A Robust $\widetilde{\mathcal{O}}(1/\sqrt{T})$ Rate for Unprojected TD Learning with Linear Function Approximation

Anonymous authors
Paper under double-blind review

Abstract

We investigate the finite-time convergence properties of Temporal Difference (TD) learning with linear function approximation, a cornerstone of reinforcement learning. We are interested in the so-called "robust" setting, where the convergence guarantee does not depend on the potential function's minimal curvature. While prior work has established convergence guarantees in this setting, these results typically rely on the artificial assumption that each iterate is projected onto a bounded set. Removing such a condition was left as an open problem by Bhandari et al. (COLT'18), hypothesizing the need for additional "regularity conditions". In this paper, we show that the simple unprojected TD(0) converges with a rate of $\widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{\theta}^*\|_2^2}{\sqrt{T}}\right)$ in expectation, even in the presence of Markovian noise. We do not require an additional regularity condition, but only a minor polylog correction to the learning rate. Our analysis reveals a novel self-bounding property of the TD updates and exploits it to guarantee bounded iterates.

1 Introduction

Temporal Difference (TD) learning (Sutton, 1988) is a cornerstone of modern reinforcement learning. It provides a model-free approach to policy evaluation, estimating the value function of a given policy within a Markov Decision Process (MDP). The versatility of TD methods has led to applications in diverse domains, including games (Silver et al., 2016), robotics (Gu et al., 2017), and autonomous systems (Chen et al., 2015). At its core, TD learning iteratively updates value function estimates based on the difference between predictions at successive time steps.

Despite its conceptual simplicity and widespread use, the theoretical analysis of TD learning, particularly with linear function approximation in large state spaces, presents considerable challenges. Early seminal work by Tsitsiklis & Van Roy (1996) established asymptotic convergence conditions by framing TD as a stochastic approximation algorithm. More recently, understanding the non-asymptotic behavior and finite-time performance of TD has become an active area of research. Challenges primarily arise from the correlated nature of samples generated by the underlying Markov chain, which can introduce bias and dependencies into the learning updates.

Several studies have provided finite-time analyses under various assumptions on the potential function, the projection step, and the stepsize. In particular, two *complementary* kinds of analyses are known, giving rise to a "robust" convergence rate of $\widetilde{\mathcal{O}}(1/\sqrt{T})$[1] (e.g., Bhandari et al., 2018; Liu & Olshevsky, 2021; Sun et al., 2021) or to a "fast" rate of $\widetilde{\mathcal{O}}(1/T)$ (e.g., Bhandari et al., 2018; Srikant & Ying, 2019; Patil et al., 2023; Samsonov et al., 2024; Mitra, 2024; Li et al., 2025). These two rates are complementary because the hidden constant in the fast rate depends on the inverse square of the curvature of the potential function which, while always present, can be arbitrarily small. Instead, the $\widetilde{\mathcal{O}}(1/\sqrt{T})$ robust rate is independent of the curvature. Hence, in non-asymptotic regimes, the fast rate can be arbitrarily worse than the robust one.[2] This mirrors

[1] $\widetilde{\mathcal{O}}$ hides polylogarithmic terms and may also hide dependencies on the mixing time.
[2] See Appendix A for an in-depth discussion of the literature on this point.

Table 1: Summary of algorithmic inputs and rates for TD(0) with linear function approximation in the literature. The quantity ω is the minimum eigenvalue of $\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi}$. All quantities are defined in Section 3 and their inputs are discussed in Appendix B.

Rate	Paper	Inputs	Without Projection	Bound independent of ω
$\widetilde{\mathcal{O}}\left(\frac{1}{T}\right)$	Bhandari et al. (2018)	ω, ϕ_∞	✗	✗
	Srikant & Ying (2019)	ω, α, ϕ_∞, $\|\boldsymbol{\theta}^*\|_2$	✓	✗
	Patil et al. (2023)	α, ϕ_∞	✓	✗
	Samsonov et al. (2024)	α, ϕ_∞	✓	✗
	Mitra (2024)	ω, α, ϕ_∞	✓	✗
	Li et al. (2025)	ω, ϕ_∞	✓	✗
$\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\right)$	Bhandari et al. (2018)	ϕ_∞	✗	✓
	Liu & Olshevsky (2021)	ϕ_∞	✗	✓
	Sun et al. (2021)	ϕ_∞	✗	✗
	This paper, Theorem 4.2	ϕ_∞	✓	✓

what happens in the stochastic approximation setting, and it is well-known that in practice the robust rate can be preferable, as motivated in Nemirovski et al. (2009).

In this work, we focus on the need for a projection step to achieve robust rates. In fact, for fast rates, the assumption of minimal curvature leads to a contraction that simplifies the analysis, eliminating the need for a projection. However, there are no known results on unprojected TD achieving the robust $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rate, whereas in practice such a projection is never used. It is worth stressing that while such a projection step is not used in practice, but only to simplify the theoretical analysis.[3] Indeed, Bhandari et al. (Section 11 2018) explicitly posed the removal of the projection step as an open problem, in both the fast and robust regime, hypothesizing that it would be possible "under additional regularity conditions." While removing the projection for the fast case was solved after one year by Srikant & Ying (2019), the unprojected robust case remained unresolved.

Contributions. In this paper, we solve one of the open problems posed by Bhandari et al. (2018): For the first time, we provide a finite-time analysis of TD(0) with linear function approximation under Markovian observations *without requiring iterate projection, while achieving a robust rate*. Moreover, we do not require any additional regularity conditions. Instead, we show that changing the learning rate from $\frac{1}{\sqrt{T}}$ or $\frac{1}{\sqrt{t}}$ to $\frac{1}{\sqrt{t}\log t \log T}$ is sufficient to guarantee a self-bounding property of TD: The iterates are constrained, in expectation, to a bounded domain around the optimal solution. Our analysis differs fundamentally from those that aimed to prove the update is a noisy contraction, and it might be of independent interest. We also show a convergence rate $\widetilde{\mathcal{O}}(\frac{\|\boldsymbol{\theta}^*\|_2^2}{\sqrt{T}})$ for the potential that guides the convergence of the TD algorithm, as defined in Liu & Olshevsky (2021). Table 1 summarizes[4] and compares our results with existing finite-time analyses of TD with linear function approximation.

2 Related Work

The initial theoretical understanding of how TD learning with linear function approximation converges over time was established by Tsitsiklis & Van Roy (1996), who framed TD methods as stochastic approximation algorithms (Kushner, 2010). That work did not derive finite-time convergence rates. Subsequent research (Korda & La, 2015; Lakshminarayanan & Szepesvári, 2018; Dalal et al., 2018) did provide such rates, but a significant limitation was the assumption that data are drawn independently from the stationary distribution. In practice, data are typically collected sequentially along a single trajectory of the Markov chain, introducing

[3]Bhandari et al. (2018) says "at this stage, we view this [projection] mainly as a tool that enables clean finite time analysis, rather than a practical algorithmic proposal."

