Towards Parameter-Free Temporal Difference Learning

Yunxiang Li
University of British Columbia

Mark Schmidt
University of British Columbia, Canada CIFAR AI Chair (Amii)

Reza Babanezhad
Samsung AI - Montreal

Sharan Vaswani
Simon Fraser University

Abstract

Temporal difference (TD) learning is a fundamental algorithm for estimating value functions in reinforcement learning. Recent finite-time analyses of TD with linear function approximation quantify its theoretical convergence rate. However, they typically require algorithm parameters that depend on problem-dependent quantities that are difficult to estimate in practice, such as the minimum eigenvalue of the feature covariance (ω), or the mixing time of the underlying Markov chain (τ_{mix}). Furthermore, some analyses require non-standard, impractical algorithmic modifications, limiting adoption. We address these limitations by using an exponential step-size schedule with the standard TD(0) algorithm. We analyze this method under two standard sampling regimes: independent and identically distributed (i.i.d.) sampling from the stationary distribution, and the more practical Markovian sampling from a single trajectory. Unlike previous works in the i.i.d. setting, the proposed algorithm requires no knowledge of such problem-dependent quantities and simultaneously attains the optimal bias-variance trade-off for the last iterate. In the Markovian setting, we propose a regularized TD(0) algorithm with exponential step-size schedule that achieves a similar convergence rate as previous works while requiring no projections, no averaging, and no knowledge of τ_{mix} or ω.

1 Introduction

Reinforcement learning (RL) is a general framework for sequential decision making under uncertainty and has been successful in various real-world applications, such as robotics (Kober et al., 2013) and aligning language models (Uc-Cetina et al., 2023). Value functions underpin value-based algorithms (Sutton and Barto, 2018) and play an important role in actor-critic methods (Konda and Tsitsiklis, 1999). Hence, efficiently calculating the value function for a given policy is a key building block in RL.

Temporal-difference (TD) learning (Sutton, 1988) is an incremental policy-evaluation algorithm that iteratively updates a value-function estimate via bootstrapping. It is efficient to implement and, as opposed to the tabular case, scales to large state-action spaces with linear function approximation. Given the algorithm's importance, a large body of theoretical work analyzes the convergence of TD with linear function approximation (Tsitsiklis and Roy, 1997; Dalal et al., 2018; Lakshminarayanan and Szepesvári, 2018; Mou et al., 2020; Bhandari et al., 2018; Patil et al., 2023; Samsonov et al., 2024) and its variants (Liu and Olshevsky, 2021; Mustafin et al., 2024). However, many analyzed

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

algorithms (i) rely on difficult-to-estimate problem-dependent parameters and/or (ii) use nonstandard modifications such as projections onto a ball or iterate averaging. These modifications are uncommon in practice, leaving a gap between theory and practice. In this paper, we aim to *design a theoretically principled TD algorithm with minimal modifications that does not require any difficult-to-estimate problem-dependent constants*.

To this end, we first consider the independent and identically distributed (i.i.d.) sampling regime, where states are sampled from the stationary distribution of the underlying Markov chain for the evaluated policy. The i.i.d. sampling regime is often used as a testbed for designing and analyzing policy-evaluation algorithms (Lakshminarayanan and Szepesvári, 2018; Dalal et al., 2018; Bhandari et al., 2018; Patil et al., 2023; Samsonov et al., 2024). In this regime, Dalal et al. (2018) analyze the TD(0) algorithm using tools from stochastic approximation. They explicitly trade off bias (the rate at which the influence of initialization vanishes) and variance (from i.i.d. sampling) for the last iterate. Follow-up work by Bhandari et al. (2018) relates TD to stochastic gradient descent (SGD) and uses optimization tools to analyze TD. They study three step-size schedules (see Table 1) under i.i.d. sampling. Some require knowledge of the smallest eigenvalue of the state-weighted feature covariance ω, while others yield slower rates. While these rates hold for the last iterate, they do not achieve the optimal bias-variance trade-off. More recent work (Patil et al., 2023; Samsonov et al., 2024) uses tail averaging to achieve the optimal bias-variance trade-off without problem-dependent constants, but averaging iterates differs from typical practical implementations of TD.

Contribution 1. For TD(0) with linear function approximation under i.i.d. sampling, we take an optimization lens similar to Bhandari et al. (2018) and develop a TD algorithm that uses exponentially decaying step-sizes (Li et al., 2021). Such exponential decaying step-sizes have been used with SGD for minimizing smooth, strongly convex objectives (Vaswani et al., 2022). Although the TD(0) update shares certain properties with SGD, it is not the gradient of a fixed objective. Nevertheless, we prove that TD(0) with exponentially decaying step-sizes achieves the optimal bias-variance trade-off for the last iterate without requiring knowledge of problem-dependent constants such as ω (Section 3).

Direct access to the stationary distribution is unrealistic, so the i.i.d. regime is impractical. Consequently, many theoretical works analyze TD(0) with Markovian sampling (Bhandari et al., 2018; Samsonov et al., 2024; Patil et al., 2023; Mou et al., 2020). In this setting, data are collected along a single Markovian trajectory, introducing temporal dependence that complicates analysis. To enable analysis, prior work often assumes fast mixing so the state distribution approaches stationarity exponentially quickly. Under this assumption, Bhandari et al. (2018) analyze a projected variant of TD(0) under three step-size schedules. Similar to the i.i.d. case, the algorithm does not achieve the optimal trade-off between bias and dependence on the mixing time. Moreover, the projection step is nonstandard in practice and requires ω. More recent Markovian analyses fall into two categories: (i) Srikant and Ying (2019); Mitra (2025), which control correlations between consecutive samples and prove convergence without projection; and (ii) Samsonov et al. (2024); Patil et al. (2023), which study TD with data drop, a nonstandard variant that does not explicitly analyze the consecutive sample correlations. Both approaches can achieve the optimal trade-off between bias and mixing-time dependence, but they require knowledge of the mixing time (hard to estimate) and prove only average-iterate convergence. Furthermore, algorithms that discard samples (Samsonov et al., 2024; Patil et al., 2023) are sample-inefficient.

Contribution 2. In the Markovian sampling regime, we show that standard TD(0) with linear function approximation, using similar exponentially decaying step-sizes, achieves the optimal bias-mixing time trade-off. Furthermore, the algorithm requires neither impractical modifications, such as such as projections, iterate averaging or data drop, nor knowledge of the mixing time (Section 4.1). However, it still requires ω. To remove this requirement, we consider a regularized variant of TD(0) (Patil et al., 2023) with exponentially decaying step-sizes. Unlike Patil et al. (2023), who use regularization to improve constants, we use it to make the algorithm parameter free. In Section 4.2, we show that regularized TD(0) retains the benefits of standard TD(0) while removing the need for any hard-to-estimate problem-dependent constants.

Table 1 compares our results with prior work by convergence rate, required parameters, projection, and whether average- or last-iterate convergence is guaranteed. The rest of the paper is organized as follows: Section 2 formalizes the problem and notation and introduces TD(0) with exponentially decaying step-sizes. Section 3 presents the i.i.d. analysis. Section 4 extends the analysis to Markovian sampling. Section 4.2 establishes our parameter-free regularized TD(0) guarantees.

Sampling	Step-size	Convergence rate	Parameters needed	Projection	Last or Average iterate convergence
i.i.d.	$O(t+1)^{-z}, z \in (0,1)$ (Dalal et al., 2018)	$O\big(\exp(-\omega T^{1-z}) + 1/T^z\big)$	None	No	Last
	$1/\sqrt{T}$ (Bhandari et al., 2018)	$O\Big(\frac{\sigma^2}{\sqrt{T}}\Big)$	None	No	Average
	$O(\omega)$ (Bhandari et al., 2018)	$O\big(\exp(-\omega^2 T) + \sigma^2\big)$	ω	No	Last
	$O\Big(\frac{1}{1+t\,\omega}\Big)$ (Bhandari et al., 2018)	$O\Big(\frac{\sigma^2}{T\omega}\Big)$	ω	No	Last
	$O(1)$ (Samsonov et al., 2024)	$\tilde{O}\Big(\exp(-\omega T) + \frac{\sigma^2}{\omega^2 T}\Big)$	None	No	Average
	$O\Big(\frac{1}{T}^{t/T}\Big)$ (**Ours**)	$\tilde{O}\Big(\exp(-\omega T) + \frac{\sigma^2}{\omega^2 T}\Big)$	None	No	Last
Markovian samples	$1/\sqrt{T}$ (Bhandari et al., 2018)	$O\Big(\frac{\left(1+\tau_{\text{mix}}(1/\sqrt{T})\right)}{\omega^2\sqrt{T}}\Big)$	No	Yes	Average
	$O(1/\omega)$ (Bhandari et al., 2018)	$O(\exp(-2\eta\omega\, T)) + O\Big(\frac{\eta\left(1+\tau_{\text{mix}}(\eta)\right)}{\omega^3}\Big)$	ω	Yes	Last
	$O(1/(\omega\,(t+1)))$ (Bhandari et al., 2018)	$O\Big(\frac{\left(1+\tau_{\text{mix}}(\alpha_T)\right)}{\omega^3}\cdot\frac{1+\log T}{T}\Big)$	ω	Yes	Average
	$O(1)$ for TD with data drop (Samsonov et al., 2024)	$\tilde{O}\big(\exp(-\omega T) + \frac{\tau_{\text{mix}}}{\omega^2 T}\big)$	τ_{mix}	No	Average
	$O\Big(\frac{\omega}{\tau_{\text{mix}}}\Big)$ (Mitra, 2025)	$O\Big(\exp\Big(-\frac{\omega^2(T+1)}{\tau_{\text{mix}}}\Big)\Big) + \tilde{O}\Big(\frac{\tau_{\text{mix}}}{\omega^2(T+1)}\Big)$	$\tau_{\text{mix}}, \omega$	No	Average
	$O\Big(\frac{1}{\ln T}\frac{1}{T}^{t/T}\Big)$ (**Ours**)	$O\Big(\exp\Big(-\frac{\omega T}{\ln^3(T)}\Big) + \frac{\ln^4(T)}{\omega^2 T}\exp\Big(\frac{m}{\ln(1/\rho)}\Big)\Big)$	ω	No	Last
	$O\Big(\frac{1}{\sqrt{T}\ln T}\frac{1}{T}^{t/T}\Big)$ for regularized TD (**Ours**)	$O\Big(\exp\Big(-\frac{\omega\sqrt{T}}{\ln^3(T)}\Big) + \frac{\ln^3(T)}{\omega^2 T}\exp\Big(\frac{m}{\ln(1/\rho)}\Big)\Big)$	None	No	Last

Table 1: Comparison of our method and other methods. ω is the smallest eigenvalue of the state feature covariance matrix, T is the number of updates, τ_{mix} is the mixing time of the Markov chain as defined in Eq. (4), m and ρ are τ_{mix} related constants defined in Definition 4.1, Our i.i.d. result and our regularized TD(0) result under Markovian sampling require no projections, no prior knowledge of τ_{mix} or ω, and no iterate averaging.

2 Problem Formulation

In this section, we formalize the setting and notation: the Markov decision process (MDP) and TD(0), linear value-function approximation and assumptions, and an exponential step-size schedule.

Markov decision process. We consider a discounted MDP $M = (\mathcal{S}, \mathcal{A}, \pi, P_\pi, \mu_0, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, π is a fixed policy mapping each state $s \in \mathcal{S}$ to a distribution over actions in $\mathcal{A}$, $P_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the transition matrix induced by π with entries $(P_\pi)_{ij} \triangleq \mathbb{P}(s_{t+1} = s_j \mid s_t = s_i)$, μ_0 is the initial state distribution, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in (0, 1]$ is the discount factor. At time t, an action $a_t \sim \pi(\cdot \mid s_t)$ is selected, the next state $s_{t+1} \sim P_\pi(\cdot \mid s_t)$ is sampled, and a reward $r(s_t, a_t)$ is received. Because the policy π is fixed, we define the expected immediate reward as $r(s) \triangleq \mathbb{E}_{a\sim\pi(\cdot|s)}[r(s, a)]$. For simplicity, we assume $r(s) \in [0, 1]$. Iterating this interaction produces a trajectory $\tau = (s_0, a_0, s_1, \dots)$ with distribution $p_\pi(\tau)$ under policy π. Let μ_π denote the stationary state distribution induced by π. The initial state distribution μ_0 may differ from μ_π. The state distribution at time t is $P_\pi^t \mu_0$. The value function V^π gives the expected cumulative discounted reward starting from μ_0 and following policy π i.e. $V^\pi(s) = \mathbb{E}_{\tau\sim p_\pi(\cdot|\mu_0)}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t)\right]$.

Sampling is considered in three regimes: mean-path, i.i.d., and Markovian. In the mean-path setting, expectations are evaluated exactly under μ_π. In the i.i.d. setting, samples are independent draws from μ_π. In the Markovian setting, the process starts from μ_0 and evolves along a single trajectory.

With these sampling regimes in place, we now turn to policy evaluation for π using TD methods. Given data from the Markov chain induced by π, TD learning estimates the value function. In particular, given a sampled transition (s_t, a_t, s_t') at iteration t, TD with one-step bootstrapping (referred to as TD(0)) updates the value estimate using the following equation:

$$V(s_t) \leftarrow V(s_t) + \eta_t \left(r(s_t) + \gamma V(s_t') - V(s_t)\right), \qquad \text{(TD(0) update)}$$

where $\eta_t > 0$ is the step-size at time t. For simplicity, we omit the superscript π in V^π.

Linear value function approximation. The above TD(0) update is done on a per-state basis, and becomes computationally expensive for MDPs with a large state space. Consequently, previous works (Tsitsiklis and Roy, 1997; Bhandari et al., 2018) consider a linear approximation of the value function. Specifically, for parameters $w \in \mathbb{R}^d$ to be estimated, these works assume that $V_w(s) = w^\top \phi(s)$, where $\phi(s) \in \mathbb{R}^d$ is the known feature vector of state s. Given a sampled transition (s_t, a_t, s'_t), the linear TD(0) update (Sutton, 1988) is given as:

$$\begin{aligned} w_{t+1} &= w_t + \eta_t \left(r(s_t) + \gamma w_t^\top \phi(s'_t) - w_t^\top \phi(s_t) \right) \phi(s_t) \\ &= w_t + \eta_t g_t(w_t) \,. \end{aligned} \tag{1}$$

where $g_t(w_t) := \left(r(s_t) + \gamma w_t^\top \phi(s'_t) - w_t^\top \phi(s_t) \right) \phi(s_t)$ is the TD(0) direction at iteration t and η_t is the corresponding step-size. To analyze convergence, we adopt the following standard MDP assumptions:

Assumption 2.1. The Markov chain induced by policy π is irreducible and aperiodic, and there exists a unique stationary distribution μ_π.

It is convenient to define the expected TD(0) update direction $g(w)$ where the expectation is over the stationary distribution μ_π. In particular, $g(w) := \mathbb{E}_{s\sim\mu_\pi, s'\sim P(\cdot|s)} \left[\phi(s) \left(r(s) + (\gamma\phi(s') - \phi(s))^\top w \right) \right]$, referred to as the *mean-path update*.

The next assumption concerns the feature vectors used in the linear approximation. Let n denote the number of states. Define $\Phi \in \mathbb{R}^{n\times d}$ as the feature matrix whose i-th row is $\phi(s_i)^\top$, and $D := \mathrm{diag}(\mu_\pi(s_1), \ldots, \mu_\pi(s_n))$ as the diagonal matrix of stationary state probabilities. Finally, let $\Sigma := \Phi^\top D \Phi$, and let ω denote its smallest eigenvalue.

Assumption 2.2. The feature matrix $\Phi = [\phi(s_1)^\top, \ldots, \phi(s_n)^\top] \in \mathbb{R}^{n\times d}$ has full column rank, which ensures a unique solution w^*. In addition, $\|\phi(s)\|^2 \le 1$ for all s.

Under Assumption 2.1 and Assumption 2.2, TD(0) with suitable step-size η_t converges to the unique fixed point w^* with $g(w^*) = 0$ (Bhandari et al., 2018). We next study how to choose η_t in the TD(0) update (1) for robust and efficient convergence.

Exponential step-size schedule. We adopt an exponential schedule (Li et al., 2021). For a fixed number of iterations T, set $\eta_t = \eta_0 \, \alpha^t$ with $\alpha = (1/T)^{1/T}$. This schedule is effective for smooth, strongly convex problems and adapts to noise without prior knowledge of its level.

