, \mathcal{F}_t\right] = \mathbb{E}\left[\|(\mathbf{I} - \beta \mathbf{M}_t) y\|^2 \,\middle|\, \mathcal{F}_t\right] \leq (1 - \beta\mu) \|y\|_2^2 \,. \tag{49}$$

Taking square root on both sides of (49) yields the second claim

$$\mathbb{E}\left[\|(\mathbf{I} - \beta \mathbf{M}_t) y\| \mid \mathcal{F}_t\right] \leq (1 - \beta\mu)^{\frac{1}{2}} \|y\|_2 \leq \left(1 - \frac{\beta\mu}{2}\right) \|y\|_2 \,, \tag{50}$$

where (50) follows by using the inequality $(1-x)^{\frac{1}{2}} \leq 1 - \frac{x}{2}$, for $x \geq 0$ with $x = \beta\mu$. □

Now, we bound the bias term as follows:

$$\begin{aligned}
z_t^{\text{bias}} &= \mathbb{E}\left[\left\|\mathbf{C}^{t:0} z_0\right\|^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\mathbf{C}^{t-1:0} z_{t-1}^{\text{bias}}\right)^\top (\mathbf{I} - \beta\mathbf{M}_t)^\top (\mathbf{I} - \beta\mathbf{M}_t)\left(\mathbf{C}^{t-1:0} z_{t-1}^{\text{bias}}\right) | \mathcal{F}_t\right]\right] \\
&\overset{(i)}{\leq} (1-\beta\mu)\,\mathbb{E}\left[\left\|\mathbf{C}^{t-1:0} z_{t-1}^{\text{bias}}\right\|^2\right] \\
&\leq (1-\beta\mu)^t\,\mathbb{E}\left[\|z_0\|^2\right] \qquad (51)\\
&\leq \exp\left(-\beta\mu t\right)\mathbb{E}\left[\|z_0\|^2\right], \qquad (52)
\end{aligned}$$

where (i) follows by Lemma 7.1; (51) follows by unrolling the recursion and using Lemma 7.1 repetitively; and (52) follows by using the inequality below

$$(1-\beta\mu)^t = \exp(t\log(1-\beta\mu)) \leq \exp(-\beta\mu t).$$

Step 3: Bounding the variance term For the variance bound, we require an upper bound for $\|h_t(\bar{w})\|^2$, which we derive below.

$$\begin{aligned}
\|h_t(\bar{w})\|^2 &= \|r_t\phi(s_t) - \mathbf{M}_t\bar{w}\|^2 \\
&\overset{(a)}{\leq} 2\|r_t\phi(s_t)\|^2 + 2\|\mathbf{M}_t\bar{w}\|_2^2 \\
&\overset{(b)}{\leq} 2R_{\mathsf{max}}^2\left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^u)^2\right) + 2\|\mathbf{M}_t\|^2\|\bar{w}\|_2^2 \\
&\overset{(c)}{\leq} 2R_{\mathsf{max}}^2\left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^u)^2\right) + 2((\phi_{\mathsf{max}}^v)^4(1+\gamma)^2 + (\phi_{\mathsf{max}}^u)^4\left(1+\gamma^2\right)^2 \\
&\quad + 4\gamma^2 R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^v)^2(\phi_{\mathsf{max}}^u)^2)\|\bar{w}\|_2^2 \\
&= \sigma^2, \qquad (53)
\end{aligned}$$

where (a) follows using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; (b) follows by bounds on features and rewards (Assumptions 3 and 4); and (c) follows by expanding the upper bound on $\|\mathbf{M}_t\|^2$.

Next, we bound the variance term in (31) as follows:

$$\begin{aligned}
z_t^{\text{variance}} &= \mathbb{E}\left[\left\|\sum_{k=0}^{t}\mathbf{C}^{t:k+1}h_k(\bar{w})\right\|_2^2\right] \\
&\overset{(a)}{\leq} \sum_{k=0}^{t}\mathbb{E}\left[\left\|\mathbf{C}^{t:k+1}h_k(\bar{w})\right\|_2^2\right] \\
&\overset{(b)}{\leq} \sum_{k=0}^{t}\mathbb{E}\left[\left\|\mathbf{C}^{t:k+1}\right\|^2\|h_k(\bar{w})\|^2\right] \\
&\overset{(c)}{\leq} \sigma^2\sum_{k=0}^{t}\mathbb{E}\left[\left\|\mathbf{C}^{t:k+1}\right\|_2^2\right] \\
&\overset{(d)}{\leq} \sigma^2\sum_{k=0}^{t}\mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{C}^{t:k+1}\right\|_2^2 \,\middle|\, \mathcal{F}_t\right]\right] \\
&\overset{(e)}{\leq} \sigma^2\sum_{k=0}^{t}\mathbb{E}\left[\mathbb{E}\left[\left\|(\mathbf{I}-\beta\mathbf{M}_t)\mathbf{C}^{t-1:k+1}\right\|_2^2 \,\middle|\, \mathcal{F}_t\right]\right]
\end{aligned}$$

$$\overset{(f)}{\leq} \sigma^2 \sum_{k=0}^{t} \mathbb{E}\left[\mathbb{E}\left[\|(\mathbf{I}-\beta\mathbf{M}_t)\|^2 \;\middle|\; \mathcal{F}_t\right] \left\|\mathbf{C}^{t-1:k+1}\right\|_2^2\right]$$
$$\overset{(g)}{\leq} \sigma^2 \sum_{k=0}^{t} (1-\beta\mu)\, \mathbb{E}\left[\left\|\mathbf{C}^{t-1:k+1}\right\|_2^2\right]$$
$$\overset{(h)}{\leq} \sigma^2 \sum_{k=0}^{t} (1-\beta\mu)^{t-k}$$
$$\overset{(i)}{\leq} \frac{\sigma^2}{\beta\mu}, \tag{54}$$

where (a) follows by triangle inequality and linearity of expectations; (b) follows by using the inequality $\|\mathbf{A}x\| \leq \|\mathbf{A}\| \, \|x\|$; (c) follows by a bound on $\|h_k(\bar{w})\|^2$ in (53); (d) follows by the tower property of conditional expectations; (e) follows by unrolling the product of matrices $\mathbf{C}^{t:k+1}$ by one factor; (f) follows by using the inequality $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$; (g) follows by Lemma 7.1; (h) follows by unrolling the the product of matrices; and (i) follows by computing the upper bound for the finite geometric series.

Step 4: Clinching argument

The main claim follows by combining the bounds on the bias (52) and variance (54) terms in (31) as follows:

$$\mathbb{E}[\|z_{t+1}\|^2] \leq 2z_t^{\text{bias}} + 2\beta^2 z_t^{\text{variance}}$$
$$\leq 2\exp(-\beta\mu t)\, \mathbb{E}\left[\|z_0\|^2\right] + \frac{2\beta\sigma^2}{\mu}.$$

□

8 Proof of Theorem 3.2

Proof.

Step 1: Bias-variance decomposition for tail averaging

The tail averaged error when starting at $k+1$, at time t is given by

$$z_{k+1:t} = \frac{1}{N}\sum_{i=k+1}^{k+N} z_i = \frac{1}{t-k}\sum_{i=k+1}^{t} z_i.$$

By taking expectations, $\|z_{k+1:t}\|^2$ can be expressed as:

$$\mathbb{E}\left[\|z_{k+1:t}\|_2^2\right] = \frac{1}{N^2}\sum_{i,j=k+1}^{k+N} \mathbb{E}\left[z_i^\top z_j\right]$$
$$\overset{(a)}{\leq} \frac{1}{N^2}\left(\sum_{i=k+1}^{k+N} \mathbb{E}\left[\|z_i\|_2^2\right] + 2\sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N} \mathbb{E}\left[z_i^\top z_j\right]\right), \tag{55}$$

where (a) follows from isolating the diagonal and off-diagonal terms.

Next, we state and prove a result that bounds the second term in (55).

Lemma 8.1. *For all $i \geq 1$, we have*

$$\sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N} \mathbb{E}\left[z_i^\top z_j\right] \leq \frac{2}{\beta\mu}\sum_{i=k+1}^{k+N} \mathbb{E}\left[\|z_i\|_2^2\right]. \tag{56}$$

Proof.

$$\begin{aligned}
\sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N}\mathbb{E}\left[z_i^\top z_j\right] &\overset{(a)}{=} \sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N}\mathbb{E}\left[z_i^\top(\mathbf{C}^{j:i+1}z_i + \beta\sum_{l=i+1}^{j-i-1}\mathbf{C}^{j:l+1}h_l(\bar{w}))\right]\\
&\overset{(b)}{=} \sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N}\mathbb{E}\left[z_i^\top \mathbf{C}^{j:i+1}z_i\right]\\
&\overset{(c)}{\le} \sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N}\mathbb{E}\left[\|z_i\|\,\mathbb{E}\left[\|\mathbf{C}^{j:i+1}z_i\| \mid \mathcal{F}_j\right]\right]\\
&\overset{(d)}{\le} \sum_{i=k+1}^{k+N-1}\sum_{j=i+1}^{k+N}\left(1-\frac{\beta\mu}{2}\right)^{j-i}\mathbb{E}\left[\|z_i\|_2^2\right]\\
&\le \sum_{i=k+1}^{k+N}\mathbb{E}\left[\|z_i\|_2^2\right]\sum_{j=i+1}^{\infty}\left(1-\frac{\beta\mu}{2}\right)^{j-i}\\
&\overset{(e)}{\le} \frac{2}{\beta\mu}\sum_{i=k+1}^{k+N}\mathbb{E}\left[\|z_i\|_2^2\right],
\end{aligned}$$

where (a) follows by expanding z_j using (31); (b) follows from the observation that

$$\mathbb{E}[h_t(\bar{w}) \mid \mathcal{F}_t] = \mathbb{E}[r_t\phi_t - \mathbf{M}_t\bar{w} \mid \mathcal{F}_t] = \xi - \mathbf{M}\bar{w} = 0;$$

(c) follows by using Cauchy-Schwarz inequality and tower property of expectations; (d) follows from a repetitive application of Lemma 7.1; and (e) follows by computing the limit of the infinite geometric series. □

Substituting the result of Lemma 8.1 in (55), we obtain

$$\begin{aligned}
\mathbb{E}\left[\|z_{k+1:t}\|_2^2\right] &\le \frac{1}{N^2}\left(\sum_{i=k+1}^{k+N}\mathbb{E}\left[\|z_i\|_2^2\right] + \frac{4}{\beta\mu}\sum_{i=k+1}^{k+N}\mathbb{E}\left[\|z_i\|_2^2\right]\right)\\
&= \frac{1}{N^2}\left(1+\frac{4}{\beta\mu}\right)\sum_{i=k+1}^{k+N}\mathbb{E}\left[\|z_i\|_2^2\right]\\
&\overset{(a)}{\le} \underbrace{\frac{2}{N^2}\left(1+\frac{4}{\beta\mu}\right)\sum_{i=k+1}^{k+N}z_i^{\mathsf{bias}}}_{z_{k+1,N}^{\mathsf{bias}}} + \underbrace{\frac{2}{N^2}\left(1+\frac{4}{\beta\mu}\right)\beta^2\sum_{i=k+1}^{k+N}z_i^{\mathsf{variance}}}_{z_{k+1:t}^{\mathsf{variance}}},
\end{aligned} \tag{57}$$

where (a) follows from the bias-variance decomposition of $\mathbb{E}[\|z_i\|_2^2]$ in (31).

Step 2: Bounding the bias

First term, $z_{k+1:t}^{\mathsf{bias}}$ in (57) is bounded as follows:

$$\begin{aligned}
z_{k+1:t}^{\mathsf{bias}} &\le \frac{2}{N^2}\left(1+\frac{4}{\beta\mu}\right)\sum_{i=k+1}^{\infty}z_i^{\mathsf{bias}}\\
&\overset{(a)}{\le} \frac{2}{N^2}\left(1+\frac{4}{\beta\mu}\right)\sum_{i=k+1}^{\infty}(1-\beta\mu)^i\mathbb{E}\left[\|z_0\|_2^2\right]\\
&\overset{(b)}{=} \frac{2\mathbb{E}\left[\|z_0\|_2^2\right]}{\beta\mu N^2}(1-\beta\mu)^{k+1}\left(1+\frac{4}{\beta\mu}\right),
\end{aligned}$$

where (a) follows from (51), which provides a bound on z_i^{bias}; (b) follows from the bound on the summation of a geometric series.

Step 4: Bounding the variance

Next, the second term $z_{k+1:t}^{\text{variance}}$ in (57) is bounded as follows:

$$\begin{aligned} z_{k+1:t}^{\text{variance}} &\overset{(a)}{\leq} \frac{2\beta^2}{N^2}\left(1+\frac{4}{\beta\mu}\right)\sum_{i=k+1}^{k+N}\frac{\sigma^2}{\beta\mu} \\ &\leq \frac{2\beta^2}{N^2}\left(1+\frac{4}{\beta\mu}\right)\sum_{i=0}^{N}\frac{\sigma^2}{\beta\mu} \\ &= \left(1+\frac{4}{\beta\mu}\right)\frac{2\beta\sigma^2}{\mu N}, \end{aligned}$$

where (a) follows from (54), which provides a bound on z_i^{variance}.

Step 5: Clinching argument

Finally substituting the bounds on $z_{k+1:t}^{\text{bias}}$ and $z_{k+1:t}^{\text{variance}}$ in (57), we get

$$\begin{aligned} \mathbb{E}[\|z_{k+1:t}\|_2^2] &\leq \left(1+\frac{4}{\beta\mu}\right)\left(\frac{2}{\beta\mu N^2}(1-\beta\mu)^{k+1}\mathbb{E}[\|z_0\|_2^2]+\frac{2\beta\sigma^2}{\mu N}\right), \\ &\overset{(a)}{\leq} \left(1+\frac{4}{\beta\mu}\right)\left(\frac{2\exp(-k\beta\mu)}{\beta\mu N^2}\mathbb{E}[\|z_0\|_2^2]+\frac{2\beta\sigma^2}{\mu N}\right) \\ &\overset{(b)}{\leq} \frac{10\exp(-k\beta\mu)}{\beta^2\mu^2N^2}\mathbb{E}\left[\|z_0\|_2^2\right]+\frac{10\sigma^2}{\mu^2 N}, \end{aligned}$$

where (a) follows from $(1+x)^y = \exp(y\log(1+x)) \leq \exp(xy)$; (b) uses $\beta\mu < 1$ as $\beta \leq \beta_{\max}$ defined in Theorem 3.1, which implies that $1+\frac{4}{\beta\mu} \leq \frac{5}{\beta\mu}$. □

9 Proof of Theorem 3.3

For proving Theorem 3.3, we first establish an upper bound on the mean squared error (MSE) of the difference between the tail-averaged TD iterate and the regularized TD fixed point. The result below provides this bound, which we subsequently use to prove Theorem 3.3.

