

# A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation

Jalaj Bhandari, Daniel Russo and Raghav Singal

Columbia University

## Abstract

Temporal difference learning (TD) is a simple iterative algorithm used to estimate the value function corresponding to a given policy in a Markov decision process. Although TD is one of the most widely used algorithms in reinforcement learning, its theoretical analysis has proved challenging and few guarantees on its statistical efficiency are available. In this work, we provide a *simple and explicit finite time analysis* of temporal difference learning with linear function approximation. Except for a few key insights, our analysis mirrors standard techniques for analyzing stochastic gradient descent algorithms, and therefore inherits the simplicity and elegance of that literature. Final sections of the paper show how all of our main results extend to the study of TD learning with eligibility traces, known as TD( $\lambda$ ), and to Q-learning applied in high-dimensional optimal stopping problems.

**Keywords:** Reinforcement learning, temporal difference learning, finite time analysis, stochastic gradient descent.

## 1 Introduction

Originally proposed by Sutton [1988], temporal difference learning (TD) is one of the most widely used reinforcement learning algorithms and a foundational idea on which more complex methods are built. The algorithm operates on a stream of data generated by applying some policy to a poorly understood Markov decision process. The goal is to learn an approximate value function, which can then be used to track the net present value of future rewards as a function of the system’s evolving state. TD maintains a parametric approximation to the value function, making a simple incremental update to the estimated parameter vector each time a state transition occurs.

While easy to implement, theoretical analysis of TD is subtle. Reinforcement learning researchers in the 1990s gathered both limited convergence guarantees [Jaakkola et al., 1994] and examples of divergence [Baird, 1995]. Many issues were then clarified in the work of Tsitsiklis and Van Roy [1997], which establishes precise conditions for the asymptotic convergence of TD with linear function approximation and gives examples of divergent behavior when key conditions are violated. With guarantees of asymptotic convergence in place, a natural next step is to understand the algorithm’s statistical efficiency. How much data is required to guarantee a given level of accuracy? Can one give uniform bounds on this, or could data requirements explode depending on the problem instance? Twenty years after the work of Tsitsiklis and Van Roy [1997], such questions remain largely unsettled.

### 1.1 Contributions

This paper develops a *simple and explicit non-asymptotic analysis of TD with linear function approximation*. The resulting guarantees provide assurances of robustness. They explicitly bound the worst-case dependence on problem features like the discount factor, the conditioning of the feature covariance matrix, and the mixing time of the underlying Markov chain. Our analysis reveals rigorous connections between TD and stochastic gradient descent algorithms, provides a template for finite time analysis of incremental algorithms with Markovian noise, and applies without modification to analyzing a class of high-dimensional optimal stopping problems. We elaborate on these contributions below.

- *Links with gradient descent:* Despite a cosmetic connection to stochastic gradient descent (SGD), incremental updates of TD are not (stochastic) gradient steps with respect to any fixed loss function. It is therefore difficult to show that it makes consistent, quantifiable, progress toward its asymptotic limit point. Nevertheless, Section 6 shows that expected TD updates obey crucial properties mirroring those of gradient descent on a particular quadratic loss function. In a model where the observations are corrupted by i.i.d. noise, these gradient-like properties of TD allow us to give state-of-the-art convergence bounds by essentially mirroring standard analyses of stochastic gradient descent (SGD). This approach may be of broader interest as SGD analyses are commonly taught in machine learning courses and serve as a launching point for a much broader literature on first-order optimization. Rigorous connections with the optimization literature can facilitate research on principled improvements to TD.
- *Non-asymptotic treatment with Markovian noise:* TD is usually applied online to a single Markovian data stream. However, to our knowledge, there has been no successful<sup>1</sup> non-asymptotic analysis in the setting with Markovian observation noise. Instead, many papers have studied such algorithms under the simpler i.i.d noise model mentioned earlier [Sutton et al., 2009b,a, Liu et al., 2015, Touati et al., 2018, Dalal et al., 2018b, Lakshminarayanan and Szepesvári, 2018]. One reason is that the dependent nature of the data introduces a substantial technical challenge: the algorithm’s updates are not only noisy, but can be severely biased. We use information theoretic techniques to control the magnitude of bias, yielding bounds that are essentially scaled by a factor of the mixing time of the underlying Markov process relative to those attained for i.i.d. model. Our analysis in this setting applies only to a variant of TD that projects the iterates onto a norm ball. This projection step imposes a uniform bound on the noise of gradient updates, which is needed for tractability. For similar reasons, projection operators are widely used throughout the stochastic approximation literature [Kushner, 2010, Section 2].
- *An extendable approach:* Much of the paper focuses on analyzing the most basic temporal difference learning algorithm, known as TD(0). We also extend this analysis to other algorithms. First, we establish convergence bounds for temporal difference learning with eligibility traces, known as TD( $\lambda$ ). This is known to often outperform TD(0) [Sutton and Barto, 1998], but a finite time analysis is more involved. Our analysis also applies without modification to Q-learning for a class of high-dimensional optimal stopping problems. Such problems have been widely studied due to applications in the pricing of financial derivatives [Tsitsiklis and Van Roy, 1999, Andersen and Broadie, 2004, Haugh and Kogan, 2004, Desai et al., 2012, Goldberg and Chen, 2018]. For our purposes, this example illustrates more clearly the link between value prediction and decision-making. It also shows our techniques extend seamlessly to analyzing an instance of non-linear stochastic approximation. To our knowledge, no prior work has provided non-asymptotic guarantees for either TD( $\lambda$ ) or Q-learning with function approximation.