3 Exponential step-size with i.i.d. sampling

Prior analyses under the i.i.d. sampling either set step sizes using problem-dependent constants (Bhandari et al., 2018; Mustafin et al., 2024) or provide guarantees only for an averaged iterate (Bhandari et al., 2018; Samsonov et al., 2024), limiting practical utility. We adopt an exponential step-size schedule and develop a problem-dependent parameter-free variant of TD(0) for i.i.d. sampling, and establish a last-iterate guarantee. We first present optimization style lemmas that we will use, and then show how the exponential schedule delivers our objective.

We first provide an one step expansion as follows.

$$\begin{aligned} &\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] = \|w_t - w^*\|^2 \\ &+ 2\eta_t \mathbb{E}_{s_t\sim\mu_\pi}\left[g_t(w_t)^\top (w_t - w^*)\right] + \eta_t^2 \mathbb{E}_{s_t\sim\mu_\pi}\left[\|g_t(w_t)\|^2\right]. \end{aligned}$$

The expectation is taken over the i.i.d. sampling from the stationary distribution μ_π. We omit the subscript in the following for brevity. Following we provide lemmas that give upper bounds on the red and blue terms. The red term can be analyzed using the following lemmas.

Lemma 3.1. *[Lemma 3 from Bhandari et al. (2018)] Under the i.i.d. sampling,* $\mathbb{E}\left[(w^* - w)^\top g_t(w)\right] \ge (1-\gamma)\|V_w - V_{w^*}\|_D^2, \forall w, w^* \in \mathbb{R}^d$.

This is analogous to a one-point strong monotonicity (or strong convexity) condition in optimization analysis. It lower bounds the alignment between the direction $w^* - w_t$ from the optimum to the current iterate, and the update direction $g_t(w)$ by the value-space error $\|V_w - V_{w^*}\|_D^2$.

Lemma 3.2. *[Extended Lemma 1 from Bhandari et al. (2018)] Under the i.i.d. sampling,* $\omega \|w_1 - w_2\|^2 \leq \|V_{w_1} - V_{w_2}\|_D^2 = \|w_1 - w_2\|_\Sigma^2 \leq \|w_1 - w_2\|^2, \forall w_1, w_2 \in \mathbb{R}^d$.

This lemma lower bounds $\|V_{w_1} - V_{w_2}\|_D^2$ with $\omega \|w_1 - w_2\|^2$, allowing us to use strong convexity like arguments from the optimization literature. Using these lemmas, we can bound the red term as follows:

$$2\eta_t \mathbb{E}\left[g_t(w_t)^\top (w_t - w^*)\right] \leq -2\eta_t (1-\gamma)\omega \|w_t - w^*\|^2 .$$

In order to bound the blue term, we use the following lemma.

Lemma 3.3. *[Lemma 5 from Bhandari et al. (2018)] Under the i.i.d. sampling,* $\mathbb{E}\left[\|g_t(w)\|^2\right] \leq 2\sigma^2 + 8\|V_w - V_{w^*}\|_D^2$, *where* $\sigma^2 = \mathbb{E}\left[\|g_t(w^*)\|^2\right]$.

σ^2 is the variance of the TD update at the optimum. This lemma is analogous to the variance control in optimization.

Combining these bounds on the red and blue terms and setting $\eta_0 \leq \frac{1-\gamma}{8}$, we get the following bound:

$$\begin{aligned}&\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right]\\ \leq &\|w_t - w^*\|^2 \left(1 - 2\eta_0 \alpha^t (1-\gamma)\omega\right) + 2\eta_0^2 \alpha^{2t} \sigma^2.\end{aligned}$$

Taking expectation over $t \in [T]$ and using the fact $(1-x) \leq e^{-x}$, we have

$$\begin{aligned}&\mathbb{E}\left[\|w_T - w^*\|^2\right]\\ \leq &\|w_0 - w^*\|^2 \exp\left(-\eta_0 \omega (1-\gamma) \sum_{t=1}^{T} \alpha^t\right)\\ &+ 2\sigma^2 \eta_0^2 \sum_{t=1}^{T} \alpha^{2t} \exp\left(-\eta_0 \omega (1-\gamma) \sum_{i=t+1}^{T} \alpha^i\right).\end{aligned}$$

As shown in Appendix F, the exponential step-size yields $\sum_{t=1}^{T} \alpha^t \geq \frac{\alpha T}{\ln T} - \frac{1}{\ln T}$ and $\sum_{t=1}^{T} \alpha^{2t} \exp\left(-\sum_{i=t+1}^{T} \alpha^i\right) \leq O\left(\frac{(\ln(T))^2}{\alpha^2 T}\right)$. The exponential step-size achieves bias-variance trade-off without iterate averaging (Patil et al., 2023; Samsonov et al., 2024). Combining these bounds with the TD(0) recursion under i.i.d. sampling gives the final convergence rate:

Theorem 3.4. *Under Assumption 2.1 and 2.2, TD(0) under the i.i.d. sampling from stationary distribution with* $\eta_t = \eta_0 \alpha_t$, *where* $\eta_0 = \frac{1-\gamma}{8}$, $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$, $\alpha = \frac{1}{T}^{1/T}$, *has the following convergence:*

$$\begin{aligned}&\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right]\\ \leq &\|w_1 - w^*\|^2 e \exp\left(-\eta_0 \omega (1-\gamma) \frac{\alpha T}{\ln T}\right)\\ &+ \frac{8\sigma^2}{e\left(\omega(1-\gamma)\right)^2} \frac{\ln^2 T}{\alpha^2 T},\end{aligned}$$

where $\sigma^2 = \mathbb{E}\left[\|g_t(w^*)\|^2\right]$.

The proof of this theorem is in Appendix D. Compared with other methods in Table 1, TD(0) with an exponential step-size under i.i.d. sampling attains a fast convergence rate without requiring problem-dependent parameters, and it enables the optimal bias-variance trade-off. Moreover, our guarantee holds for the last iterate rather than an averaged iterate, which better reflects common practice.

4 Exponential step-size with Markovian sampling

We now relax the i.i.d. assumption and consider Markovian sampling, where the TD update uses the samples drawn sequentially from a single trajectory of the Markov chain. This setting is more realistic because it does not assume that the samples are being drawn from the stationary distribution. It is also more practical since the TD algorithm can effectively use all the samples obtained by interacting with the environment. However, since the samples are temporally correlated, the update direction is biased relative to the mean-path update. In particular, when the chain is not at stationarity then in general $\mathbb{E}\left[g_t(w_t)\right] \neq g(w_t)$. This will require controlling an additional error term.

In order to do so, we will use the property that the Markov chain is fast-mixing. This is a standard assumption in the analysis of the TD algorithm (Bhandari et al., 2018; Mitra, 2025). In particular, under Assumption 2.1, the t-step state distribution $\mu_0 P_\pi^t$ started from any μ_0 converges to the stationary distribution μ_π geometrically fast, *i.e.*,

$$\sup_{\mu_0} d_{\text{TV}}\left(P_\pi^t \mu_0, \mu_\pi\right) \leq m\rho^t, \forall t \in \mathbb{N}_0, \tag{2}$$

where d_{TV} is the total variation distance, and initial distance m and mixing speed $\rho \in (0,1)$ are positive constants that depend on the underlying Markov chain. This deviation from stationarity is quantified via the *mixing time*.

Definition 4.1. Define the mixing time as $\tau_\delta = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \delta\}$, where $\delta \in (0,1)$.

We define τ_{mix} as τ_δ for an appropriate δ to be determined later.

Using this property of fast-mixing, Bhandari et al. (2018); Mitra (2025) controlled the error term from Markvovian sampling. In particular, the analysis in Bhandari et al. (2018) requires projecting the iterates onto a bounded set containing w^*. This projection step is non-standard in practice, and requires the knowledge of ω. Mitra (2025) avoids this projection step by using an induction argument to show the iterates remain bounded. However, they can only prove convergence for the average iterate (obtained by Polyak-Rupert averaging) and require the knowledge of both τ_{mix} and ω. Unlike the analysis in Mitra (2025), we show that using exponential step-sizes allow us to prove convergence for the *last iterate* without knowledge of τ_{mix}. Moreover, to the remove dependence on ω, we use the regularized TD(0) update in Patil et al. (2023). The regularized TD(0) update at iteration t is given by

$$\begin{aligned} w_{t+1} &= w_t + \eta_t g_t^r(w), \quad \text{where,} \\ g_t^r(w) &:= \phi(s_t)\left(r(s_t) + \left(\gamma\phi(s_{t+1}) - \phi(s_t)\right)^\top w\right) - \lambda w \\ &= g_t(w) - \lambda w, \end{aligned}$$

where $\lambda > 0$ is the strength of regularization. The corresponding mean-path regularized direction is defined as

$$\begin{aligned} g^r(w) &:= \mathbb{E}\left[\phi(s_t)\left(r(s_t) + \left(\gamma\phi(s_{t+1}) - \phi(s_t)\right)^\top w\right)\right] - \lambda w \\ &= g(w) - \lambda w. \end{aligned}$$

We define w_r^* as the fixed point of the regularized TD(0) update satisfying $g^r(w_r^*) = 0$. Note that the standard TD(0) update involving $g_t(w)$ is a special case of $g_t^r(w)$ with $\lambda = 0$. We now establish the properties shared by both the regularized and standard TD(0) variants.

In particular, we first show that the following two lemmas provide upper bounds of $\|g_t^r(w)\|$ and $\|g^r(w)\|$ for the regularized and standard TD(0) variants.

Lemma 4.2. *For stochastic update g_t^r, we have*

$$\|g_t^r(w)\| \leq (2+\lambda)\|w - w_r^*\| + (3+\lambda)\zeta,$$

where $\zeta = \max\{1, \|w_r^*\|\}$.

For standard TD(0) corresponding to $\lambda = 0$, $\|g_t^r(w)\| \leq 2\|w - w_r^*\| + 3\zeta$. Since $w^* = w_r^*$ in this case, $\zeta = \max\{1, \|w^*\|\}$.

Lemma 4.3. *For mean-path update g^r, we have*

$$\|g^r(w)\| \leq (2+\lambda)\|w - w_r^*\|,$$

where $\zeta = \max\{1, \|w_r^*\|\}$. For standard TD(0) corresponding to $\lambda = 0$, $\|g^r(w)\| \leq 2\|w - w_r^*\|$.

For Markovian sampling, we show that the fast-mixing property in Eq. (2) implies the following result.

Lemma 4.4. *For any initial state distribution μ_0, state distribution as time t is $P_\pi^t \mu_0$. For any w, when $t \geq \tau_\delta$,*

$$\left\| \mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[g_t^r(w) \right] - g^r(w) \right\| \leq 2(2+\lambda)\delta \left(\|w\| + 1 \right). \tag{3}$$

For standard TD(0) corresponding to $\lambda = 0$, $\left\| \mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[g_t(w) \right] - g(w) \right\| \leq 4\delta \|w\| + 1$.

For the initial step-size η_0 in the TD(0) update, for the fixed T, we set δ and define τ_{mix} as follows:

$$\delta = \frac{\eta_0}{2(2+\lambda)T} \quad ; \quad \tau_{\text{mix}} = \tau_\delta. \tag{4}$$

By Eq. (2), when $T \geq \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}$ ensures that $T \geq \tau_{\text{mix}} := \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}$. Furthermore, using Lemma 4.4 ensures that the folowing property holds for all $t \geq \tau_{\text{mix}}$,

$$\begin{aligned} \left\| \mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[g_t^r(w) \right] - g^r(w) \right\| &\leq \eta_0 \frac{1}{T} \left(\|w\| + 1 \right) \\ &= \eta_T \left(\|w\| + 1 \right), \end{aligned} \tag{5}$$

where η_0 and η_T are the step-sizes at iterations $t = 0$ and $t = T$ in the exponential step-size schedule $\eta_t = \eta_0 \alpha^t$, respectively. The above equation shows that the deviation of expectation between the update direction at iteration t, $g_t(w)$ and the corresponding mean-path update $g(w)$ becomes small once $t \geq \tau_{\text{mix}}$.

In the next step of the proof, we bound one-step progress similar to the i.i.d. case in Section 3. In particular, we separate the Markovian component $g_t^r(w_t) - g^r(w_t)$ from the corresponding mean-path term, and use the following bound.

$$\begin{aligned} &\|w_{t+1} - w_r^*\|^2 \\ &\overset{\text{(i)}}{=} \|w_t - w_r^*\|^2 + 2\eta_t \langle g_t^r(w_t),\, w_t - w_r^* \rangle + \eta_t^2 \|g_t^r(w_t)\|^2 \\ &= \|w_t - w_r^*\|^2 + 2\eta_t \langle g_t^r(w_t) - g^r(w_t),\, w_t - w_r^* \rangle \\ &\quad + \eta_t^2 \|g_t^r(w_t)\|^2 + 2\eta_t \langle g^r(w_t),\, w_t - w_r^* \rangle \\ &\implies \mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[\|w_{t+1} - w_r^*\|^2 \right] \\ &\overset{\text{(ii)}}{\leq} \|w_t - w_r^*\|^2 + 2\eta_t \langle g^r(w_t),\, w_t - w_r^* \rangle + 2\eta_t^2 \|g^r(w_t)\|^2 \\ &\quad + 2\eta_t \mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[\langle g_t^r(w_t) - g^r(w_t),\, w_t - w_r^* \rangle \right] \\ &\quad + 2\eta_t^2 \mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[\|g_t^r(w_t) - g^r(w_t)\|^2 \right], \end{aligned} \tag{6}$$

where (i) uses the TD(0) update and (ii) uses the fact that $\|a\|^2 \leq 2\|a-b\|^2 + 2\|b\|^2$. Note that the two terms involving $g_t^r(w_t) - g^r(w_t)$ in red and blue capture the Markovian noise, while the remaining terms in green only depend on mean-path quantities. Unlike the i.i.d. setting, since g_t and w_t are correlated even after conditioning on the randomness at iteration t, $\mathbb{E}_{s_t \sim P_\pi^t \mu_0} \left[\langle g_t^r(w_t) - g^r(w_t),\, w_t - w_r^* \rangle | w_t \right] \neq 0$. Consequently, we follow the proof in Mitra (2025) and use a strong induction argument to simultaneously control the red and blue terms and show that the iterates remain bounded i.e. for a constant $B(\tau_{\text{mix}})$ that depends on the mixing time, $\|w_t - w_r^*\|^2 \leq B(\tau_{\text{mix}})$.

For this, we first note that for $t \geq \tau_{\text{mix}}$, Eq. (5) allows us to bound the average discrepancy between $g_t^r(w)$ and $g^r(w)$ in terms of $\|w\|$. Hence, in order to setup the induction we use $t = \tau_{\text{mix}}$ as the base case and first show that $\|w_t - w_r^*\|$ is bounded by $B(\tau_{\text{mix}})$ for *all* $t \leq \tau_{\text{mix}}$.

Lemma 4.5. *For the regularized TD(0) update with exponential step-sizes $\eta_t = \eta_0 \alpha_t$, where $\eta_0 \leq \frac{1-\gamma}{16 \ln(T)}$, $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$, $\alpha = \frac{1}{T}^{1/T}$, if $T \geq \max\{3, 1/\eta_0\}$,*

$$\forall t \leq \tau_{mix}, \quad \|w_t - w_r^*\|^2 \leq B(\tau_{mix}) \quad \textbf{\textit{(Base case)}},$$

where $B(\tau_{mix}) := \exp(2(2+\lambda)\max\{a,b\}) \cdot \|w_1 - w_r^* + \zeta\|^2$, where $a = \frac{1}{\ln(1/\rho)}$, $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$, $\zeta = \max\{1, \|w_r^*\|\}$.

With Lemma 4.5 giving the bound on $\|w_t - w_r^*\|^2$ for iterations $t \leq \tau_{\text{mix}}$, the base case for the induction is set up. We now state the inductive hypothesis for iteration t.

Inductive Hypothesis: For a fixed t, for all $k \leq t$, $\|w_k - w_r^*\|^2 \leq B(\tau_{\text{mix}})$.

Inductive Step: To complete the induction, we now need to show that for a fixed t, for all $k \leq t+1$, $\|w_k - w_r^*\|^2 \leq B(\tau_{\text{mix}})$. In order to do so, we use the inductive hypothesis and first prove a lemma that controls the size of the update across τ_{mix} iterations. Specifically, for $t \geq \tau_{\text{mix}}$, we bound w_t in terms of $w_{t-\tau_{\text{mix}}}$ as follows.