Theorem 9.1. *Suppose Assumptions 1 to 4 hold. Let $\breve{w}_{k+1:t} = \frac{1}{N}\sum_{i=k+1}^{k+N}\breve{w}_i$ denote the tail-averaged regularized iterate with $N = t-k$. Suppose the step size $\breve{\beta}$ satisfies*

$$\begin{aligned} \breve{\beta} \leq& \breve{\beta}_{\max} = \frac{\zeta}{\breve{c}}, \text{ where} \\ \breve{c} =& \zeta^2 + 2\zeta\left((\phi^v_{\max})^4(1+\gamma)^2 + (\phi^u_{\max})^4(1+\gamma^2)^2 + 4\gamma^2 R^2_{\max}(\phi^v_{\max})^2(\phi^u_{\max})^2\right)^{\frac{1}{2}} \\ &+ \max\left\{4(\phi^v_{\max})^4 + 4\gamma^2 R^2_{\max}(\phi^u_{\max})^2(\phi^v_{\max})^2, 4(\phi^u_{\max})^4\right\} \\ &+ 2\gamma R_{\max}((\phi^v_{\max})^2(\phi^u_{\max})^2 + (\phi^u_{\max})^4). \end{aligned}$$

Then,

$$\mathbb{E}\left[\|\breve{w}_{k+1:t} - \bar{w}_{\text{reg}}\|_2^2\right] \leq \frac{10\exp\left(-k\breve{\beta}(2\mu+\zeta)\right)}{\breve{\beta}^2\left(2\mu+\zeta\right)^2N^2}\mathbb{E}\left[\|\breve{w}_0 - \bar{w}_{\text{reg}}\|_2^2\right] + \frac{10\breve{\sigma}^2}{(2\mu+\zeta)^2N}, \tag{58}$$

where $N = t-k$, $\mu = \lambda_{\min}(\frac{\mathbf{M}^\top+\mathbf{M}}{2})$, and

$$\begin{aligned} \breve{\sigma}^2 =& 2R^2_{\max}\left((\phi^v_{\max})^2 + R^2_{\max}(\phi^u_{\max})^2\right) + 4\big(\zeta^2 + (\phi^v_{\max})^4(1+\gamma)^2 + (\phi^u_{\max})^4\left(1+\gamma^2\right)^2 \\ &+ 4\gamma^2R^2_{\max}(\phi^v_{\max})^2(\phi^u_{\max})^2\big)\|\bar{w}_{\text{reg}}\|_2^2 \end{aligned} \tag{59}$$

Proof. Our proof incorporates techniques from Patil et al. (2024). However, as described earlier, the analysis of mean-variance TD involves additional cross-terms, which necessitate significant deviations in the proof.

Step 1: Bias-variance decomposition with regularization

For regularized TD, we solve the following linear system:

$$-(\mathbf{M}+\zeta\mathbf{I})\bar{w}_{\text{reg}}+\xi=0, \tag{60}$$

The corresponding TD updates in Algorithm 1 to solve (60) would be:

$$\begin{aligned} v_{t+1} &= (\mathbf{I}-\check{\beta}\zeta)v_t+\check{\beta}\;\check{\delta}_t\;\phi_v(s_t), \\ u_{t+1} &= (\mathbf{I}-\check{\beta}\zeta)u_t+\check{\beta}\;\check{\epsilon}_t\;\phi_u(s_t), \end{aligned} \tag{61}$$

where $\check{\delta}_t, \check{\epsilon}_t$ are defined as

$$\begin{aligned} \check{\delta}_t =& r(s_t,a_t)+\gamma\check{v}_t^\top\phi_v(s_{t+1})-\check{v}_t^\top\phi_v(s_t) \\ \check{\epsilon}_t =& r(s_t,a_t)^2+2\gamma r(s_t,a_t)\;\check{v}_t^\top\phi_v(s_{t+1})+\gamma^2\check{u}_t^\top\phi_u(s_{t+1})-\check{u}_t^\top\phi_u(s_t). \end{aligned} \tag{62}$$

We rewrite the updates in the alternative form as:

$$\check{w}_{t+1}=\check{w}_t+\check{\beta}(r_t\phi_t-(\zeta\mathbf{I}+\mathbf{M}_t)\check{w}_t), \tag{63}$$

where $\mathbf{M}_t, r_t, \phi_t$ are defined in (8).

Letting $\check{h}_t(w_t)=r_t\phi_t-(\zeta\mathbf{I}+\mathbf{M}_t)\check{w}_t$, we have

$$\check{w}_{t+1}=\check{w}_t+\check{\beta}\check{h}_t(\check{w}_t). \tag{64}$$

As in the case of 'vanilla' mean-variance TD, we arrive at a one-step recursion for the centered error $\check{z}_{t+1}=\check{w}_{t+1}-\bar{w}_{\mathsf{reg}}$ as follows:

$$\begin{aligned} \check{z}_{t+1} &= \check{w}_t-\bar{w}_{\mathsf{reg}}+\check{\beta}(r_t\phi_t-\mathbf{M}_t\check{w}_t)+\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_t)\bar{w}_{\mathsf{reg}}-\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_t)\bar{w}_{\mathsf{reg}} \\ &= (\mathbf{I}-\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_t))(w_t-\bar{w}_{\mathsf{reg}})+\check{\beta}(r_t\phi_t-(\zeta\mathbf{I}+\mathbf{M}_t)\bar{w}_{\mathsf{reg}}) \\ &= (\mathbf{I}-\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_t))z_t+\check{\beta}\check{h}_t(\bar{w}_{\mathsf{reg}}). \end{aligned} \tag{65}$$

Unrolling the equation above, we obtain

$$\check{z}_{t+1}=\check{\mathbf{C}}^{t:0}\check{z}_0+\check{\beta}\sum_{k=0}^{t}\check{\mathbf{C}}^{t:k+1}\check{h}_k(\bar{w}_{\mathsf{reg}}), \tag{66}$$

where

$$\check{\mathbf{C}}^{i:j}=\begin{cases}(\mathbf{I}-\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_i))(\mathbf{I}-\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_{i-1}))\ldots(\mathbf{I}-\check{\beta}(\zeta\mathbf{I}+\mathbf{M}_j)) & \text{if } i\geq j \\ \mathbf{I} & \text{otherwise.}\end{cases}$$

Taking expectations and using $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$, we obtain,

$$\begin{aligned} \mathbb{E}\left[\|\check{z}_{t+1}\|^2\right] &\leq 2\mathbb{E}\left(\left\|\check{\mathbf{C}}^{t:0}\check{z}_0\right\|^2\right)+2\check{\beta}^2\mathbb{E}\left[\left\|\sum_{k=0}^{t}\check{\mathbf{C}}^{t:k+1}\check{h}_k(\bar{w}_{\mathsf{reg}})\right\|^2\right], \\ &\leq 2\check{z}_t^{\text{bias}}+2\check{\beta}^2\check{z}_t^{\text{variance}}, \end{aligned} \tag{67}$$

where $\check{z}_t^{\text{bias}}=\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t:0}\check{z}_0\right\|^2\right]$ and $\check{z}_t^{\text{variance}}=\mathbb{E}\left[\left\|\sum_{k=0}^{t}\check{\mathbf{C}}^{t:k+1}\check{h}_k(\bar{w}_{\mathsf{reg}})\right\|^2\right]$.

Step 2: Bounding the bias term

Before we bound the bias term, we first state and prove some useful lemmas.

Lemma 9.2.

$$\|\mathbf{M}\| \leq \left((\phi^v_{\mathsf{max}})^4(1+\gamma)^2 + (\phi^u_{\mathsf{max}})^4(1+\gamma^2)^2 + 4\gamma^2 R^2_{\mathsf{max}}(\phi^v_{\mathsf{max}})^2(\phi^u_{\mathsf{max}})^2\right)^{\frac{1}{2}}.$$

Proof. Recall that $\mathbf{M} = \mathbb{E}[\mathbf{M}_t \mid \mathcal{F}_t]$ where

$$\mathbf{M}_t \triangleq \begin{pmatrix} \mathbf{a}_t & \mathbf{o} \\ \mathbf{c}_t & \mathbf{b}_t \end{pmatrix} \text{ with } \mathbf{a}_t \triangleq \phi_v(s_t)\phi_v(s_t)^\top - \gamma\phi_v(s_t)\phi_v(s_{t+1})^\top,$$
$$\mathbf{b}_t \triangleq \phi_u(s_t)\phi_u(s_t)^\top - \gamma^2\phi_u(s_t)\phi_u(s_{t+1})^\top,$$
$$\mathbf{c}_t \triangleq -2\gamma r_t\phi_u(s_t)\phi_v(s_{t+1})^\top.$$

We bound the norm of the matrices $\mathbf{a}_t, \mathbf{b}_t, \mathbf{c}_t$ using bound on features and rewards (Assumptions 3 and 4) as:

$$\|\mathbf{a}_t\| \leq (1+\gamma)(\phi^v_{\mathsf{max}})^2, \|\mathbf{b}_t\| \leq (1+\gamma^2)(\phi^u_{\mathsf{max}})^2, \|\mathbf{c}_t\| \leq 2\gamma R_{\mathsf{max}}\phi^v_{\mathsf{max}}\phi^u_{\mathsf{max}}. \tag{68}$$

Next, we derive the result as follows:

$$\begin{aligned}\|\mathbf{M}\| = \|\mathbb{E}[\mathbf{M}_t \mid \mathcal{F}_t]\| &\overset{(i)}{\leq} \mathbb{E}[\|\mathbf{M}_t\| \mid \mathcal{F}_t] \\ &\overset{(ii)}{\leq} \left\| \begin{pmatrix} (1+\gamma)(\phi^v_{\mathsf{max}})^2 & 0 \\ 2\gamma R_{\mathsf{max}}\phi^v_{\mathsf{max}}\phi^u_{\mathsf{max}} & (1+\gamma^2)(\phi^u_{\mathsf{max}})^2 \end{pmatrix} \right\|_F \\ &\overset{(iii)}{\leq} \left((\phi^v_{\mathsf{max}})^4(1+\gamma)^2 + (\phi^u_{\mathsf{max}})^4(1+\gamma^2)^2 + 4\gamma^2 R^2_{\mathsf{max}}(\phi^v_{\mathsf{max}})^2(\phi^u_{\mathsf{max}})^2\right)^{\frac{1}{2}},\end{aligned}$$

where (i) follows by Jensen's inequality, (ii) follows by (68), and (iii) follows by expanding the Frobenius norm.

□

Lemma 9.3. *For any $\check{y} \in \mathbb{R}^{2q}$ measurable w.r.t $\mathcal{F}_t$ and $\check{\beta} \leq \check{\beta}_{\mathsf{max}}$ as in Theorem 9.1. The following holds:*

$$\mathbb{E}\left[\check{y}(\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))^\top(\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))\check{y} \mid \mathcal{F}_t\right] \leq \left(1 - \check{\beta}(2\mu + \zeta)\right)\|\check{y}\|_2^2,$$
$$\mathbb{E}\left[\left\|(\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))\check{y}\right\|_2 \mid \mathcal{F}_t\right] \leq \left(1 - \frac{\check{\beta}(2\mu + \zeta)}{2}\right)\|\check{y}\|_2.$$

Proof. Notice that

$$\begin{aligned}&\mathbb{E}\left[\check{y}^\top(\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))^\top(\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))\check{y} \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\check{y}^\top(\mathbf{I} - 2\check{\beta}\zeta\mathbf{I} - \check{\beta}(\mathbf{M}_t + \mathbf{M}_t^\top)) + \check{\beta}^2(\zeta^2\mathbf{I} + \zeta(\mathbf{M}_t + \mathbf{M}_t^\top) + \mathbf{M}_t^\top\mathbf{M}_t)\check{y} \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\check{y}^\top\check{y} \mid \mathcal{F}_t\right] - \check{\beta}\mathbb{E}\left[\check{y}^\top 2\zeta\mathbf{I}\check{y} \mid \mathcal{F}_t\right] - \check{\beta}\underbrace{\check{y}^\top\mathbb{E}\left[\mathbf{M}_t^\top + \mathbf{M}_t \mid \mathcal{F}_t\right]\check{y}}_{\text{Term 1}} \\ &\quad + \check{\beta}^2\underbrace{\check{y}^\top\mathbb{E}\left[\mathbf{M}_t^\top\mathbf{M}_t \mid \mathcal{F}_t\right]\check{y}}_{\text{Term 2}} + \check{\beta}^2\zeta\underbrace{\check{y}^\top\mathbb{E}\left[\mathbf{M}_t + \mathbf{M}_t^\top \mid \mathcal{F}_t\right]\check{y}}_{\text{Term 3}} + \check{\beta}^2\mathbb{E}\left[\check{y}^\top\zeta^2\mathbf{I}\check{y} \mid \mathcal{F}_t\right].\end{aligned} \tag{69}$$

We bound Term 1 in (69) as follows:

$$\check{y}^\top\mathbb{E}\left[\mathbf{M}_t^\top + \mathbf{M}_t \mid \mathcal{F}_t\right]\check{y} = \check{y}^\top(\mathbf{M}^\top + \mathbf{M})\check{y} \overset{(i)}{\geq} 2\mu\|\check{y}\|_2^2, \tag{70}$$

where (i) follows from the fact that Assumption 2 implies $\mathbf{M} + \mathbf{M}^\top$ has a minimum positive eigenvalue $\mu = \lambda_{\mathsf{min}}(\frac{\mathbf{M}^\top + \mathbf{M}}{2})$.
We bound Term 2 in (69) using the bound for T2 in (47) as follows:

$$\check{y}^\top\mathbb{E}[\mathbf{M}_t^\top\mathbf{M}_t \mid \mathcal{F}_t]\check{y}$$

$$\leq \left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R^2_{\max} (\phi^u_{\max})^2 \right) \check{v}^\top \mathbf{B} \check{v}$$
$$+ (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \check{u}^\top \mathbf{G} \check{u} + 2(\phi^u_{\max})^2 R_{\max} (\gamma(1+\gamma^2)) \check{v}^\top \left(\mathbf{B}+\mathbf{G}\right) \check{u}.$$

757 We bound Term 3 in (69) as follows:

$$\check{y}^\top \mathbb{E}[\mathbf{M}_t + \mathbf{M}_t^\top \mid \mathcal{F}_t] \check{y} \leq \left\| \mathbb{E}[\mathbf{M}_t + \mathbf{M}_t^\top \mid \mathcal{F}_t] \right\| \|\check{y}\|^2 \leq \left\| \mathbf{M} + \mathbf{M}^\top \right\| \|\check{y}\|^2$$
$$\overset{(i)}{\leq} 2\left((\phi^v_{\max})^4 (1+\gamma)^2 + (\phi^u_{\max})^4 (1+\gamma^2)^2 + 4\gamma^2 R^2_{\max} (\phi^v_{\max})^2 (\phi^u_{\max})^2 \right)^{\frac{1}{2}} \|\check{y}\|^2,$$

758 where (i) follows by Lemma 9.2.