[4]See Appendix B for a precise discussion of the inputs for each algorithm.

temporal correlations between samples. These correlations make it challenging to analyze even the basic TD(0) method.

Bhandari et al. (2018) provided the first finite-time analysis of TD learning under more realistic Markovian data, drawing parallels to stochastic gradient descent. However, their analysis, as well as that of Liu & Olshevsky (2021), requires a projection step to control the magnitude of iterates/updates. Sun et al. (2021) examined Adam-inspired (Kingma, 2014) adaptive TD variants, but they require a projection as well. Here, we remove the need to project, while obtaining the robust rate of $\widetilde{\mathcal{O}}(1/\sqrt{T})$ obtained by Bhandari et al. (2018).

Another line of work leverages the curvature of the potential function. This allowed Srikant & Ying (2019) to be the first to provide finite-time error bounds for TD learning with linear function approximation under Markovian sampling without a projection step, employing a control-theoretic approach based on Lyapunov theory. While elegant, the analysis in Srikant & Ying (2019) relies on stepsizes that depend on the strong-convexity (curvature) parameter of the potential function. Since this parameter is typically unknown, their result only implies the existence of a good but unknown learning rate. Subsequently, Patil et al. (2023) removed the dependence of stepsizes on this strong-convexity parameter, yielding a more practical algorithm, but at the price of requiring a data-dropping variant of TD. Later, Samsonov et al. (2024) improved the analysis of Patil et al. (2023) to obtain high-probability bounds.

In parallel, Mitra (2024) provided a simpler analysis using an inductive two-step argument, while Li et al. (2025) utilized an exponentially decaying stepsize to remove the data-dropping steps. Moreover, Sun et al. (2022) extended the fast analysis to neural networks in the NTK regime. However, in all these results, the non-asymptotic convergence rate can become arbitrarily slow due to ill-conditioned linear mappings. We discuss this caveat in more detail in Section 4.1.

Our proof method is fundamentally different from the above ones, which removed the projection by proving a contraction. Instead, we show that the iterates are bounded for reasons analogous to what happens in Stochastic Gradient Descent (SGD). In fact, SGD can have bounded iterates even for non-strongly convex objectives, as shown, for example, by Xiao (2010); Orabona & Pál (2021); Telgarsky (2022); Ivgi et al. (2023) under various update schemes and assumptions on the potential and stepsizes.

Another minor difference with prior work is our choice of the potential function: We study the potential function proposed in Liu & Olshevsky (2021), which improves earlier formulations by adding a term proportional to the discount factor γ.

3 Notation and Assumptions

We briefly review the required background on Markov Decision Processes (MDPs) and TD learning with linear function approximation. For a comprehensive treatment, we refer the reader to Sutton & Barto (1998) and Mannor et al. (2022).

In the following, vectors are denoted in bold, and all norms are ℓ_2 (i.e., Euclidean) norms unless stated otherwise.

3.1 Discounted Markov Decision Processes

We define a *discounted-reward MDP* as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, comprising a finite state space $\mathcal{S}$ of size n, a finite action space $\mathcal{A}$, a discount factor $\gamma \in (0,1)$, and a transition kernel $\mathcal{P}$, where $P(s' \mid s, a)$ denotes the probability of transitioning from s to s' given action a. The reward $r(s, s')$ for each transition is bounded by r_∞. Given a trajectory starting at s_0, we denote the state at time t by s_t, with the reward after transitioning defined as $r_t \coloneqq r(s_t, s_{t+1})$.

Induced Markov chain. A stationary policy $\mu : \mathcal{S} \to \Delta^{|\mathcal{A}|-1}$ induces a Markov chain with transition probabilities

$$P^\mu(s' \mid s) \;\coloneqq\; \sum_{a \in \mathcal{A}} \mu(a \mid s)\, P(s' \mid s, a), \quad \forall s, s' \in \mathcal{S}.$$

We let $\boldsymbol{P}^\mu \in \mathbb{R}^{n\times n}$ denote the corresponding transition matrix, where the entry at row s and column s' is $[\boldsymbol{P}^\mu]_{s,s'} = P^\mu(s' \mid s)$. Throughout, we denote the expected single-step reward at state s by $r(s) \coloneqq \sum_{s'} P^\mu(s' \mid s)\, r(s,s')$.

In this paper, we focus on the task of *policy evaluation*, where the goal is to compute the value function defined as the expected discounted sum of rewards.

Value functions and Bellman operators. The value function $\boldsymbol{V}^\mu \in \mathbb{R}^n$ associated with policy μ is defined component-wise as $\boldsymbol{V}^\mu(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t\, r_t \mid s_0 = s]$, where the expectation is taken over the trajectory generated by $\boldsymbol{P}^\mu$ starting from $s_0 = s$. The Bellman operator $T^\mu : \mathbb{R}^n \to \mathbb{R}^n$ is defined as

$$(T^\mu \boldsymbol{V})(s) \;\coloneqq\; r(s) + \gamma \sum_{s'\in\mathcal{S}} P^\mu(s' \mid s)\, \boldsymbol{V}(s'), \quad \forall s \in \mathcal{S}.$$

The operator T^μ is a γ-contraction in the ℓ_∞-norm; hence $\boldsymbol{V}^\mu$ is its unique fixed point.

To study the finite-time behavior of the Markov chain, we impose the following standard ergodic condition.

Assumption 3.1. The Markov chain induced by policy μ is irreducible and aperiodic.

Under Assumption 3.1, the Markov chain admits a unique stationary distribution $\pi \in \Delta^{n-1}$ and its vector form $\boldsymbol{\pi}$ satisfies $\boldsymbol{\pi}^\top \boldsymbol{P}^\mu = \boldsymbol{\pi}^\top$ and mixes geometrically:

Theorem 3.2 (Levin & Peres 2017, Theorem 4.9)**.** *There exist constants $1 < C \le 2$ and $\alpha \in [1/2, 1)$ such that*

$$\max_{s\in\mathcal{S}} \left\| (P^\mu)^t(\cdot \mid s) - \pi \right\|_{\mathrm{TV}} \le C\, \alpha^t, \quad \forall t \ge 0,$$

where $\|\cdot\|_{\mathrm{TV}}$ denotes the total variation distance and $(P^\mu)^t(\cdot \mid s)$ is the state distribution after t steps starting from state s.

Based on this, we define the *mixing time* for a tolerance ϵ as $\tau(\epsilon) \coloneqq \min\{t \in \mathbb{N} \mid C\alpha^t \le \epsilon\}$.

3.2 TD(0) with Linear Function Approximation

We consider the approximation of $\boldsymbol{V}^\mu$ with linear mappings and estimate the weights $\boldsymbol{\theta} \in \mathbb{R}^d$ via TD learning.