Lemma 4.6. *Let $T \geq \max\{3, \frac{1}{\eta_0}\}$, and T large enough such that $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, and $\ln(T) \geq \max\{a, b\}$. Suppose for all $t \geq \tau_{mix}$, if $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ for all $k \in [t]$, then,*

$$\|w_t - w_{t-\tau_{mix}}\|^2 \leq c_1^2 \, B(\tau_{mix}) \, \eta_t^2 \, \ln^4(T),$$

where $c_1^2 = 2560(2+\lambda)^2$, $a = \frac{1}{\ln(1/\rho)}$ and $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$.

Next, we decompose the first Markovian error term in red for all $t \geq \tau_{\text{mix}}$. In particular,

$$\begin{aligned}
&\mathbb{E}_t[\langle g_t^r(w_t) - g^r(w_t), w_t - w_r^*, \rangle] = T_1 + T_2 + T_3 + T_4 \\
\text{s.t } T_1 &= \mathbb{E}_t[\langle w_t - w_{t-\tau_{\text{mix}}}, g_t^r(w_t) - g^r(w_t)\rangle], \\
T_2 &= \mathbb{E}_t[\langle w_{t-\tau_{\text{mix}}} - w_r^*, g_t^r(w_{t-\tau_{\text{mix}}}) - g^r(w_{t-\tau_{\text{mix}}})\rangle], \\
T_3 &= \mathbb{E}_t[\langle w_{t-\tau_{\text{mix}}} - w_r^*, g_t^r(w_t) - g_t^r(w_{t-\tau_{\text{mix}}})\rangle], \\
T_4 &= \mathbb{E}_t[\langle w_{t-\tau_{\text{mix}}} - w_r^*, g^r(w_{t-\tau_{\text{mix}}}) - g^r(w_t)\rangle].
\end{aligned}$$

Note that terms T_1, T_3 and T_4 can be bounded deterministically by using Young's inequality. In particular, T_1 can be bounded using Lemma 4.6 and the uniform bound on $\|g_t^r(w)\|$ and $\|g^r(w)\|$ given by Lemma 4.2 and Lemma 4.3. Terms T_3 and T_4 can be bounded by using the Lipschitzness of g^r, Lemma 4.6 and using the inductive hypothesis to bound $\|w_{t-\tau_{\text{mix}}} - w_r^*\|$. For T_2, we use the bound from Eq. (5) after conditioning on $w_{t-\tau_{\text{mix}}}$. Summing these four terms yields the following lemma:

Lemma 4.7. *Let $T \geq \max\{3, \frac{1}{\eta_0}\}$, and T large enough such that $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, and $\ln(T) \geq \max\{a, b\}$. For $t \geq \tau_{mix}$, suppose $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ for all $k \in [t]$. Then:*

$$\begin{aligned}
&\mathbb{E}_t\left[\langle g_t^r(w_t) - g^r(w_t),\, w_t - w_r^*\rangle\right] \\
\leq &C \, \eta_t \, \ln^2(T) \, B(\tau_{mix}),
\end{aligned}$$

where $C = C_1 + 3 + 2C_2$, $C_1 = \frac{c_1}{2}$ and $C_2 = \frac{c_1 c_2}{2}$, $c_1 = 2560(2+\lambda)^2$ and $c_2 = 4\,(2+\lambda)^2 + 4\,(3+\lambda)^2 + 2\,(2+\lambda)^2$.

Finally, using the inductive hypothesis and the uniform bound on $\|g_t^r(w)\|$ and $\|g^r(w)\|$, we can show that the blue term in Eq. (6) can be deterministically bounded in terms of $B(\tau_{\text{mix}})$. The following lemma provides this bound.

Lemma 4.8. *Assuming $\|w_k - w_r^*\|^2 \leq B(\tau_{mix}), \forall k \in [t]$, then we have*

$$\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[\|g_t^r(w_t) - g^r(w_t)\|^2\right] \leq C' B(\tau_{mix}),$$

where $C' = 10(3+\lambda)^2$.

In order to complete the inductive step, we will use Eq. (6) and the bounds on the red and blue terms along with the mean-path analysis for the green term to show that $\|w_{t+1} - w_r^*\|^2 \leq B(\tau_{\text{mix}})$ and therefore for all $k \leq t+1$, $\|w_k - w_r^*\|^2 \leq B(\tau_{\text{mix}})$. For this last step, the analysis for the standard and regularized TD(0) updates is different, and we handle them separately.

4.1 Standard TD(0)

We use the above lemmas with $\lambda = 0$. Consequently, $w_r^* = w^*$ and $g^r(w) = g(w)$ where $g(w^*) = 0$. The following lemma shows that if the initial step-size η_0 is small enough, then we can use the inductive hypothesis to show that $\|w_{t+1} - w_r^*\|^2 \leq B(\tau_{\text{mix}})$.

Lemma 4.9. *For the standard TD(0) update, when* $T \geq \max\{3, \frac{1}{\eta_0}, \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$*,* $\frac{\ln(T)}{T} \leq \frac{1}{b}$*, and* $\ln(T) \geq \max\{a, b\}$*, for a fixed* t*, if* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t$ *and*

$$\eta_0 \leq \frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T) + C'\right]},$$

then $\|w_{t+1} - w^*\|^2 \leq B(\tau_{mix})$*, and hence,* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t+1$*.*

This completes the induction for standard TD(0), and shows that for all $t \in [T]$, $\|w_t - w^*\|^2 \leq B(\tau_{\text{mix}})$. Using Lemma 4.7, this also implies that the red term in Eq. (6) is bounded for all $t \in [T]$. Similarly, the blue term is also bounded for all $t \in [T]$ as shown in Lemma 4.8. Putting together these results, we state the complete theorem for the standard TD(0) update.

Theorem 4.10. *The standard TD(0) update with exponential step-sizes* $\eta_t = \eta_0 \alpha_t$*, where* $\eta_0 = \frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T)+C'\right]}$*,* $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$*, and* $T \geq \max\{\frac{1}{\eta_0}, \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$*,* $\frac{\ln(T)}{T} \leq \frac{1}{b}$*, and* $\ln(T) \geq \max\{a, b\}$*, achieves the following convergence rate:*

$$\begin{aligned}&\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right]\\ =&O\left(\exp\left(-\frac{\omega^2 T}{\ln^3(T)}\right) + \frac{\ln^4(T)}{\omega^2 T}\exp\left(\frac{m}{\ln(1/\rho)}\right)\right),\end{aligned}$$

where m *and* ρ *are related to mixing time as* $\tau_{mix} = \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}$*.*

The complete proof can be found in Appendix E. Comparing with other methods in Table 1, standard TD(0) with exponential step-size achieves a fast convergence rate without requiring projection onto a bounded set. In addition, our guarantee is for the last iterate. Compared with our i.i.d. sampling result in Section 3, the rate under Markovian sampling is comparable. While it requires problem-dependent parameter ω to set the initial step-size η_0. In the next subsection, we will show that regularized TD(0) update solves the ω dependence.

4.2 Regularized TD(0)

In this section, we provide analysis for regularized TD(0). The following theorem provides the step-size condition that, under the inductive hypothesis, ensures $\|w_t - w_r^*\|^2 \leq B(\tau_{\text{mix}})$ for all $t \in [T]$.

Lemma 4.11. *For the regularized TD(0) update, when* $T \geq \max\{3, \frac{1}{\eta_0}, \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$*,* $\frac{\ln(T)}{T} \leq \frac{1}{b}$*, and* $\ln(T) \geq \max\{a, b\}$*, for a fixed* t*, if* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t$*. When the step-size satisfies*

$$\eta_0 \leq \frac{\lambda}{\left[C\,\ln^2(T) + C'\right] + (8 + 2\lambda^2)},$$

then $\|w_{t+1} - w_r^*\|^2 \leq B(\tau_{mix})$*, and hence,* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t+1$*.*

This completes the induction for regularized TD(0). Comparing with the step-size requirement in Lemma 4.9, regularized TD(0) removes the requirement of ω. Similar to standard TD(0), together with Lemma 4.7 and Lemma 4.8, we have upper bounds for red and blue terms for $t \in [T]$. Having bounded all terms in the one-step expansion for all $t \in [T]$, we combine these bounds and state the final convergence rate, also accounting for the distance between w_r^* and w^*.

Theorem 4.12. *Apply reguarlized TD(0) with exponential step-size* $\eta_t = \eta_0 \alpha_t$*, where* $\eta_0 = \frac{\lambda}{\left[C\,\ln^2(T)+C'\right]+(8+2\lambda^2)}$*,* $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$*, and* $T \geq \max\{\frac{1}{\eta_0}, \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$*,* $\frac{\ln(T)}{T} \leq \frac{1}{b}$*, and* $\ln(T) \geq \max\{a, b\}$*, and* $\lambda = 1/\sqrt{T}$*, where* $a = \frac{1}{\ln(1/\rho)}$ *and*

$b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$. *Then we have the convergence rate:*

$$\begin{aligned}&\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right]\\ =&O\left(\exp\left(-\frac{\omega\sqrt{T}}{\ln^3(T)}\right) + \frac{\ln^4(T)}{\omega^2 T}\exp\left(\frac{m}{\ln(1/\rho)}\right)\right),\end{aligned}$$

where m *and* ρ *are related to mixing time as* $\tau_{mix} = \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}$.

We leave the complete proof of above lemmas and theorems to the appendix. Comparing to the methods in Table 1, our method requires no projection onto a bounded set or any prior knowledge of problem-dependent parameters. In addition, our guarantee is for the last iterate, which is often more practical than iterate averaging.

5 Conclusion

In this paper, we address a fundamental challenge in temporal difference (TD) learning: the sensitivity to step-size selection and the reliance on unknown problem parameters by using an exponential step-size $(1/T)^{t/T}$ that keeps strong theory while removing such requirements. Our main contributions are: First, our method needs no prior knowledge of problem-dependent constants under both i.i.d. and Markovian sampling and, in the Markovian case, avoids projections. Second, prove finite-time, last-iterate convergence guarantees in both settings. These results indicate a possible gain in practicality for TD learning, with some reduction in step-size tuning required.

Acknowledgments

We would like to thank Wenlong Mou, Qiushi Lin and Xingtu Liu for helpful feedback on the paper. This work was partially supported by the Canada CIFAR AI Chair Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants RGPIN-2022-03669, and enabled in part by support provided by the Digital Research Alliance of Canada (alliancecan.ca).

References

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. *Oper. Res.*, 69:950–973.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6144–6160. AAAI Press.

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

Konda, V. R. and Tsitsiklis, J. N. (1999). Actor-critic algorithms. In Solla, S. A., Leen, T. K., and Müller, K., editors, *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1008–1014. The MIT Press.

Lakshminarayanan, C. and Szepesvári, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In Storkey, A. J. and Pérez-Cruz, F., editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 1347–1355. PMLR.

Li, X., Zhuang, Z., and Orabona, F. (2021). A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6553–6564. PMLR.

Liu, R. and Olshevsky, A. (2021). Temporal difference learning as gradient splitting. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6905–6913. PMLR.

Mitra, A. (2025). A simple finite-time analysis of TD learning with linear function approximation. *IEEE Trans. Autom. Control.*, 70(2):1388–1394.

Mou, W., Li, C. J., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I. (2020). On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In Abernethy, J. D. and Agarwal, S., editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 2947–2997. PMLR.

Mustafin, A., Olshevsky, A., and Paschalidis, I. (2024). Closing the gap between SVRG and TD-SVRG with gradient splitting. *Transactions on Machine Learning Research*.

Patil, G., A., P. L., Nagaraj, D., and Precup, D. (2023). Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In Ruiz, F. J. R., Dy, J. G., and van de Meent, J., editors, *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, volume 206 of *Proceedings of Machine Learning Research*, pages 5438–5448. PMLR.

Samsonov, S., Tiapkin, D., Naumov, A., and Moulines, E. (2024). Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4511–4547. PMLR.

Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. *CoRR*, abs/1902.00923.

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning - an introduction, 2nd Edition*. MIT Press.

Tsitsiklis, J. N. and Roy, B. V. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Trans. Autom. Control.*, 42(5):674–690.

Uc-Cetina, V., Navarro-Guerrero, N., Martin-Gonzalez, A., Weber, C., and Wermter, S. (2023). Survey on reinforcement learning for language processing. *Artificial Intelligence Review*, 56(2):1543–1575.

Vaswani, S., Dubois-Taine, B., and Babanezhad, R. (2022). Towards noise-adaptive, problem-adaptive (accelerated) stochastic gradient descent. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S., editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 22015–22059. PMLR.

A Additional Inequalities Used in the Proofs

There are other inequalities regarding $g(w)$ that we used in the proofs, we also list them here for completeness.

$-g$ is 2-Lipschitz. *i.e.* $\|g(w_1) - g(w_2)\| \leq 2\|w_1 - w_2\|$.

Proof.

$$\begin{aligned}\|g(w_1) - g(w_2)\| &= \mathbb{E}_{s_t\sim\mu_\pi, s_{t+1}\sim P_\pi\mu_\pi}\left[\left\|\phi(s_t)(\phi(s_t) - \gamma\phi(s_{t+1}))^\top(w_1 - w_2)\right\|\right] \\ &\leq 2\|w_1 - w_2\|.\end{aligned} \tag{7}$$

□

$-g_t$ is 2-Lipschitz. *i.e.* $\|g_i(w_1) - g_i(w_2)\| \leq 2\|w_1 - w_2\|$.

Proof.

$$\begin{aligned}\|g_i(w_1) - g_i(w_2)\| &= \left\|\left[\phi(s_t)(\phi(s_t) - \gamma\phi(s_{t+1}))^\top\right](w_1 - w_2)\right\| \\ &\leq 2\|w_1 - w_2\|.\end{aligned} \tag{8}$$

□

Equation 7 and 8 in Mitra (2025).

$$\|g_t(w)\| \leq 2\|w - w^*\| + 4\zeta, \tag{9}$$

where $\zeta = \max\{1, \|w^*\|\}$. Since $g(w^*) = 0$, we have

$$\|g(w)\| \leq 2\|w - w^*\|. \tag{10}$$

B Constant step-size with mean-path update

Follow the vanilla TD(0) update proof in Bhandari et al. (2018),

Theorem B.1. *For the mean-path TD(0) update with a constant step-size $\eta \leq \frac{1-\gamma}{8}$, the algorithm achieves the following convergence rate:*

$$\|w_T - w^*\|^2 \leq \exp\left(-\eta(1-\gamma)\omega T\right)\left[\|w_1 - w^*\|_D^2\right].$$

Hence, to obtain accuracy $\|w_T - w^*\| \leq \epsilon$, we need $O\left(\left(\frac{1}{\omega}\right)\log\left(\frac{1}{\epsilon}\right)\right)$ gradient evaluations.

Proof. With the update $w_{t+1} = w_t + \eta g(w_t)$

$$\|w_{t+1} - w^*\|^2 = \|w_t - w^*\|^2 + 2\eta g(w_t)^\top(w_t - w^*) + \eta^2\|g(w_t)\|^2.$$

$$\begin{aligned}\|w_{t+1} - w^*\|^2 &= \|w_t - w^*\|^2 + 2\eta g(w_t)^\top(w_t - w^*) + \eta^2\|g(w_t)\|^2 \\ &\leq \|w_t - w^*\|^2 - \left(2\eta(1-\gamma) - 8\eta^2\right)\|V_{w^*} - V_{w_t}\|_D^2 \quad \text{(by Lemma 3.1 Lemma 3.3)} \\ &\leq \|w_t - w^*\|^2 - \eta(1-\gamma)\|V_{w^*} - V_{w_t}\|_D^2 \qquad (\eta \leq (1-\gamma)/8) \\ &\leq \|w_t - w^*\|^2 - \eta(1-\gamma)\omega\|w_t - w^*\|^2. \qquad \text{(by Lemma 3.2)} \\ &= (1 - \eta(1-\gamma)\omega)\|w_t - w^*\|^2\end{aligned}$$

Recurrsing over $t \in [T]$, we have

$$\|w_T - w^*\|^2 \leq (1 - \eta(1-\gamma)\omega)^T\left[\|w_1 - w^*\|^2\right].$$

Since $(1 - \eta(1-\gamma)\omega) \leq \exp\left(-\eta(1-\gamma)\omega\right)$,

$$\|w_T - w^*\|^2 \leq \exp\left(-\eta(1-\gamma)\omega T\right)\|w_1 - w^*\|^2.$$

Then for ϵ accuracy, the gradient computation is $O(\frac{1}{\omega}\log\left(\frac{1}{\epsilon}\right))$. □

C Constant step-size with i.i.d sampling

In the following sections, we assume that we have access to i.i.d. observations from the stationary distribution. Follow the vanilla TD(0) update proof in Bhandari et al. (2018),

Theorem C.1. *If the sampling data are from i.i.d observation, let the constant stepsize $\eta \leq (1-\gamma)/8$, we have the following convergence rate:*

$$\mathbb{E}\left[\|w_T - w^*\|^2\right] \leq \exp\left(-\eta(1-\gamma)\omega T\right)\|w_1 - w^*\|^2 + \eta\frac{2\sigma^2}{(1-\gamma)\omega}.$$

Then to obtain accuracy $\mathbb{E}\|\theta_t - \theta^*\| \leq \epsilon$, we need total gradient computation $O\left(\left(\frac{1}{\epsilon\omega^2}\right)\log\left(\frac{1}{\epsilon}\right)\right)$ with step-size satisfying $\eta \leq \min\left\{\frac{\omega(1-\gamma)}{8}, \frac{\epsilon(1-\gamma)\omega}{2\sigma^2}\right\}$.