759 Substituting the bounds for Terms 1–3 in (69), we obtain

$$\mathbb{E}[\check{y}^\top (\mathbf{I} - \check{\beta}(\zeta \mathbf{I} + \mathbf{M}_t))^\top (\mathbf{I} - \check{\beta}(\zeta \mathbf{I} + \mathbf{M}_t)) \check{y} \mid \mathcal{F}_t]$$
$$\leq \mathbb{E}[\check{y}^\top \check{y} | \mathcal{F}_t] - \check{\beta} \mathbb{E}[\check{y}^\top 2\zeta \mathbf{I} \check{y} \mid \mathcal{F}_t] - \check{\beta} 2\mu \|\check{y}\|^2$$
$$+ \check{\beta}^2 \Big(\left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R^2_{\max} (\phi^u_{\max})^2 \right) \check{v}^\top \mathbf{B} \check{v}$$
$$+ (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \check{u}^\top \mathbf{G} \check{u} + 2(\phi^u_{\max})^2 R_{\max} (\gamma(1+\gamma^2)) \check{v}^\top \left(\mathbf{B}+\mathbf{G}\right) \check{u} \Big)$$
$$+ \check{\beta}^2 \Big(2\big((\phi^v_{\max})^4 (1+\gamma)^2 + (\phi^u_{\max})^4 (1+\gamma^2)^2 + 4\gamma^2 R^2_{\max} (\phi^v_{\max})^2 (\phi^u_{\max})^2 \big)^{\frac{1}{2}} \|\check{y}\|^2 \Big)$$
$$+ \check{\beta}^2 \mathbb{E}[\check{y}^\top \zeta^2 \mathbf{I} \check{y} \mid \mathcal{F}_t].$$
$$\overset{(i)}{\leq} \|\check{y}\|_2^2 (1 - 2\check{\beta}(\mu+\zeta)) + \check{\beta}^2 \Big(\left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R^2_{\max} (\phi^u_{\max})^2 \right) \lambda_{\max(\mathbf{B})} \|\check{v}\|_2^2$$
$$+ (\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \lambda_{\max(\mathbf{G})} \|\check{u}\|_2^2 + (\phi^u_{\max})^2 R_{\max} (\gamma(1+\gamma^2)) \lambda_{\max(\mathbf{B}+\mathbf{G})} \|\check{y}\|_2^2$$
$$+ 2\zeta \big((\phi^v_{\max})^4 (1+\gamma)^2 + (\phi^u_{\max})^4 (1+\gamma^2)^2$$
$$+ 4\gamma^2 R^2_{\max} (\phi^v_{\max})^2 (\phi^u_{\max})^2 \big)^{\frac{1}{2}} \|\check{y}\|^2 + \zeta^2 \|\check{y}\|_2^2 \Big)$$
$$\leq \Big(1 - \check{\beta} \Big(2\mu + 2\zeta - \check{\beta} \big(\max\big\{ \left((\phi^v_{\max})^2 (1+\gamma)^2 + 4\gamma^2 R^2_{\max} (\phi^u_{\max})^2 \right) \lambda_{\max(\mathbf{B})},$$
$$(\phi^u_{\max})^2 \left(1+\gamma^2\right)^2 \lambda_{\max(\mathbf{G})} \big\} + (\phi^u_{\max})^2 R_{\max} (\gamma(1+\gamma^2)) \lambda_{\max(\mathbf{B}+\mathbf{G})}$$
$$+ \zeta^2 + 2\zeta \big((\phi^v_{\max})^4 (1+\gamma)^2 + (\phi^u_{\max})^4 (1+\gamma^2)^2$$
$$+ 4\gamma^2 R^2_{\max} (\phi^v_{\max})^2 (\phi^u_{\max})^2 \big)^{\frac{1}{2}} \big) \Big) \Big) \|\check{y}\|_2^2$$
$$\leq \Big(1 - \check{\beta} \Big(2\mu + 2\zeta - \check{\beta} \Big(\max \Big\{ 4(\phi^v_{\max})^4 + 4\gamma^2 R^2_{\max} (\phi^u_{\max})^2 (\phi^v_{\max})^2, 4(\phi^u_{\max})^4 \Big\}$$
$$+ 2\gamma R_{\max} \left((\phi^v_{\max})^2 (\phi^u_{\max})^2 + (\phi^u_{\max})^4 \right)$$
$$+ \zeta^2 + 2\zeta \big((\phi^v_{\max})^4 (1+\gamma)^2 + (\phi^u_{\max})^4 (1+\gamma^2)^2$$
$$+ 4\gamma^2 R^2_{\max} (\phi^v_{\max})^2 (\phi^u_{\max})^2 \big)^{\frac{1}{2}} \Big) \Big) \Big) \|\check{y}\|_2^2$$
$$\overset{(ii)}{\leq} (1 - \check{\beta}(2\mu+\zeta)) \|\check{y}\|_2^2, \tag{71}$$

760 where (i) follows from Lemma 7.2 and using $x^\top \mathbf{Q} x \leq \lambda_{\max(\mathbf{Q})} \|x\|_2^2$, and (ii) follows by choosing
761 $\check{\beta} \leq \check{\beta}_{\max}$.

762 Taking square root on both sides of (71) leads to

$$\mathbb{E}\left[\left\| (\mathbf{I} - \check{\beta}(\zeta \mathbf{I} + \mathbf{M}_t)) \check{y} \right\| \,\middle|\, \mathcal{F}_t \right] \leq (1 - \check{\beta}(2\mu+\zeta))^{\frac{1}{2}} \|\check{y}\|_2$$
$$\overset{(i)}{\leq} \left(1 - \frac{\check{\beta}(2\mu+\zeta)}{2} \right) \|\check{y}\|_2, \tag{72}$$

where (i) follows by using the inequality $(1-x)^{\frac{1}{2}} \leq 1 - \frac{x}{2}$, for $x \geq 0$ with $x = \check{\beta}(2\mu + \zeta)$. □

Now, we bound the bias term in (67) as follows:

$$\begin{aligned}
\check{z}_t^{\text{bias}} &= \mathbb{E}\left[\left\|\check{\mathbf{C}}^{t:0}\check{z}_0\right\|^2\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left(\check{\mathbf{C}}^{t-1:0}\check{z}_{t-1}^{\text{bias}}\right)^\top (\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))^\top (\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))(\check{\mathbf{C}}^{t-1:0}\check{z}_{t-1}^{\text{bias}})\right] \middle| \mathcal{F}_t\right] \\
&\overset{(i)}{\leq} \left(1 - \check{\beta}(2\mu + \zeta)\right)\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t-1:0}\check{z}_{t-1}^{\text{bias}}\right\|^2\right] \\
&\overset{(ii)}{\leq} \left(1 - \check{\beta}(2\mu + \zeta)\right)^t \mathbb{E}\left[\left\|\check{z}_0\right\|^2\right] && (73) \\
&\overset{(iii)}{\leq} \exp\left(-\check{\beta}(2\mu + \zeta)t\right)\mathbb{E}\left[\left\|\check{z}_0\right\|^2\right], && (74)
\end{aligned}$$

where (i) follows by Lemma 9.3, (ii) follows by unrolling the recursion and using Lemma 9.3 repetitively, and (iii) follows by using the inequality

$$(1 - \beta(2\mu + \zeta))^t = \exp(t\log(1 - \beta(2\mu + \zeta)) \leq \exp(-\beta(2\mu + \zeta)t).$$

Step 3: Bounding the variance term

Before we find an upper bound for the variance term, we upper bound on $\|h_t(\bar{w}_{\text{reg}})\|^2$ as follows:

$$\begin{aligned}
\left\|\check{h}_t(\bar{w}_{\text{reg}})\right\|^2 &= \|r_t\phi(s_t) - (\zeta\mathbf{I} + \mathbf{M}_t)\bar{w}_{\text{reg}}\|^2 \\
&\overset{(a)}{\leq} 2\|r_t\phi(s_t)\|^2 + 2\|(\zeta\mathbf{I} + \mathbf{M}_t)\bar{w}_{\text{reg}}\|_2^2 \\
&\overset{(b)}{\leq} 2R_{\max}^2\left((\phi_{\max}^v)^2 + R_{\max}^2(\phi_{\max}^u)^2\right) + 2\|\zeta\mathbf{I} + \mathbf{M}_t\|^2\|\bar{w}_{\text{reg}}\|_2^2 \\
&\overset{(c)}{\leq} 2R_{\max}^2\left((\phi_{\max}^v)^2 + R_{\max}^2(\phi_{\max}^u)^2\right) + 4\big(\zeta^2 + (\phi_{\max}^v)^4(1+\gamma)^2 \\
&\quad + (\phi_{\max}^u)^4\left(1+\gamma^2\right)^2 + 4\gamma^2R_{\max}^2(\phi_{\max}^v)^2(\phi_{\max}^u)^2\big)\|\bar{w}_{\text{reg}}\|_2^2 && (75) \\
&= \check{\sigma}^2, && (76)
\end{aligned}$$

where (a) follows using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, (b) follows using bound on features, rewards (Assumptions 3 and 4), and (c) follows by bound on $\mathbf{M}$ (Lemma 9.2) and using the inequality $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

Next, we bound the variance term in (67) as follows:

$$\begin{aligned}
\check{z}_t^{\text{variance}} &= \mathbb{E}\left[\left\|\sum_{k=0}^{t}\check{\mathbf{C}}^{t:k+1}\check{h}_k(\bar{w}_{\text{reg}})\right\|_2^2\right] \\
&\overset{(a)}{\leq} \sum_{k=0}^{t}\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t:k+1}\check{h}_k(\bar{w}_{\text{reg}})\right\|_2^2\right] \\
&\overset{(b)}{\leq} \sum_{k=0}^{t}\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t:k+1}\right\|^2\left\|\check{h}_k(\bar{w}_{\text{reg}})\right\|^2\right] \\
&\overset{(c)}{\leq} \check{\sigma}^2\sum_{k=0}^{t}\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t:k+1}\right\|_2^2\right] \\
&\overset{(d)}{\leq} \check{\sigma}^2\sum_{k=0}^{t}\mathbb{E}\left[\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t:k+1}\right\|_2^2|\mathcal{F}_t\right]\right]
\end{aligned}$$

$$\overset{(e)}{\le} \check{\sigma}^2 \sum_{k=0}^{t} \mathbb{E}\left[\mathbb{E}\left[\left\|(\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t))\check{\mathbf{C}}^{t-1:k+1}\right\|_2^2 |\mathcal{F}_t\right]\right]$$

$$\overset{(f)}{\le} \check{\sigma}^2 \sum_{k=0}^{t} \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{I} - \check{\beta}(\zeta\mathbf{I} + \mathbf{M}_t)\right\|^2 |\mathcal{F}_t\right] \left\|\check{\mathbf{C}}^{t-1:k+1}\right\|_2^2\right]$$

$$\overset{(g)}{\le} \check{\sigma}^2 \sum_{k=0}^{t} (1 - \check{\beta}(2\mu + \zeta))\mathbb{E}\left[\left\|\check{\mathbf{C}}^{t-1:k+1}\right\|_2^2\right]$$

$$\overset{(h)}{\le} \check{\sigma}^2 \sum_{k=0}^{t} (1 - \check{\beta}(2\mu + \zeta))^{t-k}$$

$$\overset{(i)}{\le} \frac{\check{\sigma}^2}{\check{\beta}(2\mu + \zeta)}, \tag{77}$$

where (a) follows by triangle inequality and linearity of expectations; (b) follows by using the inequality $\|\mathbf{A}x\| \le \|\mathbf{A}\| \|x\|$; (c) follows by a bound on $\left\|\check{h}_k(\bar{w}_{\mathsf{reg}})\right\|^2$, (d) follows by the tower property of conditional expectations; (e) follows by unrolling the product of matrices $\check{\mathbf{C}}^{t:k+1}$ by one-time step; (f) follows by using the inequality $\|\mathbf{AB}\| \le \|\mathbf{A}\| \|\mathbf{B}\|$; (g) follows by Lemma 9.3; (h) follows by unrolling the product of matrices; and (i) follows by computing the upper bound for the finite geometric series.

Step 4: Tail Averaging Using the parallel arguments from Section 8, we derive the bounds for tail-averaged error bounds for bias and variance terms as follows:

4 (a) Bias-variance decomposition for tail averaging

The tail averaged error when starting at $k+1$, at time t is given by

$$\check{z}_{k+1:t} = \frac{1}{N} \sum_{i=k+1}^{k+N} \check{z}_i.$$

By taking expectations, $\|\check{z}_{k+1:t}\|^2$ can be expressed as:

$$\mathbb{E}\left[\|\check{z}_{k+1:t}\|_2^2\right] = \frac{1}{N^2} \sum_{i,j=k+1}^{k+N} \mathbb{E}\left[\check{z}_i^\top \check{z}_j\right]$$
$$\overset{(a)}{\le} \frac{1}{N^2}\left(\sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right] + 2\sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \mathbb{E}\left[\check{z}_i^\top \check{z}_j\right]\right), \tag{78}$$

where (a) follows from isolating the diagonal and off-diagonal terms.

Next, we state and prove Lemma 9.4 to bound the second term in terms of the first term in (78).

Lemma 9.4. *For all $i \ge 1$, we have*

$$\sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \mathbb{E}\left[\check{z}_i^\top \check{z}_j\right] \le \frac{2}{\check{\beta}(2\mu + \zeta)} \sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right]. \tag{79}$$

Proof.

$$\sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \mathbb{E}\left[\check{z}_i^\top \check{z}_j\right] \overset{(a)}{=} \sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \mathbb{E}\left[\check{z}_i^\top (\check{\mathbf{C}}^{j:i+1}\check{z}_i + \check{\beta} \sum_{l=i+1}^{j-i-1} \check{\mathbf{C}}^{j:l+1}\check{h}_l(\bar{w}_{\mathsf{reg}}))\right]$$

$$
\begin{aligned}
&\overset{(b)}{=} \sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \mathbb{E}\left[\check{z}_i^\top \check{\mathbf{C}}^{j:i+1} z_i\right] \\
&\overset{(c)}{\leq} \sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\| \, \mathbb{E}[\|\check{\mathbf{C}}^{j:i+1}\check{z}_i\| \, |\mathcal{F}_j]\right] \\
&\overset{(d)}{\leq} \sum_{i=k+1}^{k+N-1} \sum_{j=i+1}^{k+N} \left(1 - \frac{\check{\beta}(2\mu+\zeta)}{2}\right)^{j-i} \mathbb{E}\left[\|\check{z}_i\|_2^2\right] \\
&\leq \sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right] \sum_{j=i+1}^{\infty} \left(1 - \frac{\check{\beta}(2\mu+\zeta)}{2}\right)^{j-i} \\
&\overset{(e)}{\leq} \frac{2}{\check{\beta}(2\mu+\zeta)} \sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right],
\end{aligned}
$$

where (a) follows by expanding z_j using (66), (b) follows from the observation that

$$\mathbb{E}[\check{h}_t(\bar{w}_{\mathsf{reg}}) \mid \mathcal{F}_t] = \mathbb{E}[r_t \phi_t - (\zeta \mathbf{I} + \mathbf{M}_t)\bar{w}_{\mathsf{reg}} \mid \mathcal{F}_t] = \xi - (\mathbf{M} + \zeta \mathbf{I})\bar{w}_{\mathsf{reg}} = 0,$$

(c) follows by using Cauchy-Schwarz inequality and tower property of expectations, (d) follows from a repetitive application of Lemma 9.3, and (e) follows by computing the limit of the infinite geometric series. □

Substituting the result of Lemma 9.4 in (78), we obtain

$$
\begin{aligned}
\mathbb{E}\left[\|\check{z}_{k+1:t}\|_2^2\right] &\leq \frac{1}{N^2}\left(\sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right] + \frac{4}{\check{\beta}(2\mu+\zeta)} \sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right]\right) \\
&= \frac{1}{N^2}\left(1 + \frac{4}{\check{\beta}(2\mu+\zeta)}\right) \sum_{i=k+1}^{k+N} \mathbb{E}\left[\|\check{z}_i\|_2^2\right] \\
&\overset{(a)}{\leq} \underbrace{\frac{2}{N^2}\left(1 + \frac{4}{\beta(2\mu+\zeta)}\right) \sum_{i=k+1}^{k+N} \check{z}_i^{\mathsf{bias}}}_{\check{z}^{\mathsf{bias}}_{k+1,N}} + \underbrace{\frac{2}{N^2}\left(1 + \frac{4}{\beta(2\mu+\zeta)}\right) \check{\beta}^2 \sum_{i=k+1}^{k+N} \check{z}_i^{\mathsf{variance}}}_{\check{z}^{\mathsf{variance}}_{k+1:t}},
\end{aligned} \tag{80}
$$

where (a) follows from (67).