Linear architecture. Let $\phi_i : \mathcal{S} \to \mathbb{R}$ for $i \in \{1, \dots, d\}$ be the feature mappings. For each $s \in \mathcal{S}$, define the feature vector $\boldsymbol{\phi}(s) \coloneqq [\phi_1(s), \dots, \phi_d(s)]^\top \in \mathbb{R}^d$. We define the feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n\times d}$ such that the row corresponding to state s is $\boldsymbol{\phi}(s)^\top$. The value function is approximated as $V_{\boldsymbol{\theta}}(s) \coloneqq \boldsymbol{\theta}^\top \boldsymbol{\phi}(s)$, or in vector form $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$.

We recall the following standard assumption on the features.

Assumption 3.3. The feature matrix $\boldsymbol{\Phi}$ has full column rank d, and $\|\boldsymbol{\phi}(s)\| \le \phi_\infty$ for[5] all $s \in \mathcal{S}$.

TD error and update. Given the weight $\boldsymbol{\theta}_t$ and a trajectory sample (s_t, s_{t+1}, r_t), the TD error is defined as $\delta_t \coloneqq r_t + \gamma V_{\boldsymbol{\theta}_t}(s_{t+1}) - V_{\boldsymbol{\theta}_t}(s_t)$. The TD(0) updates as

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t\, \delta_t\, \nabla_{\boldsymbol{\theta}} V_{\boldsymbol{\theta}_t}(s_t) = \boldsymbol{\theta}_t + \eta_t \left(r_t + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}(s_{t+1}) - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}(s_t)\right)\boldsymbol{\phi}(s_t) \coloneqq \boldsymbol{\theta}_t + \eta_t \boldsymbol{g}_t,$$

where $\eta_t > 0$ is the stepsize.

Projected Bellman equation. Under Assumptions 3.1–3.3 and suitable η_t, TD(0) converges to $\boldsymbol{\theta}^*$ asymptotically and $\boldsymbol{\theta}^*$ is characterized as the unique solution to the projected Bellman equation (Tsitsiklis & Van Roy, 1996) $\boldsymbol{\Phi}\boldsymbol{\theta}^* = \Pi_{\boldsymbol{D}}\, T^\mu(\boldsymbol{\Phi}\boldsymbol{\theta}^*)$, where $\Pi_{\boldsymbol{D}}$ is the orthogonal projection operator onto the subspace $\{\boldsymbol{\Phi}\boldsymbol{x} \mid \boldsymbol{x} \in \mathbb{R}^d\}$ with respect to the $\boldsymbol{D}$-norm defined below.

D-norms. Let $\boldsymbol{D} \coloneqq \mathrm{diag}(\boldsymbol{\pi})$. Since $\boldsymbol{D} \succ \boldsymbol{0}$, for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we can define the inner product $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\boldsymbol{D}} \coloneqq \boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{y}$ and the norm $\|\boldsymbol{x}\|_{\boldsymbol{D}} \coloneqq \sqrt{\langle \boldsymbol{x}, \boldsymbol{x}\rangle_{\boldsymbol{D}}}$.

[5]The knowledge of ϕ_∞ can be removed by using, for example, a feature normalization scheme.

Algorithm 1 Unprojected TD(0) with linear function approximation

Input: iteration budget T, ϕ_∞, $c > 15 + 18\sqrt{2}$, s_0
$\boldsymbol{\theta}_0 = \mathbf{0}$
for $t = 0, \dots, T-1$ **do**
  Receive trajectory sample (s_t, s_{t+1}, r_t)
  $\boldsymbol{g}_t = \big(r_t + \gamma \boldsymbol{\phi}(s_{t+1})^\top \boldsymbol{\theta}_t - \boldsymbol{\phi}(s_t)^T \boldsymbol{\theta}_t\big)\, \boldsymbol{\phi}(s_t)$
  $\eta_t = \frac{1}{c\, \phi_\infty^2 \log T \log(t+3)\sqrt{t+1}}$
  $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{g}_t$
end for
Output: $\bar{\boldsymbol{\theta}}_T \coloneqq (\sum_{k=0}^{T-1} \eta_k)^{-1} \sum_{k=0}^{T-1} \eta_k \boldsymbol{\theta}_k$

Dirichlet-seminorms. The *Dirichlet seminorm* (Diaconis & Saloff-Coste, 1996; Ollivier, 2018; Liu & Olshevsky, 2021) is defined as

$$\|\boldsymbol{V}\|_{\mathrm{Dir}}^2 \;\coloneqq\; \frac{1}{2} \sum_{s,s' \in \mathcal{S}} \pi(s)\, P^\mu(s' \mid s) \big(V(s') - V(s)\big)^2. \tag{1}$$

Since $\boldsymbol{\Phi}$ is full column rank by Assumption 3.3, the matrix $\boldsymbol{\Phi}^\top \boldsymbol{D} \boldsymbol{\Phi}$ is positive definite. Thus, the minimum eigenvalue $\omega \coloneqq \lambda_{\min}(\boldsymbol{\Phi}^\top \boldsymbol{D} \boldsymbol{\Phi})$ is strictly positive. This quantity ω plays the role of a strong-convexity (curvature) parameter in fast-rate analyses.

Stationary update. The asymptotic behavior of TD(0) is closely tied to the stationary update field

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) \;\coloneqq\; \mathbb{E}\big[\big(r(s,s') + \gamma \boldsymbol{\phi}(s')^\top \boldsymbol{\theta} - \boldsymbol{\phi}(s)^\top \boldsymbol{\theta}\big) \boldsymbol{\phi}(s)\big]\,.$$

Here, the expectation is taken over stationary transitions where $s \sim \pi$ and $s' \sim P^\mu(\cdot \mid s)$. In particular, $\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) = \mathbf{0}$.

4 Unprojected Temporal Difference Learning

In this section, we present our main result: We analyze TD(0) without any projection, as shown in Algorithm 1, and present a robust convergence result for it.

First, we explain what exactly our convergence is. In prior work (e.g., Bhandari et al., 2018), the potential function for the convergence analysis was

$$\big\|\boldsymbol{V}_{\bar{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\big\|_{\boldsymbol{D}}^2\,. \tag{2}$$

Note that when the discount factor $\gamma \to 1$, their rate $\widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{\theta}^*\|^2}{(1-\gamma)\sqrt{T}}\right)$ loses its utility in characterizing the error of estimates $\bar{\boldsymbol{\theta}}_T$ to $\boldsymbol{\theta}^*$ in T. For this reason, we focus on the better potential function proposed by Liu & Olshevsky (2021):

$$f(\boldsymbol{\theta}) \coloneqq (1-\gamma)\, \|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2 + \gamma\, \|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\mathrm{Dir}}^2\,.$$

Clearly, the point $\boldsymbol{\theta}^*$ minimizes $f(\boldsymbol{\theta})$. Moreover, any result obtained using this potential with $(1-\gamma)^{-1}$ scaling implies those obtained using (2) since $\|\boldsymbol{V}_{\boldsymbol{\theta}} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\mathrm{Dir}}^2$ is non-negative. Finally, this potential provides convergence results even when the discount factor $\gamma \to 1$; see the discussion in Liu & Olshevsky (2021).