Proof. With the update $w_{t+1} = w_t + \eta g_t(w_t)$

$$\|w_{t+1} - w^*\|^2 = \|w_t - w^*\|^2 + 2\eta g_t(w_t)^\top(w_t - w^*) + \eta^2\|g_t(w_t)\|^2.$$

Taking expectation over the i.i.d samples, we have

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \|w_t - w^*\|^2 + 2\eta\mathbb{E}_{s_t\sim\mu_\pi}\left[g_t(w_t)^\top(w_t - w^*)\right] + \eta^2\mathbb{E}\left[\|g_t(w_t)\|^2\right] \\
&\leq \|w_t - w^*\|^2 - \left(2\eta(1-\gamma) - 8\eta^2\right)\|V_{w^*} - V_{w_t}\|_D^2 + 2\eta^2\sigma^2 \\
&\qquad \text{(by Lemma 3.1 Lemma 3.3)} \\
&\leq \|w_t - w^*\|^2 - \eta(1-\gamma)\|V_{w^*} - V_{w_t}\|_D^2 + 2\eta^2\sigma^2 \qquad (\eta \leq (1-\gamma)/8) \\
&\leq \|w_t - w^*\|^2 - \eta(1-\gamma)\omega\|w_t - w^*\|^2 + 2\eta^2\sigma^2. \qquad \text{(by Lemma 3.2)}
\end{aligned}$$

Taking expectation over $t \in [T]$, we have

$$\mathbb{E}\left[\|w_T - w^*\|^2\right] \leq (1-\eta(1-\gamma)\omega)^T\left[\|w_1 - w^*\|^2\right] + 2\eta^2\sigma^2\sum_{t=0}^{\infty}(1-\eta(1-\gamma)\omega)^t$$

Since $(1-\eta(1-\gamma)\omega) \leq \exp\left(-\eta(1-\gamma)\omega\right)$,

$$\mathbb{E}\left[\|w_T - w^*\|^2\right] \leq \exp\left(-\eta(1-\gamma)\omega T\right)\|w_1 - w^*\|^2 + \eta\frac{2\sigma^2}{(1-\gamma)\omega}.$$

Then for ϵ accuracy, select $\eta \leq \min\left\{\frac{\omega(1-\gamma)}{8}, \frac{\epsilon(1-\gamma)\omega}{2\sigma^2}\right\}$, the gradient computation is $O\left(\frac{1}{\eta(1-\gamma)\omega}\log\left(\frac{1}{\epsilon}\right)\right) = O(\frac{1}{\epsilon\omega^2}\log\left(\frac{1}{\epsilon}\right))$. □

D Exponential step-size with i.i.d sampling proof

We now complete the proof of Theorem 3.4.

Theorem 3.4. *Under Assumption 2.1 and 2.2, TD(0) under the i.i.d. sampling from stationary distribution with $\eta_t = \eta_0\alpha_t$, where $\eta_0 = \frac{1-\gamma}{8}$, $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$, $\alpha = \frac{1}{T}^{1/T}$, has the following convergence:*

$$\begin{aligned}&\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right]\\ \leq&\|w_1 - w^*\|^2 e \exp\left(-\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln T}\right)\\ &+ \frac{8\sigma^2}{e\left(\omega(1-\gamma)\right)^2}\frac{\ln^2 T}{\alpha^2 T},\end{aligned}$$

where $\sigma^2 = \mathbb{E}\left[\|g_t(w^*)\|^2\right]$.

Proof. With the update $w_{t+1} = w_t + \eta_t g_t(w_t)$

$$\|w_{t+1} - w^*\|^2 = \|w_t - w^*\|^2 + 2\eta_t g_t(w_t)^\top (w_t - w^*) + \eta_t^2\|g_t(w_t)\|^2.$$

Taking expectation over the i.i.d samples, we have

$$\begin{aligned}\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &= \|w_t - w^*\|^2 + 2\eta_t\mathbb{E}_{s_t\sim\mu_\pi, s_{t+1}\sim P_\pi\mu_\pi}\left[g_t(w_t)^\top(w_t - w^*)\right] + \eta_t^2\mathbb{E}_{s_t\sim\mu_\pi, s_{t+1}\sim P_\pi\mu_\pi}\left[\|g_t(w_t)\|^2\right]\\ &\leq \|w_t - w^*\|^2 - \left(2\eta_t(1-\gamma) - 8\eta_t^2\right)\|V_{w^*} - V_{w_t}\|_D^2 + 2\eta_t^2\sigma^2\\ &\qquad\qquad\text{(by Lemma 3.1 Lemma 3.3)}\\ &\leq \|w_t - w^*\|^2 - \eta_t(1-\gamma)\|V_{w^*} - V_{w_t}\|_D^2 + 2\eta_t^2\sigma^2 \qquad (\eta \leq (1-\gamma)/8)\\ &\leq \|w_t - w^*\|^2 - \eta_t(1-\gamma)\omega\|w_t - w^*\|^2 + 2\eta_t^2\sigma^2. \qquad \text{(by Lemma 3.2)}\end{aligned}$$

Taking expectation over $t \in [T]$ and unrolling, we have

$$\begin{aligned}\mathbb{E}\left[\|w_T - w^*\|^2\right] &\leq \|w_0 - w^*\|^2\prod_{t=1}^T\left(1 - \eta_0\omega(1-\gamma)\alpha^t\right) + 2\sigma^2\eta_0^2\sum_{t=1}^T\alpha^{2t}\prod_{i=t+1}^T\left(1 - \eta_0\omega(1-\gamma)\alpha^t\right)\\ &\leq \|w_0 - w^*\|^2\exp\left(-\eta_0\omega(1-\gamma)\underbrace{\sum_{t=1}^T\alpha^t}_{X}\right) + 2\sigma^2\eta_0^2\underbrace{\sum_{t=1}^T\alpha^{2t}\exp\left(-\eta_0\omega(1-\gamma)\sum_{i=t+1}^T\alpha^i\right)}_{Y}.\end{aligned}$$

Applying Lemma F.1 to the first term, we have

$$\begin{aligned}&\|w_0 - w^*\|^2\exp\left(-\eta_0\omega(1-\gamma)\sum_{t=1}^T\alpha^t\right)\\ \leq&\|w_0 - w^*\|^2\exp\left(-\eta_0\omega(1-\gamma)\left(\frac{\alpha T}{\ln(T)} - \frac{1}{\ln(T)}\right)\right)\\ \leq&\|w_0 - w^*\|^2\exp\left(\eta_0\omega(1-\gamma)\frac{1}{\ln(T)}\right)\exp\left(-\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right)\\ \leq&\|w_0 - w^*\|^2 e\exp\left(-\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right)\\ &(\eta_0 \leq 1, \omega \leq 1, (1-\gamma) \leq 1, \tfrac{1}{\ln(T)} \leq 1 \text{ when } T \geq 3, \text{thus } \exp\left(\eta_0\omega(1-\gamma)\tfrac{1}{\ln(T)}\right) \leq e)\end{aligned}$$

And applying Lemma F.2 to the second term, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha^{2t} \exp\left(-\eta_0\omega(1-\gamma)\sum_{i=t+1}^{T}\alpha^i\right) \\
\leq &\frac{4\exp\left(\eta_0\omega(1-\gamma)\frac{1}{\ln(T)}\right)}{e^2\left(\eta_0\omega(1-\gamma)\right)^2}\frac{\ln^2(T)}{\alpha^2 T} \\
\leq &\frac{4e}{e^2\left(\eta_0\omega(1-\gamma)\right)^2}\frac{\ln^2(T)}{\alpha^2 T} && \left(\exp\left(\eta_0\omega(1-\gamma)\tfrac{1}{\ln(T)}\right)\leq e\right) \\
= &\frac{4}{e\left(\eta_0\omega(1-\gamma)\right)^2}\frac{\ln^2(T)}{\alpha^2 T}
\end{aligned}$$

Putting two terms together, we have the convergence result:

$$\mathbb{E}\left[\|w_T - w^*\|^2\right] \leq \|w_0 - w^*\|^2 e\exp\left(-\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right) + \frac{8\sigma^2}{e\left(\omega(1-\gamma)\right)^2}\frac{\ln^2(T)}{\alpha^2 T}.$$

□

E Exponential step-size with Markovian sampling

Here we analyze standard TD(0) and its regularized variant with an exponential step-size under Markovian sampling. We include standard TD(0) for reference and focus on regularized TD(0) because it requires no problem-dependent parameters. The regularized update at iteration t is

$$\begin{aligned}
g_t^r(w) &= \phi(s_t)\left(r(s_t) + \left(\gamma\phi(s_{t+1}) - \phi(s_t)\right)^\top w\right) - \lambda w \\
&= g_t(w) - \lambda w.
\end{aligned}$$

The corresponding mean-path regularized update is $g^r(w) = \mathbb{E}_{s_t\sim\mu_\pi, s_{t+1}\sim P_\pi\mu_\pi}\left[\phi(s_t)\left(r(s_t) + \left(\gamma\phi(s_{t+1}) - \phi(s_t)\right)^\top w\right)\right] - \lambda w = g(w) - \lambda w$, where P_π is the transition matrix induced by π, and μ_π is the stationary distribution. We define w_r^* as the fixed point of regularized TD(0) update satisfying $g^r(w_r^*) = 0$.

It is clear that the standard TD(0) update involving $g_t(w)$ is a special case of $g_t^r(w)$ with $\lambda = 0$. We first state the properties shared by both methods that will be used in the proofs, and then highlight the points where their analyses differ.

Lemma E.1. *[Lemma 3 from Bhandari et al. (2018) with regularized update] For regularized TD(0) with mean-path sampling, the following inequality holds:*

$$\langle g^r(w), w_r^* - w\rangle \geq \left[(1-\gamma)\omega + \lambda\right]\|w - w_r^*\|^2.$$

For standard TD(0) corresponding to $\lambda = 0$, $\langle g^r(w), w_r^* - w\rangle \geq (1-\gamma)\omega\|w - w_r^*\|^2$.

Proof. Define $\xi_r = (w_r^* - w)^\top\phi(s)$ and $\xi_r' = (w_r^* - w)^\top\phi(s_{t+1})$. Since both s and s' are sampled from the stationary distribution, ξ_r and ξ_r' have the same marginal distribution. Using the expression for g^r,

$$\begin{aligned}
g^r(w) &= g^r(w) - g^r(w_r^*) && \text{(since } g^r(w_r^*) = 0\text{)} \\
&= g(w) - g(w_r^*) - \lambda(w - w_r^*) && \text{(by definition of } g^r\text{)} \\
&= \mathbb{E}_{s_t\sim\mu_\pi}\left[\phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t))(w - w_r^*)\right] - \lambda(w - w_r^*) && \text{(by definition of } g\text{)} \\
&= \mathbb{E}_{s_t\sim\mu_\pi}\left[\phi(s)(\xi_r - \gamma\xi_r')\right] - \lambda(w - w_r^*).
\end{aligned}$$

Therefore

$$\begin{aligned}
\langle g^r(w), w_r^* - w\rangle &= \mathbb{E}_{s_t\sim\mu_\pi, s_{t+1}\sim P_\pi\mu_\pi}\left[\xi_r(\xi_r - \gamma\xi_r')\right] + \lambda\left\|w - w_r^*\right\|^2 \\
&\qquad\qquad \text{(since } \xi_r = \langle\phi(s), w_r^* - w\rangle\text{)} \\
&= \mathbb{E}_{s_t\sim\mu_\pi}\left[\xi_r^2\right] - \gamma\mathbb{E}_{s_t\sim\mu_\pi, s_{t+1}\sim P_\pi\mu_\pi}\left[\xi_r\xi_r'\right] + \lambda\left\|w - w_r^*\right\|^2 \\
&\geq \mathbb{E}_{s_t\sim\mu_\pi}\left[\xi_r^2\right] - \gamma\sqrt{\mathbb{E}_{s_t\sim\mu_\pi}\left[\xi_r^2\right]}\sqrt{\mathbb{E}_{s_{t+1}\sim\mu_\pi}\left[(\xi_r')^2\right]} + \lambda\left\|w - w_r^*\right\|^2 \\
&\qquad\qquad \text{(by Cauchy-Schwarz)} \\
&= \mathbb{E}_{s_t\sim\mu_\pi}\left[\xi_r^2\right] - \gamma\mathbb{E}_{s_t\sim\mu_\pi}\left[\xi_r^2\right] + \lambda\left\|w - w_r^*\right\|^2 \\
&\qquad\qquad \text{(since } \xi_r \text{ and } \xi_r' \text{ have the same marginal distribution)} \\
&= (1-\gamma)(w_r^* - w)^\top\mathbb{E}_{s_t\sim\mu_\pi}\left[\phi(s_t)\phi(s_t)^\top\right](w_r^* - w) + \lambda\left\|w - w_r^*\right\|^2 \\
&\qquad\qquad \text{(by the definition of } \xi_r\text{)} \\
&\geq \left((1-\gamma)\omega + \lambda\right)\left\|w - w_r^*\right\|^2. \qquad \text{(since } \mathbb{E}_{s_t\sim\mu_\pi}\left[\phi(s_t)\phi(s_t)^\top\right] \succeq \omega I_d\text{)}
\end{aligned}$$

□

Lemma E.2. *[Lemma 4 from Bhandari et al. (2018) with regularized update] For regularized TD(0) with mean-path sampling,*

$$\left\|g^r(w)\right\|^2 \leq (8 + 2\lambda^2)\left\|w - w_r^*\right\|^2.$$

For standard TD(0) corresponding to $\lambda = 0$, $\left\|g_r(w)\right\|^2 \leq 8\left\|w - w_r^*\right\|^2$.