4 (b) Bounding the bias term

First term, $\check{z}^{\mathsf{bias}}_{k+1:t}$in (80) is bounded as follows:

$$
\begin{aligned}
\check{z}^{\mathsf{bias}}_{k+1:t} &\leq \frac{2}{N^2}\left(1 + \frac{4}{\check{\beta}(2\mu+\zeta)}\right) \sum_{i=k+1}^{\infty} \check{z}_i^{\mathsf{bias}} \\
&\overset{(a)}{\leq} \frac{2}{N^2}\left(1 + \frac{4}{\check{\beta}(2\mu+\zeta)}\right) \sum_{i=k+1}^{\infty} (1 - \check{\beta}(2\mu+\zeta))^i \mathbb{E}\left[\|\check{z}_0\|_2^2\right] \\
&\overset{(b)}{=} \frac{2\mathbb{E}\left[\|\check{z}_0\|_2^2\right]}{\check{\beta}(2\mu+\zeta)N^2} \left(1 - \check{\beta}(2\mu+\zeta)\right)^{k+1} \left(1 + \frac{4}{\check{\beta}(2\mu+\zeta)}\right),
\end{aligned}
$$

where (a) follows from (73), which provides a bound on $\check{z}_i^{\mathsf{bias}}$ and (b) follows from the bound on the summation of a geometric series.

4 (c) Bounding the variance term

Next, the second term $z^{\text{variance}}_{k+1:t}$ in (80) is bounded as follows:

$$\begin{aligned}
\check{z}^{\text{variance}}_{k+1:t} &\overset{(a)}{\leq} \frac{2\check{\beta}^2}{N^2}\left(1+\frac{4}{\check{\beta}(2\mu+\zeta)}\right)\sum_{i=k+1}^{k+N}\frac{\check{\sigma}^2}{\check{\beta}(2\mu+\zeta)} \\
&\leq \frac{2\check{\beta}^2}{N^2}\left(1+\frac{4}{\check{\beta}(2\mu+\zeta)}\right)\sum_{i=0}^{N}\frac{\check{\sigma}^2}{\check{\beta}(2\mu+\zeta)} \\
&= \left(1+\frac{4}{\check{\beta}(2\mu+\zeta)}\right)\frac{2\check{\beta}\check{\sigma}^2}{(2\mu+\zeta)N},
\end{aligned}$$

where (a) follows from (77), which provides a bound on $\check{z}^{\text{variance}}_i$.

Step 5: Clinching argument

Finally substituting the bounds on $\check{z}^{\text{bias}}_{k+1:t}$ and $\check{z}^{\text{variance}}_{k+1:t}$ in (80), we get

$$\begin{aligned}
&\mathbb{E}[\|\check{z}_{k+1:t}\|_2^2] \\
&\leq \left(1+\frac{4}{\check{\beta}(2\mu+\zeta)}\right)\left(\frac{2}{\check{\beta}(2\mu+\zeta)N^2}(1-\check{\beta}(2\mu+\zeta))^{k+1}\mathbb{E}[\|\check{z}_0\|_2^2]+\frac{2\check{\beta}\check{\sigma}^2}{(2\mu+\zeta)N}\right), \\
&\overset{(a)}{\leq} \left(1+\frac{4}{\check{\beta}(2\mu+\zeta)}\right)\left(\frac{2\exp(-k\check{\beta}(2\mu+\zeta))}{\check{\beta}(2\mu+\zeta)N^2}\mathbb{E}[\|z_0\|_2^2]+\frac{2\check{\beta}\check{\sigma}^2}{(2\mu+\zeta)N}\right) \\
&\overset{(b)}{\leq} \frac{10\exp(-k\check{\beta}(2\mu+\zeta))}{\check{\beta}^2(2\mu+\zeta)^2N^2}\mathbb{E}\left[\|\check{z}_0\|_2^2\right]+\frac{10\check{\sigma}^2}{\check{\beta}(2\mu+\zeta)^2N},
\end{aligned} \tag{81}$$

where (a) follows from $(1+x)^y = \exp(y\log(1+x)) \leq \exp(xy)$, and (b) uses $\check{\beta}(2\mu+\zeta) < 1$ as $\check{\beta} \leq \check{\beta}_{\max}$ defined in Theorem 9.1, which implies that

$$1+\frac{4}{\check{\beta}(2\mu+\zeta)} \leq \frac{5}{\check{\beta}(2\mu+\zeta)}.$$

□

Proof of Theorem 3.3

The proof of Theorem 3.3 builds on Theorem 9.1 and a bound on $\|\check{w}_{k+1:t}-\bar{w}_{\text{reg}}\|_2^2$, incorporating techniques from (Patil et al., 2024, Corollary 1,2).

Proof. Notice that

$$\mathbb{E}\left[\|\check{w}_{k+1:t}-\bar{w}\|_2^2\right] \overset{(i)}{\leq} \underbrace{2\|\bar{w}_{\text{reg}}-\bar{w}\|_2^2}_{\text{Term 1}} + \underbrace{2\mathbb{E}\left[\|\check{w}_{k+1:t}-\bar{w}_{\text{reg}}\|_2^2\right]}_{\text{Term 2}}, \tag{82}$$

where (i) follows by using $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

We bound Term 1 below.

$$\begin{aligned}
\|\bar{w}-\bar{w}_{\text{reg}}\|_2^2 &= \left\|\mathbf{M}^{-1}\xi-(\mathbf{M}+\zeta\mathbf{I})^{-1}\xi\right\|_2^2 \\
&\overset{(a)}{\leq} \left\|\mathbf{M}^{-1}-(\mathbf{M}+\zeta\mathbf{I})^{-1}\right\|_2^2\|\xi\|_2^2 \\
&= \left\|\mathbf{M}^{-1}(\mathbf{M}+\zeta\mathbf{I}-\mathbf{M})(\mathbf{M}+\zeta\mathbf{I})^{-1}\right\|_2^2\|\xi\|_2^2 \\
&\leq \left\|\mathbf{M}^{-1}\right\|_2^2\zeta^2\left\|(\mathbf{M}+\zeta\mathbf{I})^{-1}\right\|_2^2\|\xi\|_2^2
\end{aligned}$$

$$\overset{(b)}{\leq} \frac{\zeta^2 (R_{\mathsf{max}}^2 \left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2 (\phi_{\mathsf{max}}^u)^2\right))}{\iota^2 (\zeta + \iota)^2}, \tag{83}$$

where (a) follows from $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$, and (b) follows from the fact that $\left\|\mathbf{M}^{-1}\right\| = 1/\iota_{\min}(\mathbf{M})$, where $\iota = \iota_{\min}(\mathbf{M})$ is the minimum singular value of $\mathbf{M}$.

We observe that (81) bounds Term 2. Using this bound and (83) in (82), we obtain

$$\begin{aligned}\mathbb{E}\left[\|\check{w}_{k+1:t} - \bar{w}\|_2^2\right] \leq & \frac{20 \exp(-k\check{\beta}(2\mu + \zeta))}{\check{\beta}^2 (2\mu + \zeta)^2 N^2} \mathbb{E}\left[\|\check{z}_0\|_2^2\right] + \frac{20 \check{\sigma}^2}{\check{\beta}(2\mu + \zeta)^2 N} \\ & + \frac{\zeta^2 (R_{\mathsf{max}}^2 \left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2 (\phi_{\mathsf{max}}^u)^2\right))}{\iota^2 (\zeta + \iota)^2}.\end{aligned} \tag{84}$$

For $\zeta = \frac{1}{\sqrt{N}}$, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\check{w}_{k+1:t} - \bar{w}\|_2^2\right] \leq & \frac{20 \exp\left(-k\check{\beta}(2\mu + (N)^{-1/2})\right)}{\check{\beta}^2 (2\mu + \zeta)^2 N^2} \mathbb{E}\left[\|\check{w}_0 - \bar{w}_{\mathsf{reg}}\|_2^2\right] + \frac{20 \check{\sigma}^2}{\mu^2 N} \\ & + \frac{2 (R_{\mathsf{max}}^2 \left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2 (\phi_{\mathsf{max}}^u)^2\right))}{\iota^2 N}.\end{aligned} \tag{85}$$

□

10 High Probability Bounds for Mean-Variance TD

For the high probability bound, we consider the following update rule and assumption:

$$w_{t+1} = \Gamma(w_t + \beta h_t(w_t)), \tag{86}$$

where Γ projects on to the set $\mathcal{C} \triangleq \{w \in \mathbb{R}^{2q} \mid \|w\|_2 \leq H\}$.

Assumption 9. *The projection radius H of the set $\mathcal{C}$ satisfies $H > \frac{\|\xi\|_2}{\mu}$, where $\mu = \lambda_{\mathsf{min}}(\frac{\mathbf{M}^\top + \mathbf{M}}{2})$ and ξ is as defined in (6).*

Under the additional projection-related assumption above, we state and prove a high probability bound for the tail-averaged variant of Algorithm 1 in the next section. Subsequently, we analyze the regularized mean-variance TD variant to derive high-probability bounds.

10.1 Bounds for vanilla (un-regularized) mean-variance TD

Theorem 10.1. *Suppose Assumptions 1 to 6 hold. Run Algorithm 1 for t iterations with step size β as defined in Theorem 3.2. Then, for any $\delta \in (0, 1]$, we have the following bound for the projected tail-averaged iterate $w_{k+1:t}$ with $N = t - k$:*

$$\mathbb{P}\left(\|w_{k+1:t} - \bar{w}\|_2 \leq \frac{2\tau}{\mu\sqrt{N}} \sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{4\exp\left(-k\beta\mu\right)}{\beta\mu N} \mathbb{E}\left[\|w_0 - \bar{w}\|_2\right] + \frac{4\tau}{\mu\sqrt{N}} \right) \geq 1 - \delta,$$

where $w_0, \bar{w}, \beta$ are defined as in Theorem 3.1, and

$$\begin{aligned}\tau = & \Big(2R_{\mathsf{max}}^2 \left((\phi_{\mathsf{max}}^v)^2 + R_{\mathrm{max}}^2 (\phi_{\mathsf{max}}^u)^2\right) + 2\Big((\phi_{\mathsf{max}}^v)^4 (1+\gamma)^2 + (\phi_{\mathsf{max}}^u)^4 \left(1+\gamma^2\right)^2 \\ & + 4\gamma^2 R_{\mathrm{max}}^2 (\phi_{\mathsf{max}}^v)^2 (\phi_{\mathsf{max}}^u)^2\Big) H^2\Big)^{\frac{1}{2}}.\end{aligned}$$

The proof follows a similar structure to Patil et al. (2024, Theorem 2) and Prashanth et al. (2021, Proposition 8.3), with necessary adaptations to account for our setting.

Proof. A martingale difference decomposition of $\left\|z_{k+1,N}\right\|_2 - \mathbb{E}[\left\|z_{k+1:t}\right\|_2]$ is as follows:

$$\left\|z_{k+1,N}\right\|_2 - \mathbb{E}\left[\left\|z_{k+1:t}\right\|_2\right] = \sum_{i=k+1}^{k+N}(g_i - g_{i-1}) = \sum_{i=k+1}^{k+N} D_i, \tag{87}$$

where $z_{k+1:t}$ denotes tail-averaged iterate error,

$$D_i \triangleq g_i - \mathbb{E}\left[g_i \mid \mathcal{G}_{i-1}\right], \; g_i \triangleq \mathbb{E}[\left\|z_{k+1:t}\right\|_2 \mid \mathcal{G}_i], \text{ and}$$

$\mathcal{G}_i$ denotes the sigma-field generated by random variables $\{w_t, \; t \leq i\}$ for $t, \; i \; \in \mathbb{Z}^+$.

Let $h_i(w) \triangleq r_i\phi_i - \mathbf{M}_i w$ denote random innovation at time i for $w_i = w$. If we show that functions g_i are L_i Lipschitz continuous in the random innovation h_i at time i, then we can see that the martingale difference D_i is a L_i Lipschitz function of the ith random innovation.

Let $\Omega_j^i(w)$ represent the iterate value at time j, evolving according to (86), starting from the value of w at time i. Let w and w' be two different iterate values at time i, dependent on h and h', respectively, as $w = w_{i-1} + \beta h$ and $w' = w_{i-1} + \beta h'$. We compute the difference between the iterate values at time j when the initial values at time i are w and w' as follows:

$$\begin{aligned}\Omega_j^i(w) - \Omega_j^i(w') &= \Omega_{j-1}^i(w) - \Omega_{j-1}^i(w') - \beta[h_j(\Omega_{j-1}^i(w)) - h_j(\Omega_{j-1}^i(w'))] \\ &= \Omega_{j-1}^i(w) - \Omega_{j-1}^i(w') - \beta\mathbf{M}_j(\Omega_{j-1}^i(w) - \Omega_{j-1}^i(w')) \\ &= (\mathbf{I} - \beta\mathbf{M}_j)(\Omega_{j-1}^i(w) - \Omega_{j-1}^i(w')).\end{aligned} \tag{88}$$

Taking expectation and since the projection Γ is non-expansive, we have the following

$$\begin{aligned}\mathbb{E}\left[\left\|\Omega_j^i(w) - \Omega_j^i(w')\right\|_2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\Omega_j^i(w) - \Omega_j^i(w')\right\|_2 \,\middle|\, \mathcal{G}_{j-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\|(\mathbf{I} - \beta M_j)(\Omega_{j-1}^i(w) - \Omega_{j-1}^i(w'))\right\|_2 \,\middle|\, \mathcal{G}_{j-1}\right]\right] \\ &\overset{(i)}{\leq} \left(1 - \frac{\beta\mu}{2}\right)\mathbb{E}[\left\|\Omega_{j-1}^i(w) - \Omega_{j-1}^i(w')\right\|_2] \\ &\overset{(ii)}{=} \left(1 - \frac{\beta\mu}{2}\right)^{j-i+1}\left\|w - w'\right\|_2, \\ &\overset{(iii)}{\leq} \beta\left(1 - \frac{\beta\mu}{2}\right)^{j-i+1}\left\|h - h'\right\|_2.\end{aligned} \tag{89}$$

where (i) follows by Lemma 7.1; (ii) follows by repeated application of (i); and (iii) follows by substituting w and w'.