From a technical point of view, the advantage of this potential is that the stationary update $\bar{\boldsymbol{g}}(\boldsymbol{\theta})$ satisfies the equality in the following theorem. In a sense, the results in Liu & Olshevsky (2021) indicate that this is the "correct" potential function for TD(0).

Lemma 4.1. *(Liu & Olshevsky, 2021, Theorem 1) For any $\boldsymbol{\theta} \in \mathbb{R}^d$, we have*

$$\langle -\bar{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle = f(\boldsymbol{\theta}) - f(\boldsymbol{\theta}^*)\,.$$

Using the above potential function, the following theorem shows that the TD(0) algorithm can converge with a rate of $\widetilde{\mathcal{O}}(\|\boldsymbol{\theta}^*\|^2/\sqrt{T})$ without projections.

Theorem 4.2. *Consider the weighted average iterate $\bar{\boldsymbol{\theta}}_T$ generated by Algorithm 1. Suppose the stepsize parameter c satisfies $c > c_0 := 15 + 18\sqrt{2}$, and the number of iterations T satisfies $\log T \geq \frac{2}{\log^3(1/\alpha)}$. Then, we have:*

(a) For any $t \leq T$, $\mathbb{E}\left[\|\boldsymbol{\theta}_t\|^2\right] \leq \rho_c^2 \max\left\{\frac{r_\infty^2}{\phi_\infty^2}, \|\boldsymbol{\theta}^*\|^2\right\}$, where $\rho_c \to 2$ as $c \to \infty$, and $\rho_c = \mathcal{O}\big(\frac{1}{c-c_0}\big)$ as $c \downarrow c_0$.

(b) $f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*) = \widetilde{\mathcal{O}}\left(\frac{c\rho_c^2 \max\{r_\infty^2, \phi_\infty^2\|\boldsymbol{\theta}^*\|^2\}}{\sqrt{T}}\right)$.

For lack of space, we give the proof of part (a) in Theorem G.1 in the Appendix, while part (b) follows as a corollary at the end of Section 5. Theorem G.1 will also detail how the stepsize parameter c determines the radius multiplier ρ_c. In short, the number c must exceed a threshold to ensure bounded iterates, and increasing c decreases ρ_c.

It is worth stressing that knowing T in advance is not a limitation, because one can use a standard doubling trick (see Appendix J).

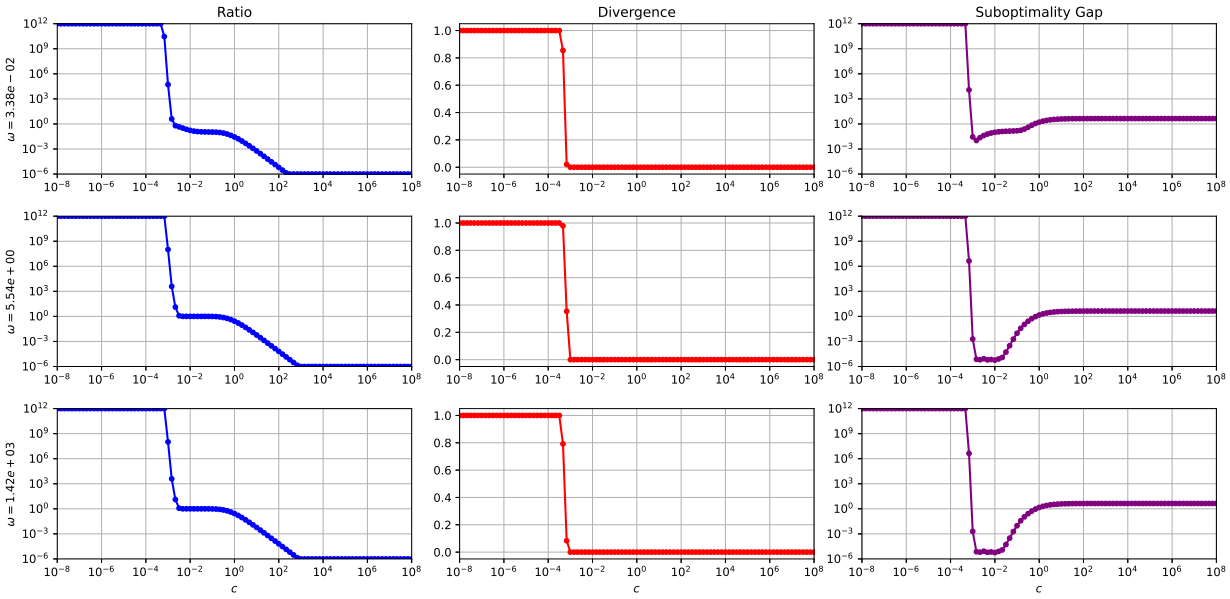

Figure 1: Instability of TD learning. Columns: boundedness ratio, divergence rate, suboptimality gap vs. stepsize scale (c). Rows: different feature scalings (changing the spectrum of $\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi}$).

For the interested reader, we also give a proof sketch in Section 5, where we explain the main steps of the proof, abstracting away the annoying details and contrasting it with previous proofs.

Dependency on α for T? Given that α is typically unknown, one might wonder about the condition on T in the theorem. First of all, *this is a limitation of all analyses on this topic, sometimes hidden.* In fact, all the prior robust bounds have a worse rate than $\widetilde{\mathcal{O}}(1/\sqrt{T})$ if T is not sufficiently large depending on α. This seems to be poorly discussed in the literature, so we devote Appendix I to proving it. Hence, to the best of our knowledge, a dependency of T on α, in an implicit or explicit way, seems unavoidable. It is probably possible to state our result in a similar way, but at the very least the expression of ρ_c would become dependent on α, and we preferred a cleaner expression of the assumptions and results. Notice that even for fast rates, we either need a sufficiently small stepsize that depends on ω (an unknown quantity), or we use

ω-agnostic algorithms (e.g. Patil et al., 2023; Samsonov et al., 2024), which rely on data dropping steps that still require knowledge of α, see Table 1.

Is the threshold on the stepsizes real? Theorem 4.2 gives a sufficient condition on c to have bounded iterates and convergence. First of all, we would like to stress that we did not try to optimize the numerical value of the threshold on c, for the simple reason that any similar analysis with such weak assumptions cannot be predictive of reality.