Proof. Similar to the proof of Lemma E.1, define $\xi_r = (w_r^* - w)^\top\phi(s_t)$ and $\xi_r' = (w_r^* - w)^\top\phi(s_{t+1})$, and note that ξ_r and ξ_r' have the same marginal distribution.

$$\begin{aligned}
&\left\|g^r(w)\right\|^2 \\
=&\left\|g^r(w) - g^r(w_r^*)\right\|^2 && \text{(since } g^r(w_r^*) = 0\text{)} \\
=&\left\|g(w) - g(w_r^*) - \lambda(w - w_r^*)\right\|^2 && \text{(by the definition of } g^r\text{)} \\
\leq&2\left\|g(w) - g(w_r^*)\right\|^2 + 2\lambda^2\left\|w - w_r^*\right\|^2 && \text{(since } (a+b)^2 \leq 2a^2 + 2b^2\text{)} \\
=&2\left\|\mathbb{E}_{s_t\sim\mu_\pi}\left[\phi(s_t)(\xi_r - \gamma\xi_r')\right]\right\|^2 + 2\lambda^2\left\|w - w_r^*\right\|^2 && \text{(by the definition of } g\text{)} \\
\leq&2\left(\sqrt{\mathbb{E}_{s_t\sim\mu_\pi}\left[\left\|\phi(s_t)\right\|^2\right]}\sqrt{\mathbb{E}_{s_t\sim\mu_\pi}\left[(\xi_r - \gamma\xi_r')^2\right]}\right)^2 + 2\lambda^2\left\|w - w_r^*\right\|^2 && \text{(by Cauchy-Schwarz)} \\
=&2\mathbb{E}_{s_t\sim\mu_\pi}\left[\left\|\phi(s_t)\right\|^2\right]\mathbb{E}_{s_t\sim\mu_\pi}\left[(\xi_r - \gamma\xi_r')^2\right] + 2\lambda^2\left\|w - w_r^*\right\|^2 \\
\leq&2\big(2\mathbb{E}_{s_t\sim\mu_\pi}\left[\xi_r^2\right] + 2\gamma^2\mathbb{E}_{s_t\sim\mu_\pi}\left[(\xi_r')^2\right]\big) + 2\lambda^2\left\|w - w_r^*\right\|^2 \\
&\qquad\qquad \text{(since } \left\|\phi(s_t)\right\|^2 \leq 1 \text{ and } (a+b)^2 \leq 2a^2 + 2b^2\text{)} \\
=&2\big(2(1+\gamma^2)(w_r^* - w)^\top\mathbb{E}_{s_t\sim\mu_\pi}\left[\phi(s_t)\phi(s_t)^\top\right](w_r^* - w)\big) + 2\lambda^2\left\|w - w_r^*\right\|^2 \\
&\qquad\qquad \text{(since } \xi_r \text{ and } \xi_r' \text{ have the same marginal distribution)} \\
\leq&(8 + 2\lambda^2)\left\|w - w_r^*\right\|^2. && \text{(since } \left\|\phi\right\|^2 \leq 1 \text{ and } \gamma \leq 1\text{)}
\end{aligned}$$

□

Lemma E.3. *The distance between the fixed points of regularized and standard TD(0) is bounded by:*

$$\left\|w^* - w_r^*\right\| \leq \frac{\lambda\left\|w^*\right\|}{\lambda + \omega(1-\gamma)}.$$

For standard TD(0) corresponding to $\lambda = 0$, $\left\|w^* - w_r^*\right\| = 0$.

Proof. By Lemma 3.1,

$$\begin{aligned}
(w^* - w)^\top g(w) &\geq (1-\gamma)\left\|V_w - V_{w^*}\right\|_D^2 \\
&\geq (1-\gamma)\omega\left\|w - w^*\right\|^2 \\
&\qquad \text{(since } \left\|V_w - V_{w^*}\right\|_D^2 = \left\|\Phi^\top(w - w^*)\right\|_D^2 \geq \omega\left\|w - w^*\right\|^2\text{)}
\end{aligned}$$

Substituting $w = w_r^*$, we obtain

$$(w^* - w_r^*)^\top g(w_r^*) \geq (1-\gamma)\omega \left\| w_r^* - w^* \right\|^2.$$

By the definition of w_r^*, we have $g_r(w_r^*) = g(w_r^*) - \lambda w_r^* = 0$, thus $g(w_r^*) = \lambda w_r^*$. Using this relation with the above inequality,

$$\begin{aligned}
& \lambda(w^* - w_r^*)^\top w_r^* \geq (1-\gamma)\omega \left\| w_r^* - w^* \right\|^2 \\
\implies \quad & \lambda(w^* - w_r^*)^\top (w_r^* - w^*) + \lambda(w^* - w_r^*)^\top w^* \geq (1-\gamma)\omega \left\| w_r^* - w^* \right\|^2 \\
& \qquad\qquad \text{(adding/subtracting)} \\
\implies \quad & -\lambda \left\| w^* - w_r^* \right\|^2 + \lambda(w^* - w_r^*)^\top w^* \geq (1-\gamma)\omega \left\| w_r^* - w^* \right\|^2 \\
\implies \quad & [\lambda + \omega(1-\gamma)] \left\| w^* - w_r^* \right\|^2 \leq \lambda(w^* - w_r^*)^\top w^* \leq \lambda \left\| w^* - w_r^* \right\| \left\| w^* \right\| \\
& \qquad\qquad \text{(by Cauchy-Schwarz)} \\
\implies \quad & \left\| w^* - w_r^* \right\| \leq \frac{\lambda \left\| w^* \right\|}{\lambda + \omega(1-\gamma)}.
\end{aligned}$$

□

The following two lemmas are analogous to Equation 7 and 8 in Mitra (2025).

Lemma 4.2. *For stochastic update g_t^r, we have*

$$\left\| g_t^r(w) \right\| \leq (2+\lambda) \left\| w - w_r^* \right\| + (3+\lambda)\zeta,$$

where $\zeta = \max\{1, \|w_r^*\|\}$.

For standard TD(0) corresponding to $\lambda = 0$, $\|g_t^r(w)\| \leq 2\|w - w_r^*\| + 3\zeta$. Since $w^* = w_r^*$ in this case, $\zeta = \max\{1, \|w^*\|\}$.

Proof.

$$\begin{aligned}
\left\| g_t^r(w) \right\| &= \left\| g_t^r(w) - g_t^r(w_r^*) + g_t^r(w_r^*) \right\| && \text{(add/subtract)} \\
&= \left\| g_t(w) - g_t(w_r^*) - \lambda(w - w_r^*) + g_t^r(w_r^*) \right\| \\
&\leq \left\| g_t(w) - g_t(w_r^*) - \lambda(w - w_r^*) \right\| + \left\| g_t^r(w_r^*) \right\| && \text{(triangle inequality)} \\
&= \left\| \left(\phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t))^\top - \lambda\right)(w - w_r^*) \right\| + \left\| \phi(s_t) \left(r(s_t) + (\gamma\phi(s_{t+1}) - \phi(s_t))^\top w_r^*\right) - \lambda w_r^* \right\| \\
&\leq \left\| \left(\phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t))^\top\right)(w - w_r^*) \right\| + \lambda \left\| w - w_r^* \right\| \\
&\quad + \left\| \phi(s_t) r(s_t) \right\| + \left\| \phi(s_t) \left((\gamma\phi(s_{t+1}) - \phi(s_t))^\top w_r^*\right) \right\| + \lambda \left\| w_r^* \right\| \\
& && (\|a+b\| \leq \|a\| + \|b\|) \\
&\leq \left\| \left(\phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t))^\top\right) \right\| \left\| w - w_r^* \right\| + \lambda \left\| w - w_r^* \right\| \\
&\quad + \left\| \phi(s_t) r(s_t) \right\| + \left\| \phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t)) \right\| \left\| w_r^* \right\| + \lambda \left\| w_r^* \right\| && \text{(by Cauchy-Schwarz)} \\
&\leq (2+\lambda) \left\| w - w_r^* \right\| + (3+\lambda)\zeta, && (r(s_t) \leq 1, \|\phi(s_t)\| \leq 1, \text{Eq. (8)})
\end{aligned}$$

where $\zeta = \max\{1, \|w_r^*\|\}$. □

The corresponding bound on the mean-path update is

Lemma 4.3. *For mean-path update g^r, we have*

$$\left\| g^r(w) \right\| \leq (2+\lambda) \left\| w - w_r^* \right\|,$$

where $\zeta = \max\{1, \|w_r^*\|\}$. For standard TD(0) corresponding to $\lambda = 0$, $\|g^r(w)\| \leq 2\|w - w_r^*\|$.

Proof.

$$\begin{aligned}
\left\| g^r(w) \right\| &= \left\| g^r(w) - g^r(w_r^*) \right\| && (g^r(w_r^*) = 0) \\
&= \left\| (g(w) - g(w_r^*)) - \lambda(w - w_r^*) \right\| && \text{(by definition of } g^r(w)) \\
&\leq \left\| g(w) - g(w_r^*) \right\| + \lambda \left\| w - w_r^* \right\| && \text{(triangle inequality)} \\
&\leq (2+\lambda) \left\| w - w_r^* \right\|. && \text{(by Eq. (7))}
\end{aligned}$$

□

In Markovian case, we also need the definition of mixing time. We restate below and provide proof.

Definition 4.1. Define the mixing time as $\tau_\delta = \min\{t \in \mathbb{N}_0 \mid m\rho^t \le \delta\}$, where $\delta \in (0,1)$.

Lemma 4.4. *For any initial state distribution μ_0, state distribution as time t is $P_\pi^t \mu_0$. For any w, when $t \ge \tau_\delta$,*

$$\left\|\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[g_t^r(w)\right] - g^r(w)\right\| \le 2(2+\lambda)\delta\left(\|w\|+1\right). \tag{3}$$

For standard TD(0) corresponding to $\lambda = 0$, $\left\|\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[g_t(w)\right] - g(w)\right\| \le 4\delta\|w\| + 1$.

Proof. Let $\|g_t^r(w)\|_\infty$ denote the supremum of $\|g_t^r(w)\|$ over states.

$$\begin{aligned}
\|g_t^r(w)\|_\infty &= \max_{s_t \in \mathcal{S}} \|g_t^r(w)\| \\
&= \max_{s_t \in \mathcal{S}} \left\|\big(r(s_t) + \gamma w^\top \phi(s_t') - w^\top \phi(s_t)\big)\phi(s_t) - \lambda w\right\| && \text{(Definition of } g_t^r(w)\text{)} \\
&\le (2+\lambda)\|w\| + 1. && \text{(since } r(s_t) \le 1, \|\phi(s_t)\| \le 1, \gamma \le 1\text{)}
\end{aligned}$$

With $\tau_\delta = \min\{t \in \mathbb{N}_0 \mid m\rho^t \le \delta\}$, and $P_\pi^t \mu_0$ representing the probability distribution over the states after t steps with initial state distribtion μ_0, we have

$$\left\|\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[g_t^r(w)\right] - g^r(w)\right\| = \left\|\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[g_t^r(w)\right] - \mathbb{E}_{s_t \sim \mu_\pi}\left[g_t^r(w)\right]\right\|$$

(Definition of $g_t^r(w)$ and $g^r(w)$)

$$\begin{aligned}
&= \left\|\sum_{s_t \in \mathcal{S}} g_t^r(w)\big(\big(P_\pi^t \mu_0\big)(s_t) - \mu_\pi(s_t)\big)\right\| \\
&\le \sum_{s_t \in \mathcal{S}} \left\|g_t^r(w)\big(\big(P_\pi^t \mu_0\big)(s_t) - \mu_\pi(s_t)\big)\right\| && \text{(triangle inequality)} \\
&\le \|g_t^r(w)\|_\infty \sum_{s_t \in \mathcal{S}} \left|\big(P_\pi^t \mu_0\big)(s_t) - \mu_\pi(s_t)\right| && (\|g_t^r(w)\| \le \|g_t^r(w)\|_\infty) \\
&\le 2\|g_t^r(w)\|_\infty \sup_{\mu_0} d_{\text{TV}}\big(P_\pi^t \mu_0, \mu_\pi\big)
\end{aligned}$$

(by the definition of the total variation distance, and taking the sup over s_0)

$$\begin{aligned}
&\le 2\|g_t^r(w)\|_\infty m\rho^t && \text{(using Eq. (2))} \\
&\le 2m\rho^t((2+\lambda)\|w\| + 1) && \text{(using the bound on } \|g_t^r(w)\|_\infty\text{)} \\
&\le 2(2+\lambda)\delta(\|w\|+1) && \text{(using the definition of } t_\delta\text{)}
\end{aligned}$$

□

Lemma E.4. *For τ_{mix} defined in Eq. (4),*

$$\tau_{mix} = a\ln(T') + b,$$

where a and b are constants with $a = \frac{1}{\ln(1/\rho)}$, $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$ and $T' = \frac{T}{\eta_0}$.

Proof. By Eq. (2), Lemma 4.4 and the definition of τ_{mix} in Eq. (4), we have the expression of τ_{mix} as:

$$\begin{aligned}
\tau_{\text{mix}} &= \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)} \\
&= \frac{\ln(T')}{\ln(1/\rho)} + \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)} && \text{(defining } T' = \tfrac{T}{\eta_0}\text{)}
\end{aligned}$$

□

Equipped with this expression of τ_{mix}, we prove the following lemmas.

Lemma E.5. *If $\alpha := \frac{1}{T}^{1/T}$, $T \ge 3$, $\eta_0 \le 1$ for τ_{mix} defined in Eq. (4), then for all $t \le \tau_{mix}$*

$$\frac{1-\alpha^t}{1-\alpha} \le 4\max\{a,b\}\ln(T'),$$

where $a = \frac{1}{\ln(1/\rho)}$, $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$ and $T' = \frac{T}{\eta_0}$.

Proof. Using the expression of $\tau_{\text{mix}} = a\ln(T') + b$ from Lemma E.4,

$$\frac{1-\alpha^t}{1-\alpha} \le \frac{1-\alpha^{\tau_{\text{mix}}}}{1-\alpha} = \frac{1-(1/T)^{\tau_{\text{mix}}/T}}{1-(1/T)^{1/T}} = \frac{1-(1/T)^{(a\ln(T')+b)/T}}{1-(1/T)^{1/T}} \qquad \text{(since } t \le \tau_{\text{mix}}\text{)}$$
$$= \frac{1-\exp\left(\frac{-a(\ln(T')\ln(T))}{T} - \frac{b\ln(T)}{T}\right)}{1-\exp\left(-\frac{\ln(T)}{T}\right)}.$$

Define $j := \frac{a\ln(T)\ln(T')+b\ln(T)}{T}$, and $k := \frac{\ln(T)}{T}$, notice that j and k have the relationship $j = a\ln(T')k + bk$. We can simplify the above inequality as follows:

$$\frac{1-\alpha^{\tau_{\text{mix}}}}{1-\alpha} = \frac{1-\exp(-j)}{1-\exp(-k)}$$

To bound the numerator and denominator separately, we use the fact that $\frac{v}{1+v} \le 1-\exp(-v) \le v$ for $v > 0$. For the numerator, setting $v = j$, we have $1-\exp(-j) \le j$. And for the denominator, setting $v = k$, we have $1-\exp(-k) \ge k/1+k$. Combining the above relations,

$$\implies \frac{1-\alpha^t}{1-\alpha} \le \frac{j(1+k)}{k}$$
$$= \frac{(a\ln(T')k + bk)(1+k)}{k} \qquad \text{(since } j = a\ln(T')k + bk\text{)}$$
$$= (a\ln(T') + b)\left(1 + \frac{\ln(T)}{T}\right) \qquad \text{(since } k = \ln(T)/T\text{)}$$
$$\le (a\ln(T') + b)(1 + 1/e) \qquad \text{(}\tfrac{\ln(T)}{T}\text{ decreases after } T = e\text{, assuming } T \ge 3\text{)}$$
$$\le 4\max\{a, b\}\ln(T'). \qquad \text{(since } \eta_0 \le 1 \text{ and } T \ge 1 \implies T' \ge 1\text{)}$$

□

Lemma E.6. *If* $\alpha := \frac{1}{T}^{1/T}$, $T \ge \max\left\{3, \frac{1}{\eta_0}\right\}$ *and* T *large enough such that* $\frac{\ln^2(T)}{T} \le \frac{1}{2a}$ *and* $\frac{\ln(T)}{T} \le \frac{1}{b}$, $\eta_0 \le 1$, *for* τ_{mix} *defined in Eq.* (4) *and for all* $t \le \tau_{mix}$,

$$\frac{\alpha^{-\tau_{mix}} - 1}{1-\alpha} \le 8\ \max\{a, b\}\ \ln(T')$$

where $a = \frac{1}{\ln(1/\rho)}$, $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$ *and* $T' = \frac{T}{\eta_0}$.