Let $\Omega_t^i(w)$ to be the value of the iterate at time t, where t ranges from the tail index $k+1$ to $k+N$. The iterate evolves according to (8) beginning from w at time $i = k+1$. Next, we define

$$\tilde{\Omega}_{k+1:t}^i(\tilde{w}, w) \triangleq \frac{(i-k)\tilde{w}}{N} + \frac{1}{N}\sum_{j=i+1}^{i+N}\Omega_j^i(w), \tag{90}$$

where $\tilde{w}$ is the value of the tail averaged iterate at time i. In the above, $\tilde{\Omega}_{k+1:t}^i(\tilde{w}, w)$ denotes the value of tail-averaged iterate at time t.

From (90) and using the triangle inequality, we have

$$\mathbb{E}\left[\left\|\tilde{\Omega}_{i+1,N}^i(\tilde{w}, w) - \tilde{\Omega}_{k+1:t}^i(\tilde{w}, w')\right\|_2\right] \leq \mathbb{E}\left[\frac{1}{N}\sum_{j=i+1}^{i+N}\left\|(\Omega_j^i(w) - \Omega_j^i(w'))\right\|_2\right]. \tag{91}$$

Using (89), we bound the term $\Omega_j^i(w) - \Omega_j^i(w')$ inside the summation of (91).

$$\mathbb{E}\left[\left\|\tilde{\Omega}_{k+1}^i(\tilde{w}, w) - \tilde{\Omega}_{k+1}^i(\tilde{w}, w')\right\|_2\right] \leq \frac{\beta}{N} \sum_{j=i+1}^{i+N} \left(1 - \frac{\beta\mu}{2}\right)^{j-i+1} \|h - h'\|_2 . \tag{92}$$

Considering the bounds on features, rewards, and the projection assumption (Assumptions 3 to 6), along with a bound on σ in (53), we obtain a uniform upper bound τ on $\|h_i(w)\|$ for all i as:

$$\tau = \Bigg(2R_{\max}^2\left((\phi_{\max}^v)^2 + R_{\max}^2(\phi_{\max}^u)^2\right) + 2\left((\phi_{\max}^v)^4(1+\gamma)^2 + (\phi_{\max}^u)^4\left(1+\gamma^2\right)^2 + 4\gamma^2 R_{\max}^2(\phi_{\max}^v)^2(\phi_{\max}^u)^2\right)H^2\Bigg)^{\frac{1}{2}}$$

Now, we use a martingale difference concentration, following Patil et al. (2024, Step 3, Theorem 2) to obtain

$$\mathbb{P}\left(\|z_{k+1,N}\|_2 - \mathbb{E}\left[\|z_{k+1,N}\|_2\right] > \epsilon\right) \leq \exp(-\eta\epsilon)\exp\left(\frac{\eta^2\tau^2\sum_{i=k+1}^{k+N} L_i^2}{2}\right).$$

Optimising over η in the above inequality leads to

$$\mathbb{P}\left(\|z_{k+1:t}\|_2 - \mathbb{E}\left[\|z_{k+1:t}\|_2\right] > \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{\tau^2\sum_{i=k+1}^{k+N} L_i^2}\right). \tag{93}$$

Using Patil et al. (2024, Lemma 13), we obtain the following bound on the Lipschitz constant,

$$\sum_{i=k+1}^{k+N} L_i^2 \leq \frac{4}{N\mu^2}. \tag{94}$$

Now, with (94) in (93), we have

$$\mathbb{P}\left(\|z_{k+1:t}\|_2 - \mathbb{E}[\|z_{k+1:t}\|_2] > \epsilon\right) \leq \exp\left(-\frac{N\mu^2\epsilon^2}{4\tau^2}\right), \tag{95}$$

For any $\delta \in (0, 1]$ the inequality (95) can be expressed in high-confidence form as:

$$\mathbb{P}\left(\|z_{k+1:t}\|_2 - \mathbb{E}[\|z_{k+1:t}\|_2] \leq \frac{2\tau}{\mu\sqrt{N}}\sqrt{\log\left(\frac{1}{\delta}\right)}\right) \geq 1 - \delta. \tag{96}$$

The final bound follows by substituting the bound on $\mathbb{E}\left[\|z_{k+1:t}\|_2\right]$ obtained by applying Jensen's inequality to Theorem 3.2 in (96). □

10.2 Bounds for mean-variance TD with regularization

Theorem 10.2. *Suppose Assumptions 1 to 4, and 6 hold. Run Algorithm 1 for t iterations with a step size $\breve{\beta}$ as specified in Theorem 9.1. Then, for any $\delta \in (0, 1]$, we have the following bound for the projected tail-averaged regularized TD iterate:*

$$\mathbb{P}\Bigg(\|\breve{w}_{k+1:t} - \bar{w}_{\text{reg}}\|_2 \leq \frac{2\breve{\tau}}{(2\mu+\zeta)\sqrt{N}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{4\exp\left(-k\breve{\beta}\left(2\mu+\zeta\right)\right)}{\breve{\beta}\left(2\mu+\zeta\right)N}\mathbb{E}\left\|w_0 - \bar{w}_{\text{reg}}\right\|_2 + \frac{4\breve{\tau}}{(2\mu+\zeta)\sqrt{N}}\Bigg) \geq 1 - \delta,$$

where $N, \breve{w}_0, \bar{w}_{\mathsf{reg}}, \mu$. are as specified in Theorem 9.1 and

$$\check{\tau} = \Big(2R_{\mathsf{max}}^2\left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^u)^2\right) + 4\big(\zeta^2 + (\phi_{\mathsf{max}}^v)^4\left(1+\gamma\right)^2 + (\phi_{\mathsf{max}}^u)^4\left(1+\gamma^2\right)^2 + 4\beta^2 R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^v)^2(\phi_{\mathsf{max}}^u)^2\big)H^2\Big)^{\frac{1}{2}}.$$

The proof for the regularized case follows using arguments similar to those in the proof of Theorem 3.4 with changes indicated below.

Proof. Let $\check{\Omega}_j^i(\breve{w})$ represent the iterate value at time j, evolving following (86), starting from the value of $\breve{w}$ at time i. We compute the difference between the iterate values at time j when the initial values at time i are $\breve{w}$ and $\breve{w}'$, respectively. Let $\breve{w}$ and $\breve{w}'$ be two different parameter values at time i which depend on $\check{h}$ and $\check{h}'$ as $\breve{w} = \breve{w}_{i-1} + \breve{\beta}\check{h}$, and $\breve{w}' = \breve{w}_{i-1} + \breve{\beta}h'$. We obtain the difference as:

$$\begin{aligned}\check{\Omega}_j^i(\breve{w}) - \check{\Omega}_j^i(\breve{w}') &= \check{\Omega}_{j-1}^i(\breve{w}) - \check{\Omega}_{j-1}^i(\breve{w}') - \breve{\beta}[\check{h}_j(\check{\Omega}_{j-1}^i(\breve{w})) - \check{h}_j(\check{\Omega}_{j-1}^i(\breve{w}'))] \\ &= (\mathbf{I} - \breve{\beta}(\zeta\mathbf{I} + \mathbf{M}_j))(\check{\Omega}_{j-1}^i(\breve{w}) - \check{\Omega}_{j-1}^i(\breve{w}')). \end{aligned} \tag{97}$$

Taking expectation and since the projection Γ is non-expansive, we have the following

$$\begin{aligned}\mathbb{E}\left[\left\|\check{\Omega}_j^i(\breve{w}) - \check{\Omega}_j^i(\breve{w}')\right\|_2\right] &= \mathbb{E}\left[\mathbb{E}\left[\left\|\check{\Omega}_j^i(\breve{w}) - \check{\Omega}_j^i(\breve{w}')\right\|_2 \,\middle|\, \check{\mathcal{G}}_{j-1}\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\left\|(\mathbf{I} - \breve{\beta}\mathbf{M}_j)(\check{\Omega}_{j-1}^i(\breve{w}) - \check{\Omega}_{j-1}^i(\breve{w}'))\right\|_2 \,\middle|\, \check{\mathcal{G}}_{j-1}\right]\right] \\ &\overset{(i)}{\leq} \left(1 - \frac{\breve{\beta}(2\mu+\zeta)}{2}\right)\mathbb{E}\left[\left\|\check{\Omega}_{j-1}^i(w) - \check{\Omega}_{j-1}^i(\breve{w}')\right\|_2\right] \\ &\overset{(ii)}{=} \left(1 - \frac{\breve{\beta}(2\mu+\zeta)}{2}\right)^{j-i+1} \left\|\breve{w} - \breve{w}'\right\|_2, \\ &\leq \breve{\beta}\left(1 - \frac{\breve{\beta}(2\mu+\zeta)}{2}\right)^{j-i+1} \left\|\check{h} - \check{h}'\right\|_2. \end{aligned} \tag{98}$$

where (i) follows by Lemma 9.3; (ii) follows by repeated application of (i); and (98) follows by substituting the values of w and w'.

Let $\check{\Omega}_t^i(\breve{w})$ be the value of the iterate at time t where t ranges from the tail index $k+1$ to $k+N$. The iterate evolves according to (14) starting at the value $\breve{w}$ at time $i = k+1$. Next, we define

$$\bar{\Omega}_{k+1:t}^i(\hat{w}, \breve{w}) \triangleq \frac{(i-k)\tilde{\breve{w}}}{N} + \frac{1}{N}\sum_{j=i+1}^{i+N} \check{\Omega}_j^i(\breve{w}), \tag{99}$$

where $\hat{w}$ is the value of the tail-averaged iterate at time i.

Now, we prove that Lipschitz continuity in the random innovation $\check{h}_i$ at time i with constant $\check{L}_i$.

$$\mathbb{E}\left[\left\|\tilde{\check{\Omega}}_{i+1,N}^i(\tilde{\breve{w}}, \breve{w}) - \tilde{\check{\Omega}}_{k+1:t}^i(\tilde{\breve{w}}, \breve{w}')\right\|_2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{j=i+1}^{i+N}\left\|(\check{\Omega}_j^i(\breve{w}) - \check{\Omega}_j^i(\breve{w}'))\right\|_2\right]. \tag{100}$$

Using (98), we bound the term $\check{\Omega}_j^i(\breve{w}) - \check{\Omega}_j^i(\breve{w}')$ in (100).

$$\mathbb{E}\left[\left\|\tilde{\Omega}_{k+1}^i(\tilde{w}, w) - \tilde{\Omega}_{k+1}^i(\tilde{w}, w')\right\|_2\right] \leq \frac{\beta}{N}\sum_{j=i+1}^{i+N}\left(1 - \frac{\breve{\beta}(2\mu+\zeta)}{2}\right)^{j-i+1}\left\|\check{h} - \check{h}'\right\|_2. \tag{101}$$

Considering the bounds on features, rewards, and the projection assumption (Assumptions 3 to 6), along with a bound on $\breve{\sigma}$ in (59), we find an upper bound $\check{\tau}$ on $\left\|\check{h}_i(\breve{w}_i)\right\|$ as follows:

$$\check{\tau} = \Big(2R_{\mathsf{max}}^2\left((\phi_{\mathsf{max}}^v)^2 + R_{\mathsf{max}}^2(\phi_{\mathsf{max}}^u)^2\right)$$

$$+ 4\left(\zeta^2 + (\phi^v_{\max})^4 (1+\gamma)^2 + (\phi^u_{\max})^4 \left(1+\gamma^2\right)^2 + 4\beta^2 R^2_{\max} (\phi^v_{\max})^2 (\phi^u_{\max})^2\right) H^2\Big)^{\frac{1}{2}}.$$

Using Patil et al. (2024, Lemma 20), we obtain the following bound on the Lipschitz constant,

$$\sum_{i=k+1}^{k+N} \check{L}_i^2 \leq \frac{4}{N(2\mu+\zeta)^2}. \tag{102}$$

The rest of the proof follows by making parallel arguments to those in Subsection 10.1. □

11 Outline of Actor Analysis

Proof. (Sketch) As visualized in Figure 1, the proof begins by establishing the smoothness of the

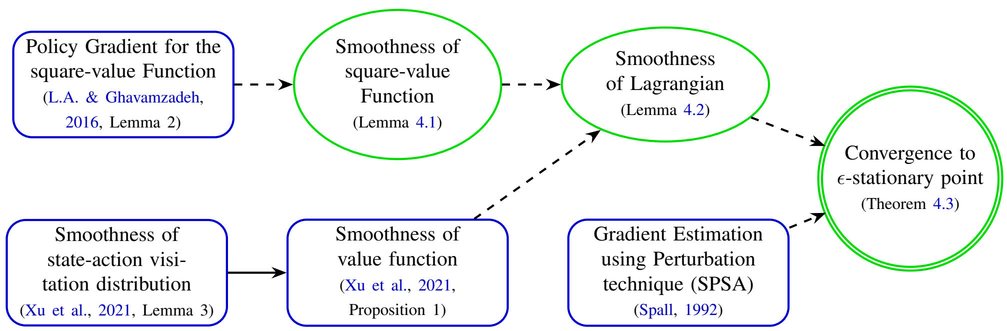

Figure 1: Logical dependency graph for proving Theorem 4.3. Rectangular nodes (blue) represent established results from prior work, elliptical nodes (green) denote our novel contributions, and dashed lines illustrate the logical dependencies we establish to derive the final result (green circle).

policy gradient for the square-value function:

$$\nabla U(\theta) = \frac{1}{1-\gamma^2}\Big(\underbrace{\textstyle\sum_{s,a} \tilde{\nu}_\theta(s,a) \nabla \log \pi_\theta(a|s) W_\theta(s,a)}_{T_1(\theta)} + 2\gamma \underbrace{\textstyle\sum_{s,a,s'} \tilde{\nu}_\theta(s,a) P(s'|s,a) \nabla V_\theta(s')}_{T_2(\theta)}\Big). \tag{103}$$

We decompose the expression in (103) into $T_1(\theta)$ and $T_2(\theta)$. $T_1(\theta)$ consists of three terms: the state-action visitation distribution, the score function, and the square-value function. To obtain a smoothness constant for $T_1(\theta)$ (36), we use the following: (i) the smoothness result for the state-action visitation distribution (Lemma 12.1), as stated in (Xu et al., 2021, Lemma 3); (ii) the boundedness and smoothness of the policy (Assumption 7).

$T_2(\theta)$ is the product of the state-action visitation distribution and the policy gradient of the value function. To establish the smoothness constant for $T_2(\theta)$, we apply the smoothness result for the value function from (Xu et al., 2021, Proposition 1).

Combining the results for $T_1(\theta)$ and $T_2(\theta)$ gives the smoothness constants for the square-value function. By splitting the terms in the Lagrangian into the gradients of the value function and the square-value function and appropriately bounding the gradient norms, we obtain the smoothness constant L in (21) for the Lagrangian.

The proof broadly follows a standard SGD analysis framework (Ghadimi & Lan, 2013; Kumar et al., 2023). However, key modifications are required to account for the use of SPSA-based gradient estimates, particularly in handling the perturbation parameter p_t and critic batch size m.