Instead, the more interesting question is to check if the condition is necessary too. That is, does TD(0) without projection have bounded iterates in finite time with arbitrary stepsizes satisfying the condition in Tsitsiklis & Van Roy (1996)? To test this effect, we conducted an experiment in which we ran TD(0) on a synthetic problem (details in Appendix H). In Figure 1, first column, we show the expected boundedness ratio, defined as $\frac{\max_{t\leq T}\mathbb{E}\left[\|\boldsymbol{\theta}_t\|^2\right]}{\|\boldsymbol{\theta}^*\|^2}$, which is large if the iterates blow up. The second column shows the divergence rate, that is, the fraction of runs with $\|\boldsymbol{\theta}_t\|^2 > 10^{12}$, while the third column shows the suboptimality gap. Overall, we have that the threshold on the stepsizes is indeed real. In fact, from both the expected boundedness ratio and the divergence rate, it is clear that if c is too small, the iterates of the algorithm are not controlled in finite time. Moreover, the explosion in both of these measures nicely mirrors the theoretical behavior in our function ρ_c in Theorem 4.2.

The rows correspond to different spectral characteristics of $\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi}$. Hence, changing the spectral characteristics does not change much the iterates, as our theory suggests. However, due to the complementarity of the robust and fast rates, it does influence the suboptimality gap, suggesting that a robust rate of $\widetilde{\mathcal{O}}(1/\sqrt{T})$ is pessimistic when the strong convexity is large.

In Appendix H, we also report experiments with a fixed stepsize that show the same behaviors.

4.1 Detailed Comparison with Previous Results

Here, we highlight a few technical differences between our analysis and existing finite-time results for TD(0) with linear function approximation.

Comparison with robust $\widetilde{\mathcal{O}}(1/\sqrt{T})$ rates with projections. Bhandari et al. (2018, Theorem 3.(a)) proves that *projected* TD with stepsize $\eta_t = \frac{1}{\sqrt{T}}$ satisfies

$$\mathbb{E}\left[\|\boldsymbol{V}_{\bar{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2\right] = \widetilde{\mathcal{O}}\left(\frac{R^2}{(1-\gamma)\sqrt{T}}\right),$$

where R is the projection radius and requires $R \geq \|\boldsymbol{\theta}^*\|$. Our bound implies the same type of D-norm guarantee, since $\|\cdot\|_{\mathrm{Dir}}^2 \geq 0$. Also, their choice of the learning rate and ours have the same dependency on T, up to polylogarithmic terms, that is, we do not gain more stability by using a much smaller learning rate, but rather with a refined analysis. Moreover, our bound depends explicitly on $\|\boldsymbol{\theta}^*\|$ rather than on an a priori radius R.

We stress once again that the use of a projection step in Bhandari et al. (2018) is only for the purpose of analysis, but not a real practical possibility. In fact, while one can obtain a (potentially very loose) upper bound on $\|\boldsymbol{\theta}^*\|$ in terms of $\omega = \lambda_{\min}(\boldsymbol{\Phi}^\top \boldsymbol{D}\boldsymbol{\Phi})$ (Bhandari et al., 2018, Lemma 1), this approach is impractical from an algorithm design perspective. In fact, the goal of TD learning is precisely to avoid the computational complexity that scales with the number of states n, which is highly non-trivial when estimating ω involves computing the stationary matrix $\boldsymbol{D}$.

Comparison with fast $\widetilde{\mathcal{O}}(1/T)$ rates. It is instructive to compare our results with the known fast rates, to underline their complementarity. In the fast regime (Srikant & Ying, 2019; Patil et al., 2023; Mitra, 2024; Samsonov et al., 2024; Li et al., 2025), contraction-based arguments yield bounds of the form

$$\mathbb{E}\left[\|\boldsymbol{V}_{\bar{\boldsymbol{\theta}}_T} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2\right] = \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{\theta}^*\|^2}{(1-\gamma)^2\,\omega^2\,T}\right).$$

Some algorithms choose stepsizes without prior knowledge of ω (Patil et al., 2023; Samsonov et al., 2024), matching the same information passed to the analyses achieving a robust rate. Nevertheless, in non-asymptotic regimes, the fast guarantee can be much worse than an ω-independent $\widetilde{\mathcal{O}}(1/\sqrt{T})$ guarantee, because ω can be arbitrarily small. Indeed, ignoring logarithmic factors, the fast bound only improves over $\widetilde{\mathcal{O}}(\|\boldsymbol{\theta}^*\|^2/((1-\gamma)\sqrt{T}))$ when $T \gtrsim 1/((1-\gamma)^2\omega^4)$. We now show that one can construct problems where ω is arbitrarily small.

Consider a two-state Markov chain with states s_1 and s_2, and transition matrix

$$\boldsymbol{P} = \begin{bmatrix} \frac{1-\alpha}{2} & \frac{1+\alpha}{2} \\ \frac{1+\alpha}{2} & \frac{1-\alpha}{2} \end{bmatrix}, \qquad \frac{1}{2} < \alpha < 1 \ .$$

Then, this chain is irreducible and aperiodic, and its stationary distribution is uniform: $\boldsymbol{D} = \frac{1}{2}\mathbf{I}$. Now consider the feature matrix

$$\boldsymbol{\Phi} = \begin{bmatrix} \epsilon & 1 \\ -\epsilon & 1 \end{bmatrix}, \qquad 0 < \epsilon < 1 \ .$$

The matrix $\boldsymbol{\Phi}$ is full column rank, and $\boldsymbol{\phi}_\infty = \sqrt{2}$. Moreover, a direct calculation yields

$$\boldsymbol{\Phi}^\top \boldsymbol{D} \boldsymbol{\Phi} = \begin{bmatrix} \epsilon^2 & 0 \\ 0 & 1 \end{bmatrix} \ .$$

Consequently, $\omega = \lambda_{\min}\left(\boldsymbol{\Phi}^\top \boldsymbol{D} \boldsymbol{\Phi}\right) = \epsilon^2$. By choosing ϵ arbitrarily small, the curvature parameter ω can be made arbitrarily close to zero. Importantly, this degeneration arises solely from the feature representation: for different values of the mixing time α, the stationary distribution and ω remain unchanged.

5 Proof Sketch and Differences with Prior Proofs

The fact that the iterates are bounded in this explicit form is both a novel contribution and a crucial ingredient in our analysis. We now explain how this is obtained, contrasting it with previous approaches.