Proof. First note that,

$$\alpha = \left(\frac{1}{T}\right)^{1/T} = \exp\left(-\frac{1}{T}\ \ln(T)\right)$$

Using the expression of $\tau_{\text{mix}} = a\ln(T') + b$ from Lemma E.4,

$$\alpha^{\tau_{\text{mix}}} = \exp\left(-\frac{\tau_{\text{mix}}}{T}\ \ln(T)\right) = \exp\left(-\frac{a\ln(T') + b}{T}\ \ln(T)\right)$$

Define $k := \frac{\ln(T)}{T}$ and $j := [a\ln(T') + b]\ \frac{\ln(T)}{T} = [a\ln(T') + b]\ k$. Using the above relations,

$$\frac{\alpha^{-\tau_{\text{mix}}} - 1}{1-\alpha} = \frac{\exp(j) - 1}{1-\exp(-k)}$$

In order to simplify the above expression, we use the following inequalities:

$$\forall x \ge 0\,, 1-\exp(-x) \ge \frac{x}{1+x} \quad ; \quad \forall y \in (0,1)\,, \exp(2y) \le \frac{1+y}{1-y}$$

Since $T \ge 1$, $k \ge 0$ and hence we can use $x = k$ to conclude that $1-\exp(-k) \ge \frac{k}{1+k}$.
We substitute y with $j/2$, which requires $0 < j < 2$. $j = [a\ln(T') + b]\ \frac{\ln(T)}{T} = \tau_{\text{mix}}\frac{\ln(T)}{T} > 0$ is

already satisfied. Ensuring $j \leq 2$ requires that $[a \ln(T') + b] \frac{\ln(T)}{T} \leq 2$. For $\eta_0 \leq 1$ and $T \geq 1$, $T' \geq 1$. Moreover, for $T \geq \frac{1}{\eta_0}$, $\frac{\ln(T')}{\ln(T)} \leq 2$. Hence, it suffices to ensure that,

$$2\,a\,\frac{\ln^2(T)}{T} + b\,\frac{\ln(T)}{T} \leq 2$$

Hence, it suffices to ensures that,

$$\frac{\ln^2(T)}{T} \leq \frac{1}{2a} \quad \text{and} \quad \frac{\ln(T)}{T} \leq \frac{1}{b}$$

With these constraints on T, we can guarantee that $j \leq 1$. Using $y = \frac{j}{2}$ in the above inequality, we can conclude that,

$$\exp(j) - 1 \leq \frac{1 + j/2}{1 - j/2} - 1 = \frac{j}{1 - j/2} \leq 2\,j$$

Combining the above relations, we get that,

$$\frac{\alpha^{-\tau_{\text{mix}}} - 1}{1 - \alpha} \leq \frac{2\,j\,(k+1)}{k}$$

Following the same steps as in the proof of Lemma E.5, we conclude that, for $T \geq 3$ and $\eta_0 \leq 1$,

$$\frac{\alpha^{-\tau_{\text{mix}}} - 1}{1 - \alpha} \leq 8\,\max\{a, b\}\,\ln(T')$$

□

Lemma 4.5. *For the regularized TD(0) update with exponential step-sizes $\eta_t = \eta_0 \alpha_t$, where $\eta_0 \leq \frac{1-\gamma}{16 \ln(T)}$, $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$, $\alpha = \frac{1}{T}^{1/T}$, if $T \geq \max\{3, 1/\eta_0\}$,*

$$\forall t \leq \tau_{mix}, \quad \|w_t - w_r^*\|^2 \leq B(\tau_{mix}) \quad \textbf{\textit{(Base case)}}\,,$$

where $B(\tau_{mix}) := \exp(2(2+\lambda)\max\{a,b\}) \cdot \|w_1 - w_r^* + \zeta\|^2$, where $a = \frac{1}{\ln(1/\rho)}$, $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$, $\zeta = \max\{1, \|w_r^*\|\}$.

Proof.

$$\begin{aligned}\|w_{t+1} - w_r^*\| &\leq \|w_t - w_r^*\| + \eta_t \|g_t^r(w_t)\| \\ &\leq (1 + (2+\lambda)\eta_t)\,\|w_t - w_r^*\| + (3+\lambda)\eta_t\zeta. \qquad \text{(by Lemma 4.2)}\end{aligned}$$

Iterating the above inequality, we have $\forall t \leq \tau_{\text{mix}}$:

$$\begin{aligned}
\|w_t - w_r^*\| &\leq \|w_1 - w_r^*\| \prod_{i=1}^{t}(1 + (2+\lambda)\eta_i) + (3+\lambda)\zeta \sum_{i=1}^{t} \eta_i \prod_{j=i+1}^{t} (1 + (2+\lambda)\eta_j) \\
&\leq \|w_1 - w_r^*\| \exp\left((2+\lambda)\sum_{i=1}^{t}\eta_i\right) + (3+\lambda)\zeta \sum_{i=1}^{t} \eta_i \exp\left((2+\lambda)\sum_{j=i+1}^{t}\eta_j\right) \\
&\qquad\qquad \text{(since } 1 + x \leq \exp(x)\text{)} \\
&= \|w_1 - w_r^*\| \exp\left((2+\lambda)\eta_0\sum_{i=1}^{t}\alpha^i\right) + (3+\lambda)\zeta\eta_0 \sum_{i=1}^{t} \alpha^i \exp\left((2+\lambda)\eta_0\sum_{j=i+1}^{t}\alpha^j\right) \\
&= \|w_1 - w_r^*\| \exp\left((2+\lambda)\eta_0\sum_{i=1}^{t}\alpha^i\right) + (3+\lambda)\zeta\eta_0 \sum_{i=1}^{t} \alpha^i \exp\left((2+\lambda)\eta_0\frac{\alpha^{i+1} - \alpha^{t+1}}{1-\alpha}\right) \\
&= \|w_1 - w_r^*\| \exp\left((2+\lambda)\eta_0\frac{\alpha - \alpha^{t+1}}{1-\alpha}\right) + (3+\lambda)\zeta\eta_0 \sum_{i=1}^{t} \alpha^i \exp\left((2+\lambda)\eta_0\frac{\alpha^{i+1} - \alpha^{t+1}}{1-\alpha}\right) \\
&\leq \|w_1 - w_r^*\| \exp\left((2+\lambda)\eta_0\frac{\alpha - \alpha^{t+1}}{1-\alpha}\right) + (3+\lambda)\zeta\eta_0 \sum_{i=1}^{t} \alpha^i \exp\left((2+\lambda)\eta_0\frac{\alpha - \alpha^{t+1}}{1-\alpha}\right) \\
&\qquad\qquad \text{(since } \alpha^{i+1} \leq \alpha\text{)} \\
&= \exp\left((2+\lambda)\eta_0\frac{\alpha - \alpha^{t+1}}{1-\alpha}\right)\left[\|w_1 - w_r^*\| + (3+\lambda)\zeta\eta_0\sum_{i=1}^{t}\alpha^i\right] \\
&= \exp\left((2+\lambda)\eta_0\alpha\frac{1 - \alpha^{t}}{1-\alpha}\right)\left[\|w_1 - w_r^*\| + (3+\lambda)\zeta\eta_0\alpha\frac{1-\alpha^t}{1-\alpha}\right] \\
&\leq \exp(4(2+\lambda)\eta_0\alpha\max\{a,b\}\ln(T'))\left[\|w_1 - w_r^*\| + 4(3+\lambda)\zeta\eta_0\alpha\max\{a,b\}\ln(T')\right] \\
&\qquad\qquad \text{(by Lemma E.5)}
\end{aligned}$$

since $\eta_0 \leq \frac{1-\gamma}{16\ln(T)}$,

$$\begin{aligned}
&\leq \exp\left(\frac{1}{4}(2+\lambda)\alpha(1-\gamma)\max\{a,b\}\frac{\ln(T')}{\ln(T)}\right)\left[\|w_1 - w_r^*\| + \frac{1}{4}(3+\lambda)\alpha(1-\gamma)\max\{a,b\}\frac{\ln(T')}{\ln(T)}\zeta\right] \\
&\leq \exp\left(\frac{1}{2}(2+\lambda)\alpha(1-\gamma)\max\{a,b\}\right)\left[\|w_1 - w_r^*\| + \frac{1}{2}(3+\lambda)\alpha(1-\gamma)\max\{a,b\}\zeta\right] \\
&\qquad\qquad \text{(for } T \geq \tfrac{1}{\eta_0},\ \tfrac{\ln(T')}{\ln(T)} \leq 2\text{)} \\
&\leq \exp\left(\frac{1}{2}(2+\lambda)\max\{a,b\}\right)\exp\left(\frac{1}{4}(3+\lambda)\max\{a,b\}\right)\left[\|w_1 - w_r^*\| + \zeta\right] \\
&\qquad\qquad (\alpha \leq 1, (1-\gamma) \leq 1, \max\{a,b\} \leq \exp(\tfrac{1}{2}\max\{a,b\})) \\
&\leq \exp((2+\lambda)\max\{a,b\})\left[\|w_1 - w_r^*\| + \zeta\right].
\end{aligned}$$

Squaring the both sides, we get:

$$\|w_t - w_r^*\|^2 \leq \exp(2\,(2+\lambda)\max\{a,b\})\left[\|w_1 - w_r^*\| + \zeta\right]^2.$$

Let $B(\tau_{\text{mix}}) := \exp(2\,(2+\lambda)\max\{a,b\})\left[\|w_1 - w_r^*\| + \zeta\right]^2$, we have that $\|w_t - w_r^*\|^2 \leq B(\tau_{\text{mix}})$ for all $t \leq \tau_{\text{mix}}$. □

Using the above lemmas, we will follow a proof similar to that in Mitra (2025). For this, we define the following notation:

$$d_t := \mathbb{E}\left[\|w_t - w_r^*\|\right]^2,$$

and

$$e_t := \mathbb{E}\left[\langle w_t - w_r^*, g_t^r(w_t) - g^r(w_t)\rangle\right],$$

which includes the error introduced by sampling along the Markov chain. The expectations are taken with respect to state distribution at time t. For the subsequent lemmas, we omit the subscript $P_\pi^t \mu_0$ for brevity.

Lemma 4.6. *Let $T \geq \max\{3, \frac{1}{\eta_0}\}$, and T large enough such that $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, and $\ln(T) \geq \max\{a, b\}$. Suppose for all $t \geq \tau_{mix}$, if $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ for all $k \in [t]$, then,*

$$\|w_t - w_{t-\tau_{mix}}\|^2 \leq c_1^2 \, B(\tau_{mix}) \, \eta_t^2 \, \ln^4(T),$$

where $c_1^2 = 2560(2+\lambda)^2$, $a = \frac{1}{\ln(1/\rho)}$ and $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$.

Proof.

$$\begin{aligned}
\|w_t - w_{t-\tau_{\text{mix}}}\| &\leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \|w_{i+1} - w_i\| && \text{(triangle inequality)} \\
&\leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \eta_i \, \|g_i^r(w_i)\| && \text{(by the update)} \\
&\leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \eta_i((2+\lambda)\,\|w_i - w_r^*\| + (3+\lambda)\zeta) && \text{(by Lemma 4.2)} \\
&\leq \sum_{i=t-\tau_{\text{mix}}}^{t-1} \eta_i\Big((2+\lambda)\sqrt{B(\tau_{\text{mix}})} + (3+\lambda)\zeta\Big) \\
&&& \text{(assuming that } \|w_k - w_r^*\|^2 \leq B(\tau_{\text{mix}}) \text{ for } k \in [t]) \\
&= \underbrace{\Big((2+\lambda)\sqrt{B(\tau_{\text{mix}})} + (3+\lambda)\zeta\Big)}_{:=C} \, \eta_0 \sum_{i=t-\tau_{\text{mix}}}^{t-1} \alpha^i \\
&&& \text{(by definition of the exponential step-sizes)} \\
&= C\,\eta_0\,\alpha^{t-\tau_{\text{mix}}} \sum_{i=0}^{\tau_{\text{mix}}-1} \alpha^i \\
&= C\,\eta_0\,\frac{\alpha^t}{\alpha^{\tau_{\text{mix}}}}\,\frac{1-\alpha^{\tau_{\text{mix}}}}{1-\alpha} = C\,\eta_t\,\frac{\alpha^{-\tau_{\text{mix}}}-1}{1-\alpha} \\
&\leq 8\,C\,\eta_t\,\max\{a,b\}\,\ln(T') && \text{(using Lemma E.6)} \\
\implies \|w_t - w_{t-\tau_{\text{mix}}}\|^2 &\leq 64\,C^2\,\eta_t^2\,[\max\{a,b\}]^2\,\ln^2(T') \\
&= 64\left((2+\lambda)\sqrt{B(\tau_{\text{mix}})} + (3+\lambda)\zeta\right)^2 \eta_t^2\,[\max\{a,b\}]^2\,\ln^2(T') \\
&\leq 256\left((2+\lambda)\sqrt{B(\tau_{\text{mix}})} + (3+\lambda)\zeta\right)^2 \eta_t^2\,[\max\{a,b\}]^2\,\ln^2(T) \\
&&& \text{(for } \eta_0 \leq 1 \text{ and } T \geq 1,\ \tfrac{\ln(T')}{\ln(T)} \leq 2) \\
&\leq 256\,\left[2\,(2+\lambda)^2\,B(\tau_{\text{mix}}) + 2\,(3+\lambda)^2\,\zeta^2\right]\,\eta_t^2\,[\max\{a,b\}]^2\,\ln^2(T) \\
&&& \text{(since } (x+y)^2 \leq 2x^2 + 2y^2) \\
&\leq 256\,\left[2\,(2+\lambda)^2\,B(\tau_{\text{mix}}) + 2\,(3+\lambda)^2\,\zeta^2\right]\,\eta_t^2\,\ln^4(T) \\
&&& \text{(since } \ln(T) \geq \max\{a,b\}) \\
&\leq \underbrace{2560\,(2+\lambda)^2}_{:=c_1^2}\,B(\tau_{\text{mix}})\,\eta_t^2\,\ln^4(T) \\
&&& \text{(since } \zeta^2 \leq B(\tau_{\text{mix}}) \text{ and } 2\,(3+\lambda)^2 \leq 8\,(2+\lambda)^2) \\
&= c_1^2\,B(\tau_{\text{mix}})\,\eta_t^2\,\ln^4(T)
\end{aligned}$$

□

Lemma 4.7. *Let $T \geq \max\{3, \frac{1}{\eta_0}\}$, and T large enough such that $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, and $\ln(T) \geq \max\{a, b\}$. For $t \geq \tau_{mix}$, suppose $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ for all $k \in [t]$. Then:*

$$\begin{aligned}&\mathbb{E}_t\left[\langle g_t^r(w_t) - g^r(w_t),\, w_t - w_r^*\rangle\right]\\ \leq& C\,\eta_t\,\ln^2(T)\,B(\tau_{mix}),\end{aligned}$$

where $C = C_1 + 3 + 2C_2$, $C_1 = \frac{c_1}{2}$ and $C_2 = \frac{c_1 c_2}{2}$, $c_1 = 2560(2+\lambda)^2$ and $c_2 = 4\,(2+\lambda)^2 + 4\,(3+\lambda)^2 + 2\,(2+\lambda)^2$.