As $\nabla L(\theta_t)$ is L-Lipschitz (Lemma 4.2), we have

$$L(\theta_{t+1}) \geq L(\theta_t) + \langle \nabla L(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L\alpha_t^2}{2} \|\nabla \hat{L}(\theta_t)\|^2$$

In the above, $\nabla \hat{L}(\theta_t)$. is an SPSA gradient estimate.

Taking the expectation with respect to the sigma field $\mathcal{F}_t = \sigma(\theta_k, k \leq t)$, denoted by $\mathbb{E}_t$, we have

$$\begin{aligned}\mathbb{E}_t[L(\theta_{t+1})] \geq\ & \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] \\ & - \alpha_t K_1 \left(1 + \frac{2\lambda R_{\text{max}}}{1-\gamma}\right) \underbrace{\left\| \mathbb{E}_t\left[\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\|}_{\text{(A)}} \\ & - \lambda \alpha_t K_1 \underbrace{\left\| \mathbb{E}_t\left[\nabla \hat{U}(\theta_t) - \nabla U(\theta_t)\right]\right\|}_{\text{(B)}} - \frac{L}{2}\alpha_t^2 \underbrace{\mathbb{E}_t\left[\|\nabla \hat{L}(\theta_t)\|^2\right]}_{\text{(C)}}.\end{aligned}$$

Now, substituting the bounds obtained for biased SPSA gradient estimates namely: (A) in (116), (B) in (117), and (C) in (118) into the above equation, we get

$$\begin{aligned}\mathbb{E}_t[L(\theta_{t+1})] \geq\ & \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|\right] \\ & - \alpha_t K_1 \left(1 + \frac{2\lambda R_{\text{max}}}{1-\gamma}\right)\left(\frac{d^{\frac{3}{2}} L_J p_t}{2} + \frac{d^{\frac{1}{2}} \phi^v_{\text{max}} K_2}{p_t\sqrt{m}}\right) \\ & - \lambda \alpha_t K_1 \left(\frac{d^{\frac{3}{2}} L_U p_t}{2} + \frac{d^{\frac{1}{2}} \phi^u_{\text{max}} K_2}{p_t\sqrt{m}}\right) - \frac{L\alpha_t^2}{2}\left(\frac{K_3}{p_t^2}\right).\end{aligned}$$

Summing from $t = 1$ to n and dividing both sides by n, and setting $\alpha_t = \alpha$ and $p_t = p$, we get

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[\|\nabla L(\theta_t)\|^2\right] \leq \frac{C_1}{n\alpha} + C_2 p + \frac{C_3}{\sqrt{m}p} + \frac{C_4\alpha}{p^2}.$$

Setting $\alpha = n^a$, $p = n^b$, $m = n^c$, we have

$$\mathbb{E}\left[\|\nabla L(\theta_R)\|^2\right] \leq C_1 n^{-1-a} + C_2 n^b + C_3 n^{-b-c/2} + C_4 n^{a-2b}.$$

Optimizing for a, b, c, we find their values to be $a = -\frac{3}{4}$, $b = -\frac{1}{4}$, $c = 1$. Substituting these values, we get

$$\begin{aligned}\mathbb{E}\left[\|\nabla L(\theta_R)\|^2\right] &\leq C_1 n^{-1/4} + C_2 n^{-1/4} + C_3 n^{-1/4} + C_4 n^{-1/4} \\ &= O(n^{-1/4}).\end{aligned}$$

□

12 Proofs for the claims in Section 4

Before we prove the claims, we state a few useful supporting lemmas in the analysis.

Lemma 12.1 (Restatement of Lemma 3 (Xu et al., 2021))**.** *Consider the initialization distribution $\eta(\cdot)$ and the transition kernel $\mathsf{P}(\cdot|s,a)$. Let $\eta(\cdot) = \zeta(\cdot)$ or $\eta(\cdot) = \mathsf{P}(\cdot|\hat{s},\hat{a})$ for any given $(\hat{s},\hat{a}) \in \mathcal{S} \times \mathcal{A}$. Denote $\nu_{\pi_\theta,\eta}(\cdot,\cdot)$ as the state-action visitation distribution of the MDP with policy π_θ and initialization distribution $\eta(\cdot)$. Suppose the Assumption holds. Then, we have*

$$\left\|\nu_{\pi_{\theta_1},\eta}(\cdot,\cdot) - \nu_{\pi_{\theta_2},\eta}(\cdot,\cdot)\right\|_{TV} \leq C_\nu \left\|\theta_1 - \theta_2\right\|_2,$$

for all $\theta_1, \theta_2 \in \mathbb{R}^d$, where $C_\nu = C_\pi\left(1 + \lceil \log_\rho \kappa^{-1} \rceil + \frac{1}{1-\rho}\right)$.

12.1 Proof of Lemma 4.1

Proof. The first claim concerning the smoothness of $J(\cdot)$ can be inferred from Xu et al. (2021, Proposition 1).

We prove the smoothness of the square-value function below.

From (L.A. & Ghavamzadeh, 2016, Lemma 1), we have

$$\nabla U(\theta) = \frac{1}{1-\gamma^2}\left(\underbrace{\sum_{s,a}\tilde{\nu}_\theta(s,a)\nabla\log\pi_\theta(a|s)W_\theta(s,a)}_{T_1(\theta)} + 2\gamma\underbrace{\sum_{s,a,s'}\tilde{\nu}_\theta(s,a)P(s'|s,a)\nabla V_\theta(s')}_{T_2(\theta)}\right), \tag{104}$$

where

$$W_\theta(s,a) = \mathbb{E}\left[\left(\sum_{k=0}^{\infty}\gamma^k r_{t+k}\right)^2 \middle| s_t = s, a_t = a\right]$$

and $\tilde{\nu}_\theta(s,a) = (1-\gamma^2)\sum_{t=0}^{\infty}\gamma^{2t}\mathbb{P}(s_t = s, a_t = a)$ is the γ^2-discounted state-action visitation distribution, with $\mathbb{P}(s_t = s, a_t = a) = \mathbb{P}(s_t = s|s_0 = s)\pi_\theta(a|s)$.

$$\|\nabla U(\theta_1) - \nabla U(\theta_2)\|_2 \le \frac{1}{1-\gamma^2}\left(\|T_1(\theta_1) - T_1(\theta_2)\|_2 + 2\gamma\|T_2(\theta_1) - T_2(\theta_2)\|_2\right) \tag{105}$$

We now show that $T_1(\theta)$, defined in (104) is Lipschitz in θ.

$$\begin{aligned}
&\|T_1(\theta_1) - T_1(\theta_2)\|_2 \\
&= \left\|\sum_{s,a}\underbrace{\tilde{\nu}_{\theta_1}(s,a)}_{a_1}\underbrace{\nabla\log\pi_{\theta_1}(a|s)}_{b_1}\underbrace{W_{\pi_{\theta_1}}(s,a)}_{c_1} - \sum_{s,a}\underbrace{\tilde{\nu}_{\theta_2}(s,a)}_{a_2}\underbrace{\nabla\log\pi_{\theta_2}(a|s)}_{b_2}\underbrace{W_{\pi_{\theta_2}}(s,a)}_{c_2}\right\|_2 \\
&= \left\|\sum_{s,a}(a_1b_1c_1 - a_2b_2c_2)\right\| \\
&= \left\|\sum_{s,a}a_1b_1c_1 - a_2b_2c_2 + a_2b_2c_1 - a_2b_2c_1\right\| \\
&= \left\|\sum_{s,a}c_1(a_1b_1 - a_2b_2) + a_2b_2(c_1 - c_2)\right\| \\
&= \left\|\sum_{s,a}c_1(a_1b_1 - a_2b_2 + a_1b_2 - a_1b_2) + a_2b_2(c_1 - c_2)\right\| \\
&= \left\|\sum_{s,a}c_1(a_1(b_1 - b_2) + b_2(a_1 - a_2)) + a_2b_2(c_1 - c_2)\right\| \\
&\le \sum_{s,a}\left|W_{\theta_1}(s,a)\right|\left|\tilde{\nu}_{\theta_1}(s,a)\right|\|\nabla\log\pi_{\theta_1}(a|s) - \nabla\log\pi_{\theta_2}(a|s)\|_2 \\
&\quad + \sum_{s,a}\left|W_{\theta_1}(s,a)\right|\|\nabla\log\pi_{\theta_1}(a|s)\|_2\left|\tilde{\nu}_{\theta_1}(s,a) - \tilde{\nu}_{\theta_2}(s,a)\right| \\
&\quad + \sum_{s,a}\tilde{\nu}_{\theta_2}(s,a)\|\nabla\log\pi_{\theta_2}(a|s)\|_2\left|W_{\theta_1}(s,a) - W_{\theta_2}(s,a)\right| \\
&\overset{(a)}{\le} \frac{R_{\max}}{(1-\gamma)^2}\sum_{s,a}\|\nabla\log\pi_{\theta_1}(a|s) - \nabla\log\pi_{\theta_2}(a|s)\|_2
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{C_\psi R_{\max}}{(1-\gamma)^2} \sum_{s,a} |\tilde{\nu}_{\theta_1}(s,a) - \tilde{\nu}_{\theta_2}(s,a)| \\
&+ C_\psi \sum_{s,a} \Big| W_{\theta_1}(s,a) - W_{\theta_2}(s,a) \Big| \tilde{\nu}_{\theta_2}(s,a) \\
&\overset{(b)}{\leq} \frac{R_{\max} L_\psi}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 + \frac{2R_{\max} C_\psi C_\nu}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 \\
&\quad + C_\psi \sum_{s,a} |W_{\theta_1}(s,a) - W_{\theta_2}(s,a)| \tilde{\nu}_{\theta_2}(s,a) \\
&\overset{(c)}{\leq} \frac{R_{\max} L_\psi}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 + \frac{2R_{\max} C_\psi C_v}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 \\
&\quad + \frac{2R_{\max} C_\psi C_v}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 \\
&\leq \frac{R_{\max} L_\psi}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 + \frac{4R_{\max} C_\psi C_v}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2, \qquad (106)
\end{aligned}
$$

where (a) follows by $|W_\theta(s,a)||\tilde{\nu}_{\theta_1}(s,a)| \leq \frac{R_{\max}}{(1-\gamma)^2}$ for any $\theta \in \mathbb{R}^d$ and by the upper bound C_ψ on the score function, see Assumption 7; (b) follows by smoothness of the policy (Assumption 7) and C_ν- Lipschitzness of $\tilde{\nu}(s,a)$ (see (Xu et al., 2021, Lemma 3)); (c) follows by employing similar arguments for the square-value function, in place of the value function in (Xu et al., 2021, Lemma 4), as outlined below:

$$
\begin{aligned}
C_\psi \sum_{s,a} |W^\pi_{\theta_1}(s,a) - W^\pi_{\theta_2}(s,a)| \tilde{\nu}_{\theta_2}(s,a) &\leq C_\psi \frac{R_{\max}}{(1-\gamma)^2} \|P^\pi_{\theta_1}(\cdot,\cdot) - P^\pi_{\theta_2}(\cdot,\cdot)\|_{TV} \\
&\leq \frac{2R_{\max} C_\psi C_v}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2.
\end{aligned}
$$

Next, we obtain the Lipschitz constant for $T_2(\theta) = \sum_{s,a,s'} \tilde{\nu}_\theta(s,a) P(s'|s,a) \nabla V_\theta(s')$ below. The Lipschitzness of $T_2(\theta)$ together with that of $T_1(\theta)$ would lead to smoothness of $U(\cdot)$, from (104).

$$
\begin{aligned}
&\|T_2(\theta_1) - T_2(\theta_2)\|_2 \\
&\leq \left\| \sum_{s,a,s'} \tilde{\nu}_{\theta_1}(s,a) P(s'|s,a) \nabla V_{\theta_1}(s') - \sum_{s,a,s'} \tilde{\nu}_{\theta_2}(s,a) P(s'|s,a) \nabla V_{\theta_2}(s') \right\| \\
&\leq \Bigg\| \sum_{s,a,s'} \tilde{\nu}_{\theta_1}(s,a) P(s'|s,a) \nabla V_{\theta_1}(s') - \sum_{s,a,s'} \tilde{\nu}_{\theta_2}(s,a) P(s'|s,a) \nabla V_{\theta_2}(s') \\
&\qquad + \sum_{s,a,s'} \tilde{\nu}_{\theta_2}(s,a) P(s'|s,a) \nabla V_{\theta_1}(s') - \sum_{s,a,s'} \tilde{\nu}_{\theta_2}(s,a) P(s'|s,a) \nabla V_{\theta_2}(s') \Bigg\| \\
&\leq \sum_{s,a,s'} P(s'|s,a) \|\nabla V_{\theta_1}(s')\|_2 \|\tilde{\nu}_{\theta_1}(s,a) - \tilde{\nu}_{\theta_2}(s,a)\| \\
&\quad + \sum_{s,a,s'} P(s'|s,a) \tilde{\nu}_{\theta_2}(s,a) \|\nabla V_{\theta_1}(s') - \nabla V_{\theta_2}(s')\|_2 \\
&\overset{(a)}{\leq} \frac{2R_{\max} C_\psi}{(1-\gamma)^2} \sum_{s,a} \|\tilde{\nu}_{\theta_1}(s,a) - \tilde{\nu}_{\theta_2}(s,a)\| \\
&\quad + \sum_{s,a,s'} P(s'|s,a) \tilde{\nu}_{\theta_2}(s,a) \|\nabla V_{\theta_1}(s') - \nabla V_{\theta_2}(s')\|_2 \\
&\overset{(b)}{\leq} \frac{2R_{\max} C_\psi C_\nu}{(1-\gamma)^2} \|\theta_1 - \theta_2\|_2 + 2L_J \|\theta_1 - \theta_2\|_2 \qquad (107)
\end{aligned}
$$

where (a) follows by $P(s'|s,a)\|\nabla V_\theta(s')\|_2 \leq \frac{R_{\max}C_\psi}{(1-\gamma)^2}$; (b) follows by using (Xu et al., 2021, Lemma 3), where $C_\nu = (1/2)C_\pi\left(1 + \lceil \log_\rho \kappa^{-1} \rceil + (1-\rho)^{-1}\right)$.

Combining T_1 and T_2 into (105),

$$\|\nabla U(\theta_1) - \nabla U(\theta_2)\| \leq L_U \|\theta_1 - \theta_2\|_2, \text{ where}$$
$$L_U = \frac{1}{1-\gamma^2}\left(\frac{R_{\max}L_\psi}{(1-\gamma)^2} + \frac{4R_{\max}C_\psi C_v}{(1-\gamma)^2} + \frac{4\gamma R_{\max}C_\psi C_v + 4\gamma L_J}{(1-\gamma)^2}\right).$$

□

12.2 Proof of Lemma 4.2

Proof. Notice that

$$\begin{aligned}
\|\nabla L(\theta_1) - \nabla L(\theta_2)\|_2 &\leq \|\nabla J(\theta_1) - \nabla J(\theta_2)\|_2 + \lambda \|\nabla U(\theta_1) - \nabla U(\theta_2)\|_2 \\
&\quad + 2\lambda \|J(\theta_1)\nabla J(\theta_1) - J(\theta_2)\nabla J(\theta_2)\|_2 \\
&\overset{(a)}{\leq} L_J \|\theta_1 - \theta_2\|_2 + \lambda L_U \|\theta_1 - \theta_2\|_2 + 2\lambda \underbrace{\|J(\theta_1)\nabla J(\theta_1) - J(\theta_2)\nabla J(\theta_2)\|_2}_{\text{(I)}},
\end{aligned} \tag{108}$$

where (a) follows by Lemma 4.1.