## 1.2 Related Literature

**Non-asymptotic analysis of TD(0):** There has been very little non-asymptotic analysis of TD(0). To our knowledge, Korda and La [2015] provided the first finite time analysis. However, several serious errors in their proofs were pointed out by Lakshminarayanan and Szepesvári [2017]. A very recent work by Dalal et al. [2018a] studies TD(0) with linear function approximation in an i.i.d. observation model, which assumes sequential observations used by the algorithm are drawn independently from their steady-state distribution. They focus on analysis with problem independent step-sizes of the form  $1/T^\sigma$  for a fixed  $\sigma \in (0, 1)$  and establish that mean-squared error converges at a rate<sup>2</sup> of  $O(1/T^\sigma)$ . Unfortunately, while the analysis is technically non-asymptotic, the constant factors in the bound display a complex dependence on the problem instance and even scale exponentially with the eigenvalues of certain matrices. Dalal et al. [2018a] also give a high-probability bound, a nice feature that we do not address in this work.

---

<sup>1</sup>This was previously attempted by Korda and La [2015], but critical errors were shown by Lakshminarayanan and Szepesvári [2017].

<sup>2</sup>In personal communication, the authors have told us their analysis also yields a  $O(1/T)$  rate of convergence for problem dependent step-sizes, though we have not been able to easily verify this.

This paper was accepted at the 2018 Conference on Learning Theory (COLT) and published in the proceedings as a two-page extended abstract. While the paper was under review, an interesting paper by [Lakshminarayanan and Szepesvári \[2018\]](#) appeared. They study linear stochastic approximation algorithms under i.i.d noise, including TD(0), with constant step-sizes and iterate averaging. This line of work dates back to [Györfi and Walk \[1996\]](#), who show that the iterates of a constant step-size linear stochastic approximation algorithm form an ergodic Markov chain and, *in the case of i.i.d. observation noise*, their expectation in steady-state is equal to the true solution of the linear system. By a central limit theorem for ergodic sequences, the average iterate converges to the true solution, with mean-squared error decaying at rate  $O(1/T)$ . [Bach and Moulines \[2013\]](#) give a sophisticated non-asymptotic analysis of the least-mean-squares algorithm with constant step-size and iterate-averaging. [Lakshminarayanan and Szepesvári \[2018\]](#) aim to understand whether such guarantees extend to linear stochastic approximation algorithms more broadly. In the process, their work provides  $O(1/T)$  bounds for iterate-averaged TD(0) with constant step-size. A remarkable feature of their approach is that the choice of step-size is independent of the conditioning of the features (although the bounds themselves do degrade if features become ill-conditioned). It is worth noting that these results rely critically on the assumption that noise is i.i.d. In fact, [Györfi and Walk \[1996\]](#) provide a very simple example of failure under correlated noise. In this example, under a linear stochastic approximation algorithm applied with any constant step-size, the averaged-iterate will converge to the wrong limit.

The recent works of [Dalal et al. \[2018a\]](#) and [Lakshminarayanan and Szepesvári \[2018\]](#) give bounds for TD(0) only under i.i.d. observation noise. Therefore their results are most comparable to what is presented in Section 7. For the i.i.d. noise model, the main argument in favor of our approach is that it allows for extremely simple proofs, interpretable constant terms, and illuminating connections with SGD. Moreover, it is worth emphasizing that our approach gracefully extends to more complex settings, including more realistic models with Markovian noise, the analysis of TD with eligibility traces, and the analysis of Q-learning for optimal stopping problems as shown in Sections 8, 9 and 10.

While not directly comparable to our results, we point the readers to the excellent work of [Schapire and Warmuth \[1996\]](#). To facilitate theoretical analysis, they consider a slightly modified version of the TD( $\lambda$ ). The authors provide a finite time analysis for this algorithm in an adversarial model where the goal is to predict the discounted sum of future rewards from each state. Performance is measured relative to the best fixed linear predictor in hindsight. The analysis is creative, but results depend on several unknown constants and on the specific sequence of states and rewards on which the algorithm is applied. [Schapire and Warmuth \[1996\]](#) also apply their techniques to study value function approximation in a Markov decision process. In that case, the bounds are much weaker than what is established here. Their bound scales with the size of the state space—which is enormous in most practical problems—and applies only to TD(1)—a somewhat degenerate special case of TD( $\lambda$ ) in which it is equivalent to Monte Carlo policy evaluation [[Sutton and Barto, 1998](#)].