Prior work: Controlling the iterates with curvature. The standard approach to analyze TD(0) starts from the following recursion (Bhandari et al., 2018; Srikant & Ying, 2019; Sun et al., 2021; Patil et al., 2023; Mitra, 2024; Samsonov et al., 2024; Li et al., 2025):

$$\begin{aligned}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 &= \|\boldsymbol{\theta}_{t-1} + \eta_{t-1}\boldsymbol{g}_{t-1} - \boldsymbol{\theta}^*\|^2 = \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + 2\eta_{t-1}\langle \boldsymbol{g}_{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle + \eta_{t-1}^2 \|\boldsymbol{g}_{t-1}\|^2 \\ &= \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + 2\eta_{t-1}\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle + \eta_{t-1}^2 \|\boldsymbol{g}_{t-1}\|^2 + 2\eta_{t-1}\langle \boldsymbol{g}_{t-1} - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle \ .\end{aligned}$$

Then, in the fast regime (Srikant & Ying, 2019; Patil et al., 2023; Samsonov et al., 2024; Li et al., 2025), one can use the following lemma:

Lemma 5.1. *(Mitra, 2024, Lemma 1)*

$$\langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle \leq -\omega(1-\gamma)\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^d.$$

Then, plugging this into the recursion yields

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 \leq (1 - 2\eta_{t-1}\omega(1-\gamma))\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + \eta_{t-1}^2\|\boldsymbol{g}_{t-1}\|^2 + 2\eta_{t-1}\langle \boldsymbol{g}_{t-1} - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle \ .$$

One then proceeds to control the gradient term $\|\boldsymbol{g}_{t-1}\|^2$ and bias term $\langle \boldsymbol{g}_{t-1} - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle$ to form the following standard pseudo-contraction:

$$\mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2\right] = (1 - 2\eta_{t-1}\omega(1-\gamma))\mathbb{E}\left[\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2\right] + \mathcal{O}\left(\eta_{t-1}^2\|\boldsymbol{\theta}^*\|^2\right),$$

Unrolling the recursion yields $\mathbb{E}\left[\left\|\boldsymbol{V}_{\boldsymbol{\theta}^*} - \boldsymbol{V}_{\bar{\boldsymbol{\theta}}_T}\right\|_{\boldsymbol{D}}^2\right] = \widetilde{\mathcal{O}}(\frac{\|\boldsymbol{\theta}^*\|^2}{\omega^2(1-\gamma)^2 T})$. Notice that there is no need for projection steps in this line of analysis, thanks to the contraction property, which keeps the iterates bounded. However, in this approach, the rate depends on ω, which could be arbitrarily small, as we show in Section 4.1.

Prior Work: Controlling the iterates with projections. To avoid the dependency on ω, we can analyze TD(0) using Lemma 4.1 instead of Lemma 5.1. The analysis starts from controlling the magnitude of the gradient $\|\boldsymbol{g}_{t-1}\|$ using the following lemma:

Lemma 5.2. *(Bhandari et al., 2018, Lemma 6) Define* $\boldsymbol{g}_t(\boldsymbol{\theta}) \coloneqq \left(r(s_t, s_{t+1}) + \gamma\boldsymbol{\phi}(s_{t+1})^\top\boldsymbol{\theta} - \boldsymbol{\phi}(s_t)^\top\boldsymbol{\theta}\right)\boldsymbol{\phi}(s_t)$. *Then, for all* $\boldsymbol{\theta} \in \mathbb{R}^d$, $\|\boldsymbol{g}_t(\boldsymbol{\theta})\| \le r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}\|$.

Hence, the term $\|\boldsymbol{\theta}_{t-1}\|$ upper bounds the magnitude of the gradient $\|\boldsymbol{g}_{t-1}\|$. Since $\boldsymbol{\theta}_{t-1} = \boldsymbol{\theta}_{t-2} + \eta_{t-2}\boldsymbol{g}_{t-2}$, the term $\|\boldsymbol{\theta}_{t-1}\|$ in turn depends on both $\|\boldsymbol{\theta}_{t-2}\|$ and $\|\boldsymbol{g}_{t-2}\|$. This recursive dependence can create a vicious cycle, leading to an explosion in $\|\boldsymbol{\theta}_t\|$ driven by the stochasticity of $\boldsymbol{g}_{t-1}$. Previous analyses (Bhandari et al., 2018; Liu & Olshevsky, 2021) avoid this problem by imposing an artificial projection step, which guarantees $\|\boldsymbol{\theta}_t\| \le R$ for all t, where R is chosen agnostically to be larger than $\|\boldsymbol{\theta}^*\|$. Under this constraint, we have a uniform control over the magnitude of $\|\boldsymbol{g}_t\|$ by $r_\infty\phi_\infty + 2\phi_\infty^2 R$ for all t, which can be seen as a bounded gradients condition in optimization.

Our Approach: Controlling the iterates *without* projections. Now, we explain our proof method, which eliminates the need for a projection.

Let's first give our argument in a nutshell. We first ignore the presence of Markovian noise. Next, assuming the iterates are bounded up to time $t-1$, the next TD learning update is also bounded; this allows us to show that the next iterate remains bounded for a carefully chosen stepsize. This implication naturally suggests an inductive proof. For previously ignored Markovian noise, our strategy is to use the geometric convergence in Theorem 3.2. Thus, we work with the stationary update field $\bar{\boldsymbol{g}}$ and control the associated error term. By combining all these steps, we obtain the stated result.

Let's now look at the details. We decompose the updates as follows:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \eta_{t-1}\boldsymbol{g}_{t-1} = \boldsymbol{\theta}_{t-1} + \eta_{t-1}\left(\boldsymbol{g}_{t-1} - \mathbb{E}[\boldsymbol{g}_{t-1} \mid \mathcal{F}_{t-2}]\right) + \eta_{t-1}\left(\mathbb{E}[\boldsymbol{g}_{t-1} \mid \mathcal{F}_{t-2}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}) + \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1})\right).$$

Notice that $\boldsymbol{\xi}_{t-1} \coloneqq \boldsymbol{g}_{t-1} - \mathbb{E}[\boldsymbol{g}_{t-1} \mid \mathcal{F}_{t-2}]$ is a martingale difference sequence with respect to $\mathcal{F}_{t-1} \coloneqq \sigma(s_0, \dots, s_t)$, and $\boldsymbol{b}_{t-1} \coloneqq \mathbb{E}[\boldsymbol{g}_{t-1} \mid \mathcal{F}_{t-2}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1})$ is the gradient bias term. Then, we have

$$\begin{aligned}
\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2 &= \|\boldsymbol{\theta}_{t-1} + \eta_{t-1}(\boldsymbol{\xi}_{t-1} + \boldsymbol{b}_{t-1} + \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1})) - \boldsymbol{\theta}^*\|^2 \\
&= \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + 2\eta_{t-1}\langle\boldsymbol{\xi}_{t-1} + \boldsymbol{b}_{t-1} + \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle \\
&\le \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\|^2 + 2\eta_{t-1}\langle\boldsymbol{\xi}_{t-1} + \boldsymbol{b}_{t-1}, \boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}^*\rangle + 3\eta_{t-1}^2\|\boldsymbol{\xi}_{t-1}\|^2 + 3\eta_{t-1}^2\|\boldsymbol{b}_{t-1}\|^2 + 3\eta_{t-1}^2\|\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{t-1})\|^2,
\end{aligned}$$

where we use Lemma 4.1 in the last inequality. Taking expectation and telescoping gives

$$\mathbb{E}\left[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2\right] \le 2\mathbb{E}\left[\sum_{k=0}^{t-1}\eta_k\langle\boldsymbol{b}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^*\rangle\right] + \|\boldsymbol{\theta}^*\|^2 + 3\mathbb{E}\left[\sum_{k=0}^{t-1}\eta_k^2(\|\boldsymbol{\xi}_k\|^2 + \|\boldsymbol{b}_k\|^2 + \|\bar{\boldsymbol{g}}(\boldsymbol{\theta}_k)\|^2)\right]. \tag{3}$$