We bound (I) as follows:

$$\begin{aligned}
&\|J(\theta_1)\nabla J(\theta_1) - J(\theta_2)\nabla J(\theta_2)\|_2 \\
&= \|J(\theta_1)\nabla J(\theta_1) - J(\theta_1)\nabla J(\theta_2) + J(\theta_1)\nabla J(\theta_2) - J(\theta_2)\nabla J(\theta_2)\|_2 \\
&\leq |J(\theta_1)| \cdot \|\nabla J(\theta_1) - \nabla J(\theta_2)\|_2 + \|\nabla J(\theta_2)\|_2 \cdot |J(\theta_1) - J(\theta_2)| \\
&\overset{(i)}{\leq} \frac{R_{\max}L_J}{1-\gamma}\|\theta_1 - \theta_2\|_2 + \|\nabla J(\theta_2)\|_2 \cdot |J(\theta_1) - J(\theta_2)| \\
&\overset{(ii)}{\leq} \frac{R_{\max}L_J}{1-\gamma}\|\theta_1 - \theta_2\|_2 + \frac{R_{\max}C_\psi}{(1-\gamma)^2}|J(\theta_1) - J(\theta_2)| \\
&\leq \frac{R_{\max}L_J}{1-\gamma}\|\theta_1 - \theta_2\|_2 + \frac{R_{\max}C_\psi}{(1-\gamma)^2}\|\theta_1 - \theta_2\|_2,
\end{aligned} \tag{109}$$

where (i) follows by $|J(\theta)| \leq \frac{R_{\max}}{1-\gamma}$; (ii) follows by $\|\nabla J(\theta)\|_2 \leq \frac{R_{\max}C_\psi}{(1-\gamma)^2}$ for any $\theta \in \mathbb{R}^d$, we arrive at this by Policy Gradient Theorem (Sutton et al., 1999), Assumption 7 and $|Q_{\pi_\theta}(s,a)| \leq \frac{R_{\max}}{1-\gamma}$; (109) follows by taking first order Taylor expansion at θ_1, mean-value theorem $\exists\ \tilde{\theta} = \lambda\theta_1 + (1-\lambda)\theta_2$, for some $\lambda \in [0,1]$.
$J(\theta_1) = J(\theta_2) + \nabla J(\tilde{\theta})^\top(\theta_1 - \theta_2) \implies |J(\theta_1) - J(\theta_2)| \leq \frac{R_{\max}C_\psi}{(1-\gamma)^2}\|\theta_1 - \theta_2\|_2$.

Now, substituting (109) in (108), we obtain

$$\begin{aligned}
&\|\nabla L(\theta_1) - \nabla L(\theta_2)\| \leq \|\nabla J(\theta_1) - \nabla J(\theta_2)\| + 2\lambda\|J(\theta_1)\nabla J(\theta_1) - J(\theta_2)\nabla J(\theta_2)\| \\
&\qquad + \lambda\|\nabla U(\theta_1) - \nabla U(\theta_2)\| \\
&\leq L_J\|\theta_1 - \theta_2\|_2 + 2\lambda\left(\frac{R_{\max}L_J}{1-\gamma} + \frac{R_{\max}C_\psi}{(1-\gamma)^2}\right)\|\theta_1 - \theta_2\|_2 + \lambda L_U\|\theta_1 - \theta_2\|_2 \\
&\leq \left(L_J + 2\lambda\left(\frac{R_{\max}L_J}{1-\gamma} + \frac{R_{\max}C_\psi}{(1-\gamma)^2}\right) + \lambda L_U\right)\|\theta_1 - \theta_2\|_2 \\
&\leq L_o\|\theta_1 - \theta_2\|_2
\end{aligned}$$

Hence, Gradient of the Lagrangian is L-Lipschitz with $L_o = L_J + 2\lambda\left(\frac{R_{\max}L_J}{1-\gamma} + \frac{R_{\max}C_\psi}{(1-\gamma)^2}\right) + \lambda L_U$.
□

12.3 Proof of Theorem 4.3

Proof. Notice that as $\nabla L(\theta_t)$ is L-Lipschitz (Lemma 4.2), we have

$$L(\theta_{t+1}) \geq L(\theta_t) + \langle \nabla L(\theta_t), \theta_{t+1} - \theta_t \rangle - \frac{L\alpha_t^2}{2} \|\nabla \hat{L}(\theta_t)\|^2$$

Taking expectation w.r.t the sigma field $\mathcal{F}_t = \sigma(\theta_k, k \leq t)$, denoted by $\mathbb{E}_t$

$$\begin{aligned}
\mathbb{E}_t[L(\theta_{t+1})] &\geq \mathbb{E}_t[L(\theta_t)] + \mathbb{E}_t\left[\left\langle \nabla L(\theta_t), \alpha_t \nabla L(\theta_t) + \alpha_t\left(\nabla \hat{L}(\theta_t) - \nabla L(\theta_t)\right)\right\rangle\right] \\
&\quad - \mathbb{E}_t\left[\frac{L}{2}\alpha_t^2 \|\nabla \hat{L}(\theta_t)\|^2\right] \\
&= \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] + \alpha_t \mathbb{E}_t\left[\nabla L(\theta_t)^\top \left(\nabla \hat{L}(\theta_t) - \nabla L(\theta_t)\right)\right] \\
&\quad - \mathbb{E}_t\left[\frac{L}{2}\alpha_t^2 \|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\geq \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t \left|\mathbb{E}_t\left[\nabla L(\theta_t)^\top \left(\nabla \hat{L}(\theta_t) - \nabla L(\theta_t)\right)\right]\right| \\
&\quad - \mathbb{E}_t\left[\frac{L}{2}\alpha_t^2 \|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\overset{(i)}{\geq} \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t \|\nabla L(\theta_t)\| \left\|\mathbb{E}_t\left[\nabla \hat{L}(\theta_t) - \nabla L(\theta_t)\right]\right\| \\
&\quad - \mathbb{E}_t\left[\frac{L}{2}\alpha_t^2 \|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\overset{(ii)}{\geq} \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{L}(\theta_t) - \nabla L(\theta_t)\right]\right\| \\
&\quad - \frac{L}{2}\alpha_t^2 \mathbb{E}_t\left[\|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\overset{(iii)}{\geq} \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\| \\
&\quad - \lambda \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{U}(\theta_t) - \nabla U(\theta_t)\right]\right\| - 2\lambda \alpha_t K_1 \left\|\mathbb{E}_t\left[J(\theta_t)\nabla J(\theta_t) - \hat{J}(\theta_t)\nabla \hat{J}(\theta_t)\right]\right\| \\
&\quad - \frac{L}{2}\alpha_t^2 \mathbb{E}_t\left[\|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\overset{(iv)}{\geq} \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\| \\
&\quad - \lambda \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{U}(\theta_t) - \nabla U(\theta_t)\right]\right\| \\
&\quad - 2\alpha_t K_1 \lambda \left\|\mathbb{E}_t\left[J(\theta_t)\nabla J(\theta_t) - J(\theta_t)\nabla \hat{J}(\theta_t) + J(\theta_t)\nabla \hat{J}(\theta_t) - \hat{J}(\theta_t)\nabla \hat{J}(\theta_t)\right]\right\| \\
&\quad - \frac{L}{2}\alpha_t^2 \mathbb{E}_t\left[\|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\overset{(v)}{\geq} \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\| \\
&\quad - \lambda \alpha_t K_1 \left\|\mathbb{E}_t\left[\nabla \hat{U}(\theta_t) - \nabla U(\theta_t)\right]\right\| \\
&\quad - 2\alpha_t K_1 \lambda \left\|\mathbb{E}_t\left[J(\theta_t)\left(\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right)\right]\right\| - 2\alpha_t K_1 \lambda \left\|\mathbb{E}_t\left[\nabla \hat{J}(\theta_t)\left(J(\theta_t) - \hat{J}(\theta_t)\right)\right]\right\| \\
&\quad - \frac{L}{2}\alpha_t^2 \mathbb{E}_t\left[\|\nabla \hat{L}(\theta_t)\|^2\right] \\
&\overset{(vi)}{\geq} \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] - \alpha_t K_1 \left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right) \underbrace{\left\|\mathbb{E}_t\left[\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\|}_{\text{(A)}}
\end{aligned}$$

$$- \lambda\alpha_t K_1 \underbrace{\left\| \mathbb{E}_t \left[\nabla\hat{U}(\theta_t) - \nabla U(\theta_t) \right] \right\|}_{\text{(B)}} - \frac{L}{2}\alpha_t^2 \underbrace{\mathbb{E}_t \left[\|\nabla\hat{L}(\theta_t)\|^2 \right]}_{\text{(C)}}$$

$$- \alpha_t K_1 \left(\frac{2\lambda\sqrt{d}R_{\max}}{(1-\gamma)p_t} \right) \underbrace{\left| \mathbb{E}_t \left[\hat{J}(\theta_t) - J(\theta_t) \right] \right|}_{\text{(D)}}, \tag{110}$$

where (i) follows from applying the Cauchy–Schwarz inequality to the modulus of the inner product; (ii) follows from the uniform upper bound $\|\nabla L(\theta_t)\| \leq K_1$, which we establish below; (iii) follows from substituting

$$\nabla L(\theta) = -\nabla J(\theta) + \lambda(\nabla U(\theta) - 2J(\theta)\nabla J(\theta));$$

(iv) follows from adding and subtracting the cross term $J(\theta_t)\nabla\hat{J}(\theta_t)$; (v) follows from the triangle inequality; and (vi) follows from the bound $|J(\theta_t)| \leq \frac{R_{\max}}{1-\gamma}$ and $\|\nabla\hat{J}(\theta_t)\| \leq \frac{2\sqrt{d}R_{\max}}{1-\gamma}$, which is a consequence of the definition of the SPSA gradient estimate,

$$\nabla\hat{J}(\theta) = \frac{\hat{J}(\theta_t + p_t\Delta_t) - \hat{J}(\theta_t)}{p_t\Delta_t}.$$

Before we derive upper bounds for (A), (B), (C), and (D) in (110), we first establish the bound $\|\nabla L(\theta_t)\|_2 \leq K_1$, which is used in (ii), as follows:

By Policy Gradient Theorem (Sutton et al., 1999), we have

$$\nabla J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim\chi_\theta(\cdot,\cdot)}\left[\nabla\log\pi_\theta(a|s)Q_{\pi_\theta}(s,a)\right],$$

where

$$Q_{\pi_\theta}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) \mid s_0 = s, a_0 = a\right].$$

We upper bound the action-value function as $|Q_{\pi_\theta}(s,a)| \leq \frac{R_{\max}}{1-\gamma}$. Furthermore, by Assumption 7, the score function satisfies $\|\nabla\log\pi_\theta(a|s)\|_2 \leq C_\psi$. Thus, we obtain

$$\|\nabla J(\theta)\|_2 \leq \frac{R_{\max}C_\psi}{(1-\gamma)^2}, \quad \forall\theta\in\mathbb{R}^d. \tag{111}$$

In the same manner, we use (104), which is a policy gradient-style theorem for the square-value function from (L.A. & Ghavamzadeh, 2016, Lemma 1), to upper bound the norm of the square-value function below. $W_{\pi_\theta}(s,a)$ is the action-value function corresponding to the square-value function, i.e., $U(\theta) = \mathbb{E}_{a\sim\pi_\theta}[W_{\pi_\theta}(s,a)]$, similar to $Q_{\pi_\theta}(s,a)$.

$$\begin{aligned}
&\|\nabla U(\theta)\|_2 \\
&= \frac{1}{1-\gamma^2}\left\|\sum_{s,a}\tilde{\nu}_{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)W_{\pi_\theta}(s,a) + 2\gamma\sum_{s,a,s'}\tilde{\nu}_{\pi_\theta}(s,a)P(s'|s,a)\nabla V_{\pi_\theta}(s')\right\| \\
&\leq \frac{1}{1-\gamma^2}\sum_{s,a}\|\tilde{\nu}_{\pi_\theta}(s,a)\nabla\log\pi_\theta(a|s)\|\,|W_{\pi_\theta}(s,a)| \\
&\quad + \frac{2\gamma}{1-\gamma^2}\sum_{s,a,s'}\|\tilde{\nu}_{\pi_\theta}(s,a)\||P(s'|s,a)|\|\nabla V_{\pi_\theta}(s')\| \\
&\leq \frac{1}{1-\gamma^2}\|\nabla\log\pi_\theta(a|s)\|\sum_{s,a}\tilde{\nu}_{\pi_\theta}(s,a)W_{\pi_\theta}(s,a) \\
&\quad + \frac{2\gamma}{1-\gamma^2}\sum_{s,a,s'}\tilde{\nu}_{\pi_\theta}(s,a)P(s'|s,a)\|\nabla V_{\pi_\theta}(s')\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{C_\psi}{1-\gamma^2}\sum_{s,a}\tilde{\nu}_{\pi_\theta}(s,a)W_{\pi_\theta}(s,a) + \frac{2\gamma}{1-\gamma^2}\sum_{s,a,s'}\tilde{\nu}_{\pi_\theta}(s,a)P(s'|s,a)\|\nabla V_{\pi_\theta}(s')\| \\
&\leq \frac{C_\psi R_{\max}}{(1-\gamma^2)(1-\gamma)^2} + \frac{2\gamma R_{\max} C_\psi}{(1-\gamma^2)(1-\gamma)^2} \qquad (112)
\end{aligned}
$$

968 Combining (111) and (112), we obtain K_1:

$$
\begin{aligned}
\|\nabla L(\theta_t)\| \leq & \|\nabla J(\theta_t)\| + \lambda\|\nabla U(\theta_t)\| + 2\lambda|J(\theta_t)|\|\nabla J(\theta_t)\| \\
\leq & \frac{R_{\max}C_\psi}{(1-\gamma)^2} + 2\lambda\frac{R_{\max}C_\psi}{(1-\gamma)^3} + \lambda\|\nabla U(\theta_t)\| \\
\leq & \frac{R_{\max}C_\psi}{(1-\gamma)^2} + 2\lambda\frac{R_{\max}C_\psi}{(1-\gamma)^3} + \lambda\left(\frac{C_\psi R_{\max}}{(1-\gamma^2)(1-\gamma)^2} + \frac{2\gamma R_{\max}C_\psi}{(1-\gamma^2)(1-\gamma)^2}\right) \\
= & K_1
\end{aligned}
$$