Proof. Following Mitra (2025), we decompose as: $\langle w_t - w_r^*, g_t^r(w_t) - g^r(w_t)\rangle = T_1 + T_2 + T_3 + T_4$, where

$$\begin{aligned}T_1 &= \langle w_t - w_{t-\tau_{\text{mix}}}, g_t^r(w_t) - g^r(w_t)\rangle,\\ T_2 &= \langle w_{t-\tau_{\text{mix}}} - w_r^*, g_t^r(w_{t-\tau_{\text{mix}}}) - g^r(w_{t-\tau_{\text{mix}}})\rangle,\\ T_3 &= \langle w_{t-\tau_{\text{mix}}} - w_r^*, g_t^r(w_t) - g_t^r(w_{t-\tau_{\text{mix}}})\rangle, \text{ and}\\ T_4 &= \langle w_{t-\tau_{\text{mix}}} - w_r^*, g^r(w_{t-\tau_{\text{mix}}}) - g^r(w_t)\rangle.\end{aligned}$$

For T_1:

$$\begin{aligned}T_1 &\leq \|w_t - w_{t-\tau_{\text{mix}}}\|\,\|g_t^r(w_t) - g^r(w_t)\| && \text{(by Cauchy Schwarz)}\\ &\leq \frac{1}{2\,c_1\,\eta_t\,\ln^2(T)}\|w_t - w_{t-\tau_{\text{mix}}}\|^2 + \frac{c_1\,\eta_t\,\ln^2(T)}{2}\,\|g_t^r(w_t) - g^r(w_t)\|^2 && \text{(by Young's inequality)}\\ &\leq \frac{c_1\,\eta_t\,\ln^2(T)\,B(\tau_{\text{mix}})}{2} + \frac{c_1\,\eta_t\,\ln^2(T)}{2}\,\|g_t^r(w_t) - g^r(w_t)\|^2 && \text{(using Lemma 4.6)}\end{aligned}$$

Simplifying $\|g_t^r(w_t) - g^r(w_t)\|^2$,

$$\begin{aligned}\|g_t^r(w_t) - g^r(w_t)\|^2 &\leq 2\,\|g_t^r(w_t)\|^2 + 2\,\|g^r(w_t)\|^2 && \text{(since } (x+y)^2 \leq 2x^2 + 2y^2)\\ &\leq 2\,[(2+\lambda)\,\|w_t - w_r^*\| + (3+\lambda)\zeta]^2 + 2\,[(2+\lambda)\,\|w_t - w_r^*\|]^2\\ &&\text{(using Lemma 4.2 and Lemma 4.3)}\\ &\leq 4\,(2+\lambda)^2\,\|w_t - w_r^*\|^2 + 4\,(3+\lambda)^2\,\zeta^2 + 2\,(2+\lambda)^2\,\|w_t - w_r^*\|^2\\ &&\text{(since } (x+y)^2 \leq 2x^2 + 2y^2)\\ &\leq 4\,(2+\lambda)^2\,B(\tau_{\text{mix}}) + 4\,(3+\lambda)^2\,B(\tau_{\text{mix}}) + 2\,(2+\lambda)^2\,B(\tau_{\text{mix}})\\ &&\text{(since } d_k \leq B(\tau_{\text{mix}}) \text{ for all } k \in [t] \text{ and } \zeta^2 \leq B(\tau_{\text{mix}}))\\ &= \underbrace{\left(4\,(2+\lambda)^2 + 4\,(3+\lambda)^2 + 2\,(2+\lambda)^2\right)}_{:=c_2}\,B(\tau_{\text{mix}})\end{aligned}$$

$$\implies \|g_t^r(w_t) - g^r(w_t)\|^2 \leq c_2\,B(\tau_{\text{mix}})$$

Combining the above inequalities,

$$T_1 \leq \frac{c_1\,\eta_t\,\ln^2(T)\,B(\tau_{\text{mix}})}{2} + \frac{c_1\,c_2\,\eta_t\,\ln^2(T)}{2}\,B(\tau_{\text{mix}}) = \eta_t\,\ln^2(T)\,B(\tau_{\text{mix}})\,\underbrace{\left[\frac{c_1}{2} + \frac{c_1\,c_2}{2}\right]}_{:=C_1}$$

$$\implies T_1 \leq C_1\,\eta_t\,\ln^2(T)\,B(\tau_{\text{mix}})$$

For T_3:

$$\begin{aligned}
T_3 &\leq \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, \|g_t^r(w_t) - g_t^r(w_{t-\tau_{\text{mix}}})\| && \text{(by Cauchy Schwarz)} \\
&= \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, \|g_t(w_t) - g_t(w_{t-\tau_{\text{mix}}}) - \lambda(w_t - w_{t-\tau_{\text{mix}}})\| && \text{(by definition)} \\
&\leq \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (\|g_t(w_t) - g_t(w_{t-\tau_{\text{mix}}})\| + \lambda \, \|w_t - w_{t-\tau_{\text{mix}}}\|) && \text{(triangle inequality)} \\
&\leq \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (2 \, \|w_t - w_{t-\tau_{\text{mix}}}\| + \lambda \, \|w_t - w_{t-\tau_{\text{mix}}}\|) && \text{(by Eq. (8))} \\
&= \|w_{t-\tau_{\text{mix}}} - w_r^*\| \;\; (2+\lambda) \; \|w_t - w_{t-\tau_{\text{mix}}}\| \\
&\leq \frac{1}{2\, c_1 \, \eta_t \; \ln^2(T)} \, \|w_t - w_{t-\tau_{\text{mix}}}\|^2 + \frac{c_1 \, \eta_t \, (2+\lambda) \; \ln^2(T)}{2} \; \|w_{t-\tau_{\text{mix}}} - w_r^*\|^2 \\
& && \text{(by Young's inequality)} \\
&\leq \frac{c_1 \, \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}})}{2} + \frac{c_1 \, \eta_t \, (2+\lambda) \; \ln^2(T)}{2} \; \|w_{t-\tau_{\text{mix}}} - w_r^*\|^2 && \text{(by Lemma 4.6)} \\
&\leq \frac{c_1 \, \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}})}{2} + \frac{c_1 \, \eta_t \, (2+\lambda) \; \ln^2(T) \, B(\tau_{\text{mix}})}{2} \\
& && \text{(since } d_k \leq B(\tau_{\text{mix}}) \text{ for all } k \in [t]) \\
&= \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}}) \; \underbrace{\left[\frac{c_1}{2} + \frac{c_1 \, (2+\lambda)}{2}\right]}_{:= C_2}
\end{aligned}$$

$$\implies T_3 \leq C_2 \, \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}})$$

For T_4, the same analysis applies.

$$\begin{aligned}
T_4 &\leq \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, \|g^r(w_{t-\tau_{\text{mix}}}) - g^r(w_t)\| && \text{(by Cauchy Schwarz)} \\
&= \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, \|g(w_t) - g(w_{t-\tau_{\text{mix}}}) - \lambda(w_t - w_{t-\tau_{\text{mix}}})\| && \text{(by definition)} \\
&\leq \|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (2 \, \|w_t - w_{t-\tau_{\text{mix}}}\| + \lambda \, \|w_t - w_{t-\tau_{\text{mix}}}\|) && \text{(by Eq. (7))} \\
&= \|w_{t-\tau_{\text{mix}}} - w_r^*\| \;\; (2+\lambda) \; \|w_t - w_{t-\tau_{\text{mix}}}\|
\end{aligned}$$

Following the same analysis for T_3, we get that,

$$T_4 \leq C_2 \, \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}})$$

For T_2, following the proof in Mitra (2025):

$$\begin{aligned}
\mathbb{E}\left[T_2\right] &= \mathbb{E}\left[\langle w_{t-\tau_{\text{mix}}} - w_r^*, g_t^r(w_{t-\tau_{\text{mix}}}) - g^r(w_{t-\tau_{\text{mix}}})\rangle\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\langle w_{t-\tau_{\text{mix}}} - w_r^*, g_t^r(w_{t-\tau_{\text{mix}}}) - g^r(w_{t-\tau_{\text{mix}}})\rangle | w_{t-\tau_{\text{mix}}}\right]\right] \\
&= \mathbb{E}\left[\langle w_{t-\tau_{\text{mix}}} - w_r^*, \mathbb{E}\left[g_t^r(w_{t-\tau_{\text{mix}}}) - g^r(w_{t-\tau_{\text{mix}}}) | w_{t-\tau_{\text{mix}}}\right]\rangle\right] \\
&\leq \mathbb{E}[\|w_{t-\tau_{\text{mix}}} - w_r^*\| \, \|\mathbb{E}[g_t^r(w_{t-\tau_{\text{mix}}}) - g^r(w_{t-\tau_{\text{mix}}}) \mid w_{t-\tau_{\text{mix}}}]\|] \\
& && \text{(by Cauchy Schwarz)} \\
&\leq \eta_T \mathbb{E}\left[\|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (1 + \|w_{t-\tau_{\text{mix}}}\|)\right] && \text{(by Eq. (5))} \\
&\leq \eta_t \mathbb{E}\left[\|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (1 + \|w_{t-\tau_{\text{mix}}}\|)\right] && \text{(exponential step-size decreases)} \\
&\leq \eta_t \mathbb{E}\left[\|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (1 + \|w_r^*\| + \|w_{t-\tau_{\text{mix}}} - w_r^*\|)\right] && \text{(triangle inequality)} \\
&\leq \eta_t \mathbb{E}\left[\|w_{t-\tau_{\text{mix}}} - w_r^*\| \, (2\zeta + \|w_{t-\tau_{\text{mix}}} - w_r^*\|)\right] && \text{(by the definition of } \zeta) \\
&\leq 3\eta_t B(\tau_{\text{mix}}) \\
& \quad \text{(assuming } d_k \leq B(\tau_{\text{mix}}), \forall k \in [t].\ \|w_{t-\tau_{\text{mix}}} - w_r^*\| \leq B(\tau_{\text{mix}}),\ \zeta \leq B(\tau_{\text{mix}}))
\end{aligned}$$

$$\implies \mathbb{E}\left[T_2\right] \leq 3 \, \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}})$$

Combining T_1, T_2, T_3, T_4 yields:

$$e_t \leq \underbrace{(C_1 + 3 + 2C_2)}_{:= C} \; \eta_t \; \ln^2(T) \, B(\tau_{\text{mix}})$$

□

In addition to e_t, we also need to upper bound $\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[\|g_t^r(w_t) - g^r(w_t)\|^2\right]$.

Lemma 4.8. *Assuming* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix}), \forall k \in [t]$*, then we have*

$$\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[\|g_t^r(w_t) - g^r(w_t)\|^2\right] \leq C' B(\tau_{mix}),$$

where $C' = 10(3+\lambda)^2$.

Proof.

$$\begin{aligned}
&\mathbb{E}_{s_t \sim P_\pi^t \mu_0}\left[\|g_t^r(w_t) - g^r(w_t)\|^2\right] \\
\leq &\mathbb{E}\left[2\|g_t^r(w_t)\|^2 + 2\|g^r(w_t)\|^2\right]. && (\|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2) \\
\leq &2\mathbb{E}\left[((2+\lambda)\|w_t - w_r^*\| + (3+\lambda)\zeta)^2 + ((2+\lambda)\|w_t - w_r^*\|)^2\right] \\
& && \text{(by Lemma 4.2 and Lemma 4.3)} \\
\leq &\underbrace{10(3+\lambda)^2}_{:=C'} B(\tau_{\text{mix}}). && (\text{assuming that } d_k \leq B(\tau_{\text{mix}}), \forall k \in [t])
\end{aligned}$$

□

For the subsequent steps, the proofs for standard TD(0) and regularized TD(0) deviate. We will first provide the convergence rate for the standard TD(0) where the step-size depends on ω, subsequently show that regularized TD(0) removes the requirement on ω.

E.1 Standard TD(0)

Lemma 4.9. *For the standard TD(0) update, when* $T \geq \max\{3, \frac{1}{\eta_0}, \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$*, and* $\ln(T) \geq \max\{a, b\}$*, for a fixed* t*, if* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t$ *and*

$$\eta_0 \leq \frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T) + C'\right]},$$

then $\|w_{t+1} - w^*\|^2 \leq B(\tau_{mix})$*, and hence,* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t+1$.

Proof. We use the above lemmas and prove the result by induction. We assume that for any $t \geq \tau_{\text{mix}}$, $d_k \leq B(\tau_{\text{mix}}), \forall k \in [t]$. Now we show that with an appropriate choice of η_0, we have $d_{k+1} \leq B$. We continue the expansion in Eq. (6) and take expectation w.r.t the randomness in iteration t:

$$\begin{aligned}
&\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \\
&\leq \|w_t - w^*\|^2 + 2\eta_t \mathbb{E}\left[\langle g_t(w_t) - g(w_t), w_t - w^*\rangle\right] + 2\eta_t^2 \mathbb{E}\left[\|g_t(w_t) - g(w_t)\|^2\right] + 2\eta_t^2\|g(w_t)\|^2 + 2\eta_t\langle g(w_t), w_t - w^*\rangle \\
&\leq \|w_t - w^*\|^2 + 2\eta_t\langle g(w_t), w_t - w^*\rangle + 2\eta_t^2\|g(w_t)\|^2 + 2C\,\eta_t^2\,\ln^2(T)\,B(\tau_{\text{mix}}) + 2C'\,\eta_t^2\,B(\tau_{\text{mix}}) \\
& && \text{(using Lemmas 4.7 and 4.8)} \\
&= \|w_t - w^*\|^2 + 2\eta_t\langle g(w_t), w_t - w^*\rangle + 2\eta_t^2\|g(w_t)\|^2 + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}})
\end{aligned}$$

We notice that $\|w_t - w^*\|^2 + 2\eta_t\langle g(w_t), w_t - w^*\rangle + 2\eta_t^2\|g(w_t)\|^2$ is similar to the analysis for the mean-path update in Appendix B. We continue the analysis as follows:

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \|w_t - w^*\|^2 + \left(16\eta_t^2 - 2(1-\gamma)\eta_t\right)\|V_{w_t} - V_{w^*}\|^2 + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}})$$

(by Lemma 3.1 and Lemma 3.3)

Setting $\eta_0 \leq \frac{1-\gamma}{16\,\ln(T)} < 1$, we can guarantee $\eta_t \leq \frac{1-\gamma}{16}$, thus

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] &\leq \|w_t - w^*\|^2 - (1-\gamma)\eta_t\|V_{w_t} - V_{w^*}\|^2 + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}}) \\
&\leq \|w_t - w^*\|^2 - (1-\gamma)\eta_t\omega\|w_t - w^*\|^2 + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}}) \\
& && \text{(by Lemma 3.2)}
\end{aligned}$$

By assumption $d_k \leq B(\tau_{\text{mix}}), \forall k \in [t]$, and since $(1-\gamma)\eta_t\omega \leq 1$, we have

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \left(1 - (1-\gamma)\eta_t\omega + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\right) B(\tau_{\text{mix}}).$$

When $\eta_0 \leq \min\left\{\frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T)+C'\right]}, \frac{1-\gamma}{16\,\ln(T)}\right\}$, we have $\left(1 - (1-\gamma)\eta_t\omega + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\right) \leq 1$.

Furthermore, since $0 < (1-\gamma) \leq 1, 0 < \omega < 1, T \geq 3$, we know that $\eta_0 \leq 1$, and consequently, $(1-\gamma)\eta_t\omega \leq (1-\gamma)\eta_0\omega < 1$, implying that $1 - (1-\gamma)\eta_t\omega > 0$.

Hence, $\left(1 - (1-\gamma)\eta_t\omega + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\right) \in (0,1)$, and consequently,

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq B(\tau_{\text{mix}}),$$

which completes the induction.

Plugging in the value of C and C' with $\lambda = 0$, we have $\frac{(1-\gamma)\omega}{2[C\ln^2(T)+C']} = \frac{(1-\gamma)\omega}{446\ln^2(T)+180}$. Since $\omega \leq 1$, $\ln^2(T) \geq \ln(T)$, we have $\frac{(1-\gamma)\omega}{446ln^2(T)+180} \leq \frac{1-\gamma}{16\ln(T)}$. Thus it suffices to have $\eta_0 \leq \frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T)+C'\right]}$. □

The next theorem quantifies the convergence rate of standard TD(0) with exponential step-sizes under Markovian sampling.

Theorem 4.10. *The standard TD(0) update with exponential step-sizes $\eta_t = \eta_0\alpha_t$, where $\eta_0 = \frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T)+C'\right]}$, $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$, and $T \geq \max\{\frac{1}{\eta_0}, \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}\}$, and T large enough such that $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, and $\ln(T) \geq \max\{a, b\}$, achieves the following convergence rate:*

$$\begin{aligned}&\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right]\\ =&O\left(\exp\left(-\frac{\omega^2 T}{\ln^3(T)}\right) + \frac{\ln^4(T)}{\omega^2 T}\exp\left(\frac{m}{\ln(1/\rho)}\right)\right),\end{aligned}$$

where m and ρ are related to mixing time as $\tau_{mix} = \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}$.

Proof. Continuing the one-step expansion with step-size $\eta_0 \leq \frac{1-\gamma}{16}$ as in Lemma 4.9, we have that, for the absolute constants C and C' defined in Lemma 4.7 and Lemma 4.8 respectively,

$$\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \leq \|w_t - w^*\|^2 - (1-\gamma)\eta_t\omega\,\|w_t - w^*\|^2 + \underbrace{2\left[C\,\ln^2(T) + C'\right]}_{:=C(T)}\eta_t^2\,B(\tau_{\text{mix}})$$

Taking expectation over $t \in [T]$, we have:

$$\begin{aligned}\mathbb{E}\left[\|w_T - w^*\|^2\right] \leq \|w_0 - w^*\|^2 \exp\left(-\eta_0\omega(1-\gamma)\sum_{t=1}^{T}\alpha^t\right)\\ + C(T)\,B(\tau_{\text{mix}})\eta_0^2\sum_{t=1}^{T}\alpha^{2t}\exp\left(-\eta_0\omega(1-\gamma)\sum_{i=t+1}^{T}\alpha^i\right).\end{aligned}$$

This result has the same form as in Section 3. Applying Lemma F.1 and Lemma F.2, we obtain the convergence rate:

$$\begin{aligned}\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right] &\leq \|w_0 - w^*\|^2\, e\exp\left(-\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right) + \frac{8C(T)\,B(\tau_{\text{mix}})}{e(\omega(1-\gamma))^2}\frac{\ln^2(T)}{\alpha^2 T}\\ &= O\left(\exp\left(-\frac{\omega^2\,T}{\ln^3(T)}\right) + \frac{\ln^4(T)}{\omega^2 T}\exp\left(\frac{m}{\ln(1/\rho)}\right)\right),\\ &\qquad\qquad\text{(plugging in the values of } \eta_0, B(\tau_{\text{mix}}), C(T))\end{aligned}$$

where m and ρ are related to mixing time as $\tau_{\text{mix}} = \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}$.