Our analysis follows from controlling the gradient bias term $\boldsymbol{b}_k$ in the update. The difficulty in analyzing $\boldsymbol{b}_k$ comes from $\boldsymbol{g}_k$ is not an unbiased estimate of $\bar{\boldsymbol{g}}(\boldsymbol{\theta}_k)$. To be more precise, by defining $Z_k \coloneqq (s_k, s_{k+1})$ and overloading the notation $\boldsymbol{g}(\boldsymbol{\theta}_k, Z_k) \coloneqq \left(r(s_k, s_{k+1}) + \gamma\boldsymbol{\phi}(s_{k+1})^\top\boldsymbol{\theta}_k - \boldsymbol{\phi}(s_k)^\top\boldsymbol{\theta}_k\right)\boldsymbol{\phi}(s_k)$, in general, we have $\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_k, Z_k) \,|\, \mathcal{F}_{k-1}] \ne \bar{\boldsymbol{g}}(\boldsymbol{\theta}_k)$, which comes from the fact that Z_k depends on $Z_{k-1} \in \mathcal{F}_{k-1}$ and the distribution of $E[Z_k \mid Z_{k-1}]$ is not necessarily equal to the stationary distribution.

The key idea for decoupling this dependency is that it converges to the stationary distribution geometrically fast. So, for large k', the distribution of $\mathbb{E}[Z_k \mid Z_{k-k'}]$ can be seen as almost the same as the stationary distribution. This is characterized by the following lemma, whose proof is in Appendix D.

Lemma 5.3. *For any* $0 \le k' \le k$, *let* $\ell_{k-k'} \coloneqq r_\infty\phi_\infty + 2\phi_\infty^2\|\boldsymbol{\theta}_{k-k'}\|$, *then, we have with probability* 1,

$$\|\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_{k-k'}, Z_k) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-k'}) \mid \mathcal{F}_{k-k'-1}]\| \le 2\ell_{k-k'}C\alpha^{k'}.$$

Thus, we can consider the following decomposition of the gradient bias term $\mathbb{E}[\boldsymbol{b}_k]$ in expectation:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{b}_k] = \mathbb{E}[\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_{k-k'}, Z_k) \mid \mathcal{F}_{k-1}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-k'})] \\
+ \mathbb{E}[\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_k, Z_k) \mid \mathcal{F}_{k-1}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_k)] - \mathbb{E}[\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_{k-k'}, Z_k) \mid \mathcal{F}_{k-1}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-k'})].
\end{aligned}$$

In order to apply the Lemma 5.3, we employ the law of total expectation to the first term:

$$\begin{aligned}&\mathbb{E}[\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_{k-k'}, Z_k) \mid \mathcal{F}_{k-1}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-k'})] = \mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_{k-k'}, Z_k)] - \mathbb{E}[\bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-k'})] \\ &= \mathbb{E}[\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}_{k-k'}, Z_k) - \bar{\boldsymbol{g}}(\boldsymbol{\theta}_{k-k'}) \mid \mathcal{F}_{k-k'-1}]] \,. \end{aligned} \tag{4}$$

The remaining term can be controlled by observing $\|\mathbb{E}[\boldsymbol{g}(\boldsymbol{\theta}, Z_t) \mid \mathcal{F}_{t-1}] - \bar{\boldsymbol{g}}(\boldsymbol{\theta})\|$ is Lipschitz in $\boldsymbol{\theta}$ (see Lemma E.5). A comprehensive proof for controlling gradient bias terms for all $k \leq t$ can be found in Appendix F.1.

By applying (4) to bound $\mathbb{E}[\langle \boldsymbol{b}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle]$, we have $\mathbb{E}[\sum_{k=0}^{t-1} \langle \boldsymbol{b}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle]$ is of order $\mathcal{O}(\max_{i \leq t-1} \mathbb{E}[\|\boldsymbol{\theta}_i\|^2] + \|\boldsymbol{\theta}^*\|^2)$, plugging this into equation (3) and controlling other gradient-like terms with Lemma 5.2, we conclude that $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2]$ is also of order $\mathcal{O}(\max_{i \leq t-1} \mathbb{E}[\|\boldsymbol{\theta}_i\|^2] + \|\boldsymbol{\theta}^*\|^2)$. Notice that $\mathbb{E}[\|\boldsymbol{\theta}_t\|^2] \leq \mathbb{E}[(\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\| + \|\boldsymbol{\theta}^*\|)^2]$ from the triangle inequality. That is, to bound $\mathbb{E}[\|\boldsymbol{\theta}_t\|^2]$, we can bound the surrogate target $\mathbb{E}[\|\boldsymbol{\theta}_t - \boldsymbol{\theta}^*\|^2]$ using $\max_{i \leq t-1} \mathbb{E}[\|\boldsymbol{\theta}_i\|^2]$ first. Motivated by this, we consider the induction hypothesis:

$$\max_{i \leq t-1} \; \mathbb{E}\left[\|\boldsymbol{\theta}_i\|^2\right] \leq \rho_c^2 \max\left\{\frac{r_\infty^2}{\phi_\infty^2}, \|\boldsymbol{\theta}^*\|^2\right\}$$

and prove that $\max_{i \leq t} \mathbb{E}[\|\boldsymbol{\theta}_i\|^2] \leq \rho_c^2 \max\left\{\frac{r_\infty^2}{\phi_\infty^2}, \|\boldsymbol{\theta}^*\|^2\right\}$. Indeed, if the stepsize parameter c is small enough, a constant that bounds the previous iterates will also bound the next iterate. We remark again that the stepsize parameter c and ρ_c in induction are chosen carefully to ensure the inductive step proceeds. The precise derivation can be found in Theorem G.1 in the Appendix.

Convergence result. We now give the proof of Theorem 4.2 (b).