969 Next, we bound (A) in (110) as follows:

$$
\begin{aligned}
&\left\|\mathbb{E}_t\left[\nabla\hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\| \leq d^{\frac{1}{2}}\left|\mathbb{E}_t\left[\nabla_i\hat{J}(\theta_t) - \nabla_i J(\theta_t)\right]\right| \\
&\left|\mathbb{E}_t\left[\nabla_i\hat{J}(\theta_t) - \nabla_i J(\theta_t)\right]\right| \overset{(a)}{=} \left|\mathbb{E}_t\left[\frac{\phi_v(s_0)^\top v_m^+ - \phi_v(s_0)^\top v_m}{p_t\Delta_i(t)} - \nabla_i J(\theta_t)\right]\right| \\
&\overset{(b)}{=} \left|\mathbb{E}_t\left[\frac{\phi_v(s_0)^\top v_m^+ - \phi_v(s_0)^\top v_m + \phi_v(s_0)^\top\bar{v}^+ - \phi_v(s_0)^\top\bar{v}^+ - \phi_v(s_0)^\top\bar{v} + \phi_v(s_0)^\top\bar{v}}{p_t\Delta_i(t)} - \nabla_i J(\theta_t)\right]\right| \\
&\overset{(c)}{=} \left|\mathbb{E}_t\left[\frac{\phi_v(s_0)^\top(\bar{v}^+ - \bar{v})}{p_t\Delta_i(t)} + \frac{\phi_v(s_0)^\top(v_m^+ - \bar{v}^+) + \phi_v(s_0)^\top(\bar{v} - v_m)}{p_t\Delta_i(t)} - \nabla_i J(\theta_t)\right]\right| \\
&\leq \underbrace{\left|\mathbb{E}_t\left[\frac{J(\theta_t + p_t\Delta(t)) - J(\theta_t)}{p_t\Delta_i(t)} - \nabla_i J(\theta_t)\right]\right|}_{\text{(I)}} + \underbrace{\left|\mathbb{E}_t\left[\frac{\phi_v(s_0)^\top(v_m^+ - \bar{v}^+) + \phi_v(s_0)^\top(\bar{v} - v_m)}{p_t\Delta_i(t)}\right]\right|}_{\text{(II)}}, \qquad (113)
\end{aligned}
$$

970 where (a) follows by substituting value of SPSA gradient estimate $\nabla_i\hat{J}(\theta_t)$; (b) follows adding
971 and subtracting $\phi_v(s_0)^\top\bar{v}^+$ and $\phi_v(s_0)^\top\bar{v}$, where, $\bar{v}$ and $\bar{v}^+$ denote fixed points for unperturbed
972 and perturbed policies, respectively; (c) follows by rearranging the terms; (113) follows by (critic
973 approximation error at the fixed point is zero) Assumption 8, as a consequence, the first term in (I)
974 is equal to the actual value function.

975 We bound (I) in (113) as follows:

$$
\begin{aligned}
&\left|\mathbb{E}_t\left[\frac{J(\theta_t + p_t\Delta_i(t)) - J(\theta_t)}{p_t\Delta_i(t)} - \nabla_i J(\theta_t)\right]\right| \\
&\overset{(a)}{\leq} \left|\mathbb{E}_t\left[\frac{p_t(\nabla J(\theta_t))^\top\Delta(t) + \frac{L_J}{2}p_t^2\|\Delta(t)\|^2}{\Delta_i(t)p_t} - \nabla_i J(\theta_t)\right]\right| \\
&\overset{(b)}{\leq} \left|\mathbb{E}_t\left[\sum_{j\neq i}\left(\frac{\Delta_j(t)}{\Delta_i(t)}\right)\nabla_j J(\theta_t)\right]\right| + \left|\mathbb{E}_t\left[\frac{L_J p_t\|\Delta(t)\|^2}{2}\right]\right| \\
&\overset{(c)}{\leq} \frac{dL_J p_t}{2}, \qquad (114)
\end{aligned}
$$

976 where (a) follows from the second-order Taylor expansion of $J(\theta_t + p_t\Delta_i(t))$ around θ_t, lever-
977 aging the fact that $J(\theta)$ has a Lipschitz gradient (with constant L_J) to bound the quadratic term;
978 (b) follows from the triangle inequality and expanding the inner product into a summation over
979 components. Here, the first term has an expectation of zero because $\Delta(t)$ is a Rademacher vector.
980 Specifically, each component $\Delta_j(t)$ satisfies $\mathbb{E}_t[\Delta_j(t)] = 0$, and the independence of $\Delta_j(t)$ and

$\Delta_i(t)$ ensures that the expectation of the ratio $\frac{\Delta_j(t)}{\Delta_i(t)}$ is also zero. By the linearity of expectation, the entire summation contributes zero in expectation; (c) follows by bounding $\|\Delta(t)\| \leq \sqrt{d}$.

We bound (II) in (113) as follows:

$$\begin{aligned}
&\left|\mathbb{E}_t\left[\frac{\phi_v(s_0)^\top(v_m^+ - \bar{v}^+) + \phi_v(s_0)^\top(\bar{v} - v_m)}{p_t\Delta_i(t)}\right]\right| \\
&\overset{(a)}{\leq} \left|\mathbb{E}_t\left[\frac{\|\phi_v(s_0)\|\|v_m^+ - \bar{v}^+\| + \|\phi_v(s_0)\|\|\bar{v} - v_m\|}{p_t\Delta_i(t)}\right]\right| \\
&\overset{(b)}{\leq} \frac{\phi^v_{\mathsf{max}}}{p_t}\left(\mathbb{E}_t\left[\|v_m^+ - \bar{v}^+\|\right] + \mathbb{E}_t\left[\|v_m - \bar{v}\|\right]\right) \\
&\overset{(c)}{\leq} \frac{\phi^v_{\mathsf{max}}}{p_t\sqrt{m}}\underbrace{\left(\frac{10^{\frac{1}{2}}e^{\frac{-k\beta\mu}{2}}}{\gamma^2\mu}\left(\max_{\theta_{i=1,\ldots,n}}\mathbb{E}[\|w_0 - \bar{w}\|]\right)^{\frac{1}{2}} + \frac{10^{\frac{1}{2}}\sigma}{\mu}\right)}_{K_2} \\
&\overset{(d)}{\leq} \frac{\phi^v_{\mathsf{max}}K_2}{p_t\sqrt{m}},
\end{aligned} \tag{115}$$

where (a) follows from the Cauchy-Schwarz inequality; (b) follows from the upper bound on the norm of the features (Assumption 3) and linearity of expectation; (c) follows by bounding the terms using the tail-averaged critic error bound in (11); (d) follows by defining K_2 in step (c).

Combining (114) and (115) in (113), we obtain an upper bound for (A) in (110) as:

$$\left\|\mathbb{E}_t\left[\nabla\hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\| \leq \frac{d^{\frac{3}{2}}L_J p_t}{2} + \frac{d^{\frac{1}{2}}\phi^v_{\mathsf{max}}K_2}{p_t\sqrt{m}}. \tag{116}$$

We obtain the upper bound for (B) in (110) using arguments parallel to those used to derive the upper bound for (A). The only difference lies in the feature vector, where ϕ^u_{max} replaces ϕ^v_{max}.

$$\left\|\mathbb{E}_t\left[\nabla\hat{U}(\theta_t) - \nabla U(\theta_t)\right]\right\| \leq \frac{d^{\frac{3}{2}}L_U p_t}{2} + \frac{d^{\frac{1}{2}}\phi^u_{\mathsf{max}}K_2}{p_t\sqrt{m}}. \tag{117}$$

Next, we bound (C) in (110) as follows:

The SPSA gradient estimate of the Lagrangian is denoted as

$$\nabla\hat{L}(\theta_t) = \nabla\hat{J}(\theta_t) - \lambda\left(\nabla\hat{U}(\theta_t) - 2\hat{J}(\theta_t)\nabla\hat{J}(\theta_t)\right).$$

Taking the expectation with respect to the sigma field $\mathcal{F}_t = \sigma(\theta_k, k \leq t)$, denoted by $\mathbb{E}_t$, we have

$$\begin{aligned}
\mathbb{E}_t[\|\nabla\hat{L}(\theta_t)\|_2^2] &\overset{(a)}{\leq} 3\mathbb{E}_t[\|\nabla\hat{J}(\theta_t)\|_2^2] + 3\lambda^2\mathbb{E}_t[\|\nabla\hat{U}(\theta_t)\|_2^2] + 12\lambda^2\left(\frac{R_{\mathsf{max}}}{1-\gamma}\right)^2\mathbb{E}_t[\|\nabla\hat{J}(\theta_t)\|_2^2] \\
&\overset{(b)}{\leq} \max\left\{3 + 3\left(\frac{2\lambda R_{\mathsf{max}}}{1-\gamma}\right)^2, 3\lambda^2\right\}\left(\|\nabla\hat{J}(\theta_t)\|_2^2 + \|\nabla\hat{U}(\theta_t)\|_2^2\right) \\
&\overset{(c)}{\leq} \max\left\{3 + 3\left(\frac{2\lambda R_{\mathsf{max}}}{1-\gamma}\right)^2, 3\lambda^2\right\}\left(d\left(\frac{2R_{\mathsf{max}}}{1-\gamma}\right)^2\frac{1}{p_t^2} + d\left(\frac{2R^2_{\mathsf{max}}}{(1-\gamma)^2}\right)^2\frac{1}{p_t^2}\right) \\
&\overset{(d)}{=} \frac{K_3}{p_t^2},
\end{aligned} \tag{118}$$

where (a) follows from $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$; (b) follows by taking the maximum of all coefficients; (c) follows by bounding the SPSA gradient estimate $\left\|\frac{J(\theta_t + p_t\Delta_i(t)) - J(\theta_t)}{p_t\Delta_i(t)}\right\|^2 \leq$

$\left(\frac{2R_{\max}}{(1-\gamma)p_t}\right)^2$ for the first term and similarly bounding the SPSA gradient estimate of the square-value function for the second term; and (d) follows by defining K_3 as a constant, which is the coefficient of $\frac{1}{p_t^2}$ in (c).

Now, substituting the bounds obtained for (A) in (116), (B) in (117), and (C) in (118) into (110), we get

$$\begin{aligned}
\mathbb{E}_t[L(\theta_{t+1})] \geq\ & \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] \\
& - \alpha_t K_1 \left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right) \underbrace{\left\|\mathbb{E}_t\left[\nabla \hat{J}(\theta_t) - \nabla J(\theta_t)\right]\right\|}_{\text{(A)}} \\
& - \lambda \alpha_t K_1 \underbrace{\left\|\mathbb{E}_t\left[\nabla \hat{U}(\theta_t) - \nabla U(\theta_t)\right]\right\|}_{\text{(B)}} - \underbrace{\frac{L}{2}\alpha_t^2 \mathbb{E}_t\left[\|\nabla \hat{L}(\theta_t)\|^2\right]}_{\text{(C)}} \\
\geq\ & \mathbb{E}_t[L(\theta_t)] + \alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|\right] - \alpha_t K_1 \left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right)\left(\frac{d^{\frac{3}{2}} L_J p_t}{2} + \frac{d^{\frac{1}{2}} \phi^v_{\max} K_2}{p_t\sqrt{m}}\right) \\
& - \lambda \alpha_t K_1 \left(\frac{d^{\frac{3}{2}} L_U p_t}{2} + \frac{d^{\frac{1}{2}} \phi^u_{\max} K_2}{p_t\sqrt{m}}\right) - \frac{L\alpha_t^2}{2}\left(\frac{K_3}{p_t^2}\right)
\end{aligned}$$

Rearranging the terms, we obtain

$$\begin{aligned}
\alpha_t \mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] \leq\ & \mathbb{E}_t[L(\theta_{t+1})] - \mathbb{E}_t[L(\theta_t)] \\
& + \alpha_t K_1 \left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right)\left(\frac{d^{\frac{3}{2}} L_J p_t}{2} + \frac{d^{\frac{1}{2}} \phi^v_{\max} K_2}{p_t\sqrt{m}}\right) \\
& + \lambda \alpha_t K_1 \left(\frac{d^{\frac{3}{2}} L_U p_t}{2} + \frac{d^{\frac{1}{2}} \phi^u_{\max} K_2}{p_t\sqrt{m}}\right) + \frac{L_1 \alpha_t^2 K_3}{2p_t^2} \\
\overset{(a)}{\leq}\ & \mathbb{E}[H_t] - \mathbb{E}[H_{t+1}] + \frac{\alpha_t K_1 d^{\frac{3}{2}}}{2}\left(L_J\left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right) + \lambda L_U\right) p_t \\
& + \alpha_t K_1 K_2 d^{\frac{1}{2}} \left(\left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right)(\phi^v_{\max} + \lambda \phi^u_{\max})\right)\frac{1}{p_t\sqrt{m}} + \frac{\alpha_t^2 L_1 K_3}{2p_t^2},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_t\left[\|\nabla L(\theta_t)\|^2\right] \overset{(b)}{\leq}\ & \frac{1}{\alpha_t}\left(\mathbb{E}\left[H_{t+1}\right] - \mathbb{E}\left[H_t\right]\right) + \frac{K_1 d^{\frac{3}{2}}}{2}\left(L_J\left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right) + \lambda L_U\right) p_t \\
& + K_1 K_2 d^{\frac{1}{2}} \left(\left(1 + \frac{2\lambda R_{\max}}{1-\gamma}\right)(\phi^v_{\max} + \lambda \phi^u_{\max})\right)\frac{1}{p_t\sqrt{m}} + \frac{\alpha_t L_1 K_3}{2p_t^2},
\end{aligned}$$

where (a) follows by taking $H_t = L(\theta_t) - L(\theta_*)$, where θ^* is the optimal policy, and (b) follows by dividing both sides by α_t.

Summing from $t = 1$ to n, and taking the total expectation, we get

$$\sum_{t=1}^{n} \mathbb{E}\left[\|\nabla L(\theta_t)\|^2\right] \leq \frac{C_1}{\alpha_t} + C_2 \sum_{t=1}^{n} p_t + \frac{C_3}{\sqrt{m}} \sum_{t=1}^{n} \frac{1}{p_t} + C_4 \sum_{t=1}^{n} \frac{\alpha_t}{p_t^2}.$$

Here, we obtain $|L(\theta)| \leq C_1 = \frac{2R_{\max}}{1-\gamma}\left(1 + \frac{\lambda R_{\max}}{1-\gamma}\right)$ after a telescoping sum.

Dividing by n on both sides and setting $\alpha_t = \alpha, p_t = p$, we get

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}\left[\|\nabla L(\theta_t)\|^2\right] \leq \frac{C_1}{n\alpha} + C_2 p + \frac{C_3}{\sqrt{mp}} + \frac{C_4 \alpha}{p^2}.$$

Setting $\alpha = n^a$, $p = n^b$, $m = n^c$, we have

$$\mathbb{E}\left[\|\nabla L(\theta_R)\|^2\right] \leq C_1 n^{-1-a} + C_2 n^b + C_3 n^{-b-c/2} + C_4 n^{a-2b}.$$

Optimizing for a, b, c, we find their values to be $a = -\frac{3}{4}$, $b = -\frac{1}{4}$, $c = 1$. Substituting these values, we get

$$\begin{aligned}\mathbb{E}\left[\|\nabla L(\theta_R)\|^2\right] &\leq C_1 n^{-1/4} + C_2 n^{-1/4} + C_3 n^{-1/4} + C_4 n^{-1/4} \\ &= O(n^{-1/4}).\end{aligned}$$

□