Additionally, for the condition of $T \geq \max\{\frac{1}{\eta_0}, 3\}$. When $\eta_0 \leq \frac{(1-\gamma)\omega}{2\left[C\,\ln^2(T)+C'\right]}$, $T \geq 1/\eta_0$ implies $T \geq 3$, thus it suffices that $T \geq \frac{1}{\eta_0}$. □

E.2 Regularized TD(0)

Now we provide the proof for regularized TD(0), and demonstrate that it does not require ω.

Lemma 4.11. *For the regularized TD(0) update, when* $T \geq \max\{3, \frac{1}{\eta_0}, \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, *and* $\ln(T) \geq \max\{a, b\}$*, for a fixed* t*, if* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t$*. When the step-size satisfies*

$$\eta_0 \leq \frac{\lambda}{[C\,\ln^2(T) + C'] + (8 + 2\lambda^2)},$$

then $\|w_{t+1} - w_r^*\|^2 \leq B(\tau_{mix})$*, and hence,* $\|w_k - w_r^*\|^2 \leq B(\tau_{mix})$ *for all* $k \leq t+1$.

Proof. We use the above lemmas and prove the result by induction. We assume that for any $t \geq \tau_{\text{mix}}$, $d_k \leq B(\tau_{\text{mix}}), \forall k \in [t]$. Now we show that with an appropriate choice of η_0, we have $d_{k+1} \leq B$. We continue the expansion in Eq. (6) and take expectation w.r.t the randomness in iteration t:

$$\begin{aligned}
&\mathbb{E}\left[\|w_{t+1} - w^*\|^2\right] \\
&\leq \|w_t - w^*\|^2 + 2\eta_t \mathbb{E}\left[\langle g_t(w_t) - g(w_t), w_t - w^*\rangle\right] + 2\eta_t^2 \mathbb{E}\left[\|g_t(w_t) - g(w_t)\|^2\right] + 2\eta_t^2 \|g(w_t)\|^2 + 2\eta_t \langle g(w_t), w_t - w^*\rangle \\
&\leq \|w_t - w^*\|^2 + 2\eta_t \langle g(w_t), w_t - w^*\rangle + 2\eta_t^2 \|g(w_t)\|^2 + 2C\,\eta_t^2\,\ln^2(T)\,B(\tau_{\text{mix}}) + 2C'\,\eta_t^2\,B(\tau_{\text{mix}}) \\
&\qquad\qquad \text{(using Lemmas 4.7 and 4.8)} \\
&= \|w_t - w^*\|^2 + 2\eta_t \langle g(w_t), w_t - w^*\rangle + 2\eta_t^2 \|g(w_t)\|^2 + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}})
\end{aligned}$$

We notice that $\|w_t - w^*\|^2 + 2\eta_t \langle g(w_t), w_t - w^*\rangle + 2\eta_t^2 \|g(w_t)\|^2$ is similar to the analysis for a mean-path update. We continue the analysis as follows:

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - w_r^*\|^2\right] &\leq \|w_t - w_r^*\|^2 + \left[2(8 + 2\lambda^2)\eta_t^2 - 2\lambda\eta_t\right]\|w_t - w_r^*\|^2 - 2\eta_t(1-\gamma)\omega\,\|w_t - w_r^*\|^2 \\
&\quad + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}}) \qquad \text{(by Lemma E.1 and Lemma E.2)} \\
&\leq \left(1 + \left[2(8 + 2\lambda^2)\eta_t^2 - 2\lambda\eta_t\right]\right)\|w_t - w_r^*\|^2 + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\,B(\tau_{\text{mix}})
\end{aligned}$$

If $\eta_0 < \frac{1}{2\lambda}$, $\eta_t < \frac{1}{2\lambda}$ and consequently, $2(8 + 2\lambda^2)\eta_t^2 - 2\lambda\eta_t > -1$. Hence, for $\eta_0 \leq \frac{1}{2\lambda}$, $1 + 2(8 + 2\lambda^2)\eta_t^2 - 2\lambda\eta_t > 0$.

By the assumption $d_k \leq B(\tau_{\text{mix}})$ for all $k \in [t]$, and consequently,

$$\mathbb{E}\left[\|w_{t+1} - w_r^*\|^2\right] \leq \left(1 + 2(8 + 2\lambda^2)\eta_t^2 - 2\lambda\eta_t + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\right)B(\tau_{\text{mix}})$$

For $\eta_0 \leq \frac{\lambda}{[C\,\ln^2(T)+C'] + (8+2\lambda^2)}$, $\left(1 + 2(8 + 2\lambda^2)\eta_t^2 - 2\lambda\eta_t + 2\left[C\,\ln^2(T) + C'\right]\eta_t^2\right) \leq 1$. Hence,

$$\mathbb{E}\left[\|w_{t+1} - w_r^*\|^2\right] \leq B(\tau_{\text{mix}})$$

This completes the induction.

□

We now state the final convergence rate for regularized TD(0) under Markovian sampling.

Theorem 4.12. *Apply reguarlized TD(0) with exponential step-size* $\eta_t = \eta_0\alpha_t$*, where* $\eta_0 = \frac{\lambda}{[C\,\ln^2(T)+C'] + (8+2\lambda^2)}$, $\alpha_t = \alpha^t = \frac{1}{T}^{t/T}$*, and* $T \geq \max\{\frac{1}{\eta_0}, \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}\}$*, and* T *large enough such that* $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$*, and* $\ln(T) \geq \max\{a, b\}$*, and* $\lambda = 1/\sqrt{T}$*, where* $a = \frac{1}{\ln(1/\rho)}$ *and* $b = \frac{\ln(2(2+\lambda)m)}{\ln(1/\rho)}$*. Then we have the convergence rate:*

$$\begin{aligned}
&\mathbb{E}\left[\|w_{T+1} - w^*\|^2\right] \\
=&O\left(\exp\left(-\frac{\omega\sqrt{T}}{\ln^3(T)}\right) + \frac{\ln^4(T)}{\omega^2 T}\exp\left(\frac{m}{\ln(1/\rho)}\right)\right),
\end{aligned}$$

where m *and* ρ *are related to mixing time as* $\tau_{mix} = \frac{\ln(2(2+\lambda)Tm/\eta_0)}{\ln(1/\rho)}$.

Proof. As in the proof of Lemma 4.11, we obtain that if $\eta_0 \leq \min\left\{\frac{1}{2\lambda}, \frac{1-\gamma}{16\,\ln(T)}, \frac{\lambda}{[C\,\ln^2(T)+C']+(8+2\lambda^2)}\right\}$,

$$\|w_{t+1}-w_r^*\|^2 \leq \|w_t-w_r^*\|^2 + \left[2(8+2\lambda^2)\eta_t^2 - 2\lambda\eta_t\right]\|w_t-w_r^*\|^2 - 2\eta_t(1-\gamma)\omega\,\|w_t-w_r^*\|^2 + \underbrace{2\,[C\,\ln^2(T)+C']}_{:=C(T)}\,\eta_t^2\,B(\tau_{\text{mix}})$$

Moreover for $C''=10$, since $\eta_t \leq \eta_0$, if $\eta_0 \leq \frac{\lambda}{C''} \leq \frac{2\,\lambda}{2\,(8+2\lambda^2)}$, $2(8+2\lambda^2)\eta_t^2 - 2\lambda\eta_t < 0$.

Hence, for $\eta_0 = \min\left\{\frac{1}{2\lambda}, \frac{1-\gamma}{16\,\ln(T)}, \frac{\lambda}{[C\,\ln^2(T)+C']+C''}, \frac{\lambda}{C''}\right\}$,

$$\|w_{t+1}-w_r^*\|^2 \leq (1-2\eta_t(1-\gamma)\omega)\,\|w_t-w_r^*\|^2 + C(T)\,\eta_t^2 B(\tau_{\text{mix}})$$

Taking expectations over $t\in[T]$ and recursing,

$$\begin{aligned}\mathbb{E}\left[\|w_{T+1}-w_r^*\|^2\right] &\leq \|w_1-w_r^*\|^2\prod_{t=1}^{T}\left(1-2\eta_0\alpha^t(1-\gamma)\omega\right) + C(T)\,B(\tau_{\text{mix}})\eta_0^2\sum_{t=1}^{T}\alpha^{2t}\prod_{i=t+1}^{T}\left(1-2\eta_0\alpha^t(1-\gamma)\omega\right)\\ &\leq \|w_1-w_r^*\|^2\exp\left(-2\eta_0\omega(1-\gamma)\sum_{t=1}^{T}\alpha^t\right) + C(T)\,B(\tau_{\text{mix}})\eta_0^2\sum_{t=1}^{T}\alpha^{2t}\exp\left(-2\eta_0\omega(1-\gamma)\sum_{i=t+1}^{T}\alpha^i\right)\end{aligned}$$

Similar to the proof in Appendix D, applying Lemma F.1 and Lemma F.2 yields:

$$\mathbb{E}\left[\|w_{T+1}-w_r^*\|^2\right] \leq \|w_1-w_r^*\|^2\,e\exp\left(-2\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right) + C(T)\,B(\tau_{\text{mix}})\frac{4}{e^2(\omega(1-\gamma))^2}\frac{\ln^2(T)}{\alpha^2 T}.$$

Expressing the result in terms of the distance to w^*:

$$\begin{aligned}&\mathbb{E}\left[\|w_{T+1}-w^*\|^2\right]\\ &\leq 2\mathbb{E}\left[\|w_{T+1}-w_r^*\|^2\right] + 2\,\|w_r^*-w^*\|^2 && \text{(since } (x+y)^2\leq 2x^2+2y^2\text{)}\\ &\leq \|w_1-w_r^*\|^2\,2e\exp\left(-2\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right) + \frac{8\,C(T)\,B(\tau_{\text{mix}})}{e^2(\omega(1-\gamma))^2}\frac{\ln^2(T)}{\alpha^2 T} + \frac{2\lambda^2\,\|w^*\|^2}{(1-\gamma)^2\omega^2}. && \text{(by Lemma E.3)}\end{aligned}$$

Setting $\lambda = \frac{1}{\sqrt{T}} < 1$ gives:

$$\begin{aligned}\mathbb{E}\left[\|w_{T+1}-w^*\|^2\right] &\leq \|w_1-w_r^*\|^2\,2e\exp\left(-2\eta_0\omega(1-\gamma)\frac{\alpha T}{\ln(T)}\right) + \frac{8\,C(T)\,B(\tau_{\text{mix}})}{e^2(\omega(1-\gamma))^2}\frac{\ln^2(T)}{\alpha^2 T} + \frac{2\,\|w^*\|^2}{(\omega(1-\gamma))^2 T}\\ &= O\left(\exp\left(-\frac{\omega\sqrt{T}}{\ln^3(T)}\right) + \frac{\ln^4(T)}{\omega^2 T}\exp\left(\frac{m}{\ln(1/\rho)}\right)\right), && \text{(plugging in the values of } \eta_0, B(\tau_{\text{mix}}), C(T)\text{)}\end{aligned}$$

where m and ρ are related to mixing time as $\tau_{\text{mix}} = \frac{\ln(4Tm/\eta_0)}{\ln(1/\rho)}$.

Additionally, for the condition of T. When $\eta_0 \leq \frac{\lambda}{[C\,\ln^2(T)+C']+(8+2\lambda^2)}$, $T \geq 1/\eta_0$ implies $T\geq 3$, thus it suffices that $T \geq \max\{\frac{1}{\eta_0}, \frac{\ln(4Tm/\eta_0)}{\ln 1/\rho}\}$, and T large enough such that $\frac{\ln^2(T)}{T} \leq \frac{1}{2a}$, $\frac{\ln(T)}{T} \leq \frac{1}{b}$, and $\ln(T) \geq \max\{a,b\}$. □

F Helper Lemmas

Lemma F.1.

$$X := \sum_{t=1}^{T}\alpha^t \geq \frac{\alpha T}{\ln(T)} - \frac{1}{\ln(T)}.$$

Proof.

$$\sum_{t=1}^{T} \alpha^t = \frac{\alpha - \alpha^{T+1}}{1-\alpha} = \frac{\alpha}{1-\alpha} - \frac{\alpha^{T+1}}{1-\alpha}.$$

We have

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha}{T(1-\alpha)} = \frac{1}{T}\frac{1}{1/\alpha - 1} \leq \frac{1}{T}\frac{1}{\ln(1/\alpha)} = \frac{1}{\ln(T)},$$

where in the inequality we used Lemma 4 and the fact that $1/\alpha > 1$. Plugging back into X we get

$$X \geq \frac{\alpha}{1-\alpha} - \frac{1}{\ln(T)} \geq \frac{\alpha}{\ln(1/\alpha)} - \frac{1}{\ln(T)} = \frac{\alpha T}{\ln(T)} - \frac{1}{\ln(T)}.$$

□

Lemma F.2. *For $\alpha = \frac{1}{T}^{1/T}$ and any $\kappa > 0$,*

$$\sum_{t=1}^{T} \alpha^{2t} \exp\left(-a \sum_{i=t+1}^{T} \alpha^i\right) \leq \frac{4c\left(\ln(T)\right)^2}{a^2 e^2 \alpha^2 T},$$

where $c = \exp\left(a\frac{1}{\ln(T)}\right)$.

Proof. First, observe that,

$$\sum_{i=t+1}^{T} \alpha^i = \frac{\alpha^{t+1} - \alpha^{T+1}}{1-\alpha}$$

We have

$$\frac{\alpha^{T+1}}{1-\alpha} = \frac{\alpha}{T(1-\alpha)} = \frac{1}{T} \cdot \frac{1}{1/\alpha - 1} \leq \frac{1}{T} \cdot \frac{1}{\ln(1/\alpha)} = \frac{1}{\ln(T)}$$

These relations imply that,

$$\begin{aligned}
&\sum_{i=t+1}^{T} \alpha^i \geq \frac{\alpha^{t+1}}{1-\alpha} - \frac{1}{\ln(T)} \\
&\implies \exp\left(-a \sum_{i=t+1}^{T} \alpha^i\right) \leq \exp\left(-a\frac{\alpha^{t+1}}{1-\alpha} + a\frac{1}{\ln(T)}\right) = c\exp\left(-a\frac{\alpha^{t+1}}{1-\alpha}\right),
\end{aligned}$$

where $c = \exp\left(a\frac{1}{\ln(T)}\right)$. We then have

$$\begin{aligned}
\sum_{t=1}^{T} \alpha^{2t} \exp\left(-a \sum_{i=t+1}^{T} \alpha^i\right) &\leq c \sum_{t=1}^{T} \alpha^{2t} \exp\left(-a\frac{\alpha^{t+1}}{1-\alpha}\right) \\
&\leq c \sum_{t=1}^{T} \alpha^{2t} \left(\frac{2(1-\alpha)}{ea\alpha^{t+1}}\right)^2 \qquad \text{(by Lemma F.3 with } \nu = 2) \\
&= \frac{4c}{a^2 e^2 \alpha^2} T(1-\alpha)^2 \\
&\leq \frac{4c}{a^2 e^2 \alpha^2} T(\ln(1/\alpha))^2 \\
&= \frac{4c(\ln(T))^2}{a^2 e^2 \alpha^2 T}
\end{aligned}$$

□

Lemma F.3. *For all $x, \nu > 0$,*

$$\exp(-x) \leq \left(\frac{\nu}{ex}\right)^{\nu}$$

Proof. Let $x > 0$. Define $f(\nu) = \left(\frac{\nu}{ex}\right)^{\nu} - \exp(-x)$. We have

$$f(\nu) = \exp\left(\nu \ln(\nu) - \nu \ln(ex)\right) - \exp(-x)$$

and

$$f'(\nu) = \left(\nu \cdot \frac{1}{\nu} + \ln(\nu) - \ln(ex)\right) \exp\left(\nu \ln(\nu) - \nu \ln(ex)\right)$$

Thus

$$f'(\nu) \geq 0 \iff 1 + \ln(\nu) - \ln(ex) \geq 0 \iff \nu \geq \exp\left(\ln(ex) - 1\right) = x$$

So f is decreasing on $(0, x]$ and increasing on $[x, \infty)$. Moreover,

$$f(x) = \left(\frac{x}{ex}\right)^{x} - \exp(-x) = \left(\frac{1}{e}\right)^{x} - \exp(-x) = 0$$

and thus $f(\nu) \geq 0$ for all $\nu > 0$ which proves the lemma. □