Proof. For any $0 \leq t \leq T-1$, let $d_t = \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t\|$. Thus,

$$\begin{aligned} d_{t+1}^2 &= \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_t\|^2 = d_t^2 - 2\eta_t \langle \boldsymbol{g}_t, \boldsymbol{\theta}^* - \boldsymbol{\theta}_t \rangle + \eta_t^2 \|\boldsymbol{g}_t\|^2 \\ &= d_t^2 - 2\eta_t \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_t), \boldsymbol{\theta}^* - \boldsymbol{\theta}_t \rangle + 2\eta_t \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_t) - \boldsymbol{g}_t, \boldsymbol{\theta}^* - \boldsymbol{\theta}_t \rangle + \eta_t^2 \|\boldsymbol{g}_t\|^2 \;. \end{aligned}$$

Summing from $t = 0$ to $t = T-1$, taking the expectation, and using Lemma 4.1, we have

$$\begin{aligned} &\sum_{t=0}^{T-1} 2\eta_t \mathbb{E}\Big[(1-\gamma) \|\boldsymbol{V}_{\boldsymbol{\theta}_t} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2 + \gamma \|\boldsymbol{V}_{\boldsymbol{\theta}_t} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\mathrm{Dir}}^2\Big] \\ &\quad \leq \sum_{t=0}^{T-1} \left(\mathbb{E}\big[d_t^2\big] - \mathbb{E}\big[d_{t+1}^2\big]\right) + \mathbb{E}\left[\sum_{t=0}^{T-1} \eta_t^2 \|\boldsymbol{g}_t\|^2\right] + \mathbb{E}\left[\sum_{t=0}^{T-1} 2\eta_t \langle \bar{\boldsymbol{g}}(\boldsymbol{\theta}_t) - \boldsymbol{g}_t, \boldsymbol{\theta}^* - \boldsymbol{\theta}_t \rangle\right] \\ &\quad = \|\boldsymbol{\theta}^*\|^2 + \mathcal{O}\left(\rho_c^2 \max\left\{\frac{r_\infty^2}{\phi_\infty^2}, \|\boldsymbol{\theta}^*\|^2\right\}\right), \end{aligned}$$

where we used Theorem G.1 in the last inequality. Using the convexity of f, we have

$$\begin{aligned} \mathbb{E}\left[f(\bar{\boldsymbol{\theta}}_T) - f(\boldsymbol{\theta}^*)\right] &\leq \frac{1}{\sum_{i=0}^{T-1} \eta_i} \sum_{t=0}^{T-1} \eta_t \mathbb{E}\Big[(1-\gamma) \|\boldsymbol{V}_{\boldsymbol{\theta}_t} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\boldsymbol{D}}^2\Big] + \frac{1}{\sum_{i=0}^{T-1} \eta_i} \sum_{t=0}^{T-1} \eta_t \mathbb{E}\Big[\gamma \|\boldsymbol{V}_{\boldsymbol{\theta}_t} - \boldsymbol{V}_{\boldsymbol{\theta}^*}\|_{\mathrm{Dir}}^2\Big] \\ &= \widetilde{\mathcal{O}}\left(\frac{c\rho_c^2 \max\left\{r_\infty^2, \phi_\infty^2 \|\boldsymbol{\theta}^*\|^2\right\}}{\sqrt{T}}\right), \end{aligned}$$

where we use $\sum_{i=0}^{T-1} \eta_i \geq \frac{2\sqrt{T+1}-2}{c\phi_\infty^2 \log^2(T+3)}$ in the last equality. □

6 Conclusion

In this paper, we present a robust finite-time analysis of TD(0) without requiring additional projection steps. To the best of our knowledge, this is the first finite-time guarantee in this setting. In particular, we do not employ the contraction-based proof technique used in previous work; instead, we directly prove that the

iterates of TD(0) are bounded. We believe our proof is general and, for example, it can be easily extended to analyze the TD(λ) algorithm and the Q-learning setting introduced in (Chen et al., 2022).

In future work, we plan to investigate the possibility of obtaining rates that interpolate between $\widetilde{\mathcal{O}}(1/\sqrt{T})$ and $\widetilde{\mathcal{O}}(1/T)$, depending on the curvature of the potential function. Ideally, one would like to show that TD(0) adapts to the curvature of the function with a specific stepsize, as is possible with recent parameter-free schemes (Cutkosky & Orabona, 2018).

References

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

Chenyi Chen, Ari Seff, Alain Kornhauser, and Jianxiong Xiao. Deepdriving: Learning affordance for direct perception in autonomous driving. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2722–2730, 2015.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proc. of the Conference on Learning Theory (COLT)*, 2018.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pp. 1199–1233. PMLR, 2018.

Persi Diaconis and Laurent Saloff-Coste. Logarithmic Sobolev inequalities for finite Markov chains. *The Annals of Applied Probability*, 6(3):695–750, 1996.

Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine. Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3389–3396. IEEE, 2017.

Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pp. 14465–14499. PMLR, 2023.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Nathaniel Korda and Prashanth La. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International Conference on Machine Learning*, pp. 626–634. PMLR, 2015.

Harold Kushner. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):87–96, 2010.

Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pp. 1347–1355. PMLR, 2018.

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

Yunxiang Li, Mark Schmidt, Reza Babanezhad, and Sharan Vaswani. Towards parameter-free temporal difference learning. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025. URL `https://openreview.net/forum?id=BKcraYdfVd`.

Rui Liu and Alex Olshevsky. Temporal difference learning as gradient splitting. In *International Conference on Machine Learning*, pp. 6905–6913. PMLR, 2021.

Shie Mannor, Yishay Mansour, and Aviv Tamar. *Reinforcement Learning: Foundations*. -, 2022. URL `https://sites.google.com/view/rlfoundations/home`.

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

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

Yann Ollivier. Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*, 2018.

Francesco Orabona and David Pál. Parameter-free stochastic optimization of variationally coherent functions. *arXiv preprint arXiv:2102.00236*, 2021.

Gandharv Patil, L. A. Prashanth, Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, pp. 5438–5448. PMLR, 2023.

Sergey Samsonov, Daniil Tiapkin, Alexey Naumov, and Eric Moulines. Improved high-probability bounds for the temporal difference learning algorithm via exponential stability. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 4511–4547. PMLR, 2024.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

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

Tao Sun, Han Shen, Tianyi Chen, and Dongsheng Li. Adaptive temporal difference learning with linear function approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8812–8824, 2021.

Tao Sun, Dongsheng Li, and Bao Wang. Finite-time analysis of adaptive temporal difference learning with deep neural networks. *Advances in Neural Information Processing Systems*, 35:19592–19604, 2022.

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

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*, volume 1. MIT Press, 1998.

Matus Telgarsky. Stochastic linear optimization never overfits with quadratically-bounded losses on general data. In *Conference on Learning Theory*, pp. 5453–5488. PMLR, 2022.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in Neural Information Processing Systems*, 9, 1996.

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.