**Asymptotic analysis of stochastic approximation:** There is a well developed asymptotic theory of stochastic approximation, a field that studies noisy recursive algorithms like TD [[Kushner and Yin, 2003](#), [Borkar, 2009](#), [Benveniste et al., 2012](#)]. Most asymptotic convergence proofs in reinforcement learning use a technique known as the ODE method [[Borkar and Meyn, 2000](#)]. Under some technical conditions and appropriate decaying step-sizes, this method ensures the almost-sure convergence of stochastic approximation algorithms to the invariant set of a certain ‘mean’ differential equation. The technique greatly simplifies asymptotic convergence arguments, since it completely circumvents issues with noise in the system and issues of step-size selection. But this also makes it a somewhat coarse tool, unable to generate insight into an algorithm’s sensitivity to noise, ill-conditioning, or step-size choices. A more refined set of techniques begin to address these issues. Under fairly broad conditions, a central limit theorem for stochastic approximation algorithms characterizes their limiting variance. Such a central limit theorem has been specifically provided for TD by [Konda \[2002\]](#) and [Devraj and Meyn \[2017\]](#).

Despite the availability of such asymptotic techniques, the modern literature on first-order stochastic optimization focuses heavily on non-asymptotic analysis [[Bottou et al., 2018](#), [Bubeck, 2015](#), [Jain and Kar, 2017](#)]. One reason is that such asymptotic analysis necessarily focuses on a regime where step-sizes are arbitrarily small relative to problem features and the iterates have already converged to a small neighborhood

of the optimum. However, the use of a first-order method in the first place signals that a practitioner is mostly interesting in cheaply reaching a reasonably accurate solution, rather than the rate of convergence in the neighborhood of the optimum. In practice, it is common to use constant step-sizes, so iterates never truly converge to the optimum. A non-asymptotic analysis requires grappling with the algorithm’s behavior in practically relevant regimes where step-sizes are still relatively large and iterates are not yet close to the true solution.

**Analysis of related algorithms:** A number of papers analyze algorithms related to and inspired by the classic TD algorithm. First, among others, [Antos et al. \[2008\]](#), [Lazaric et al. \[2010\]](#), [Ghavamzadeh et al. \[2010\]](#), [Pires and Szepesvari \[2012\]](#), [Prashanth et al. \[2013\]](#) and [Tu and Recht \[2018\]](#) analyze least-squares temporal difference learning (LSTD). [Yu and Bertsekas \[2009\]](#) study the related least-squares policy iteration algorithm. The asymptotic limit point of TD is a minimizer of a certain population loss, known as the mean-squared projected Bellman error. LSTD solves a least-squares problem, essentially computing the exact minimizer of this loss on the empirical data. It is easy to derive a central limit theorem for LSTD. Finite time bounds follow from establishing uniform convergence rates of the empirical loss to the population loss. Unfortunately, such techniques appear to be quite distinct from those needed to understand the online TD algorithms studied in this paper. Online TD has seen much wider use due to significant computational advantages [[Sutton and Barto, 1998](#)].

Gradient TD methods are another related class of algorithms. These were derived by [Sutton et al. \[2009b,a\]](#) to address the issue that TD can diverge in so-called off-policy settings, where data is collected from a policy different from the one for which we want to estimate the value function. Unlike the classic TD(0) algorithm, gradient TD methods are designed to mimic gradient descent with respect to the mean squared projected Bellman error. [Sutton et al. \[2009b,a\]](#) propose asymptotically convergent two-time scale stochastic approximation schemes based on this and more recently [Dalal et al. \[2018b\]](#) give a finite time analysis of two time scale stochastic approximation algorithms, including several variants of gradient TD algorithms. A creative paper by [Liu et al. \[2015\]](#) reformulates the original optimization as a primal-dual saddle point problem and leverages convergence analysis form that literature to give a non-asymptotic analysis. This work was later revisited by [Touati et al. \[2018\]](#), who established a faster rate of convergence. The works of [Dalal et al. \[2018b\]](#), [Liu et al. \[2015\]](#) and [Touati et al. \[2018\]](#) all consider only i.i.d. observation noise. One interesting open question is whether our techniques for treating the Markovian observation model will also apply to these analyses. Finally, it is worth highlighting that, to the best of our knowledge, substantial new techniques are needed to analyze the widely used TD(0), TD( $\lambda$ ) and the Q-learning algorithm studied in this paper. Unlike gradient TD methods, they do not mimic noisy gradient steps with respect to any fixed objective<sup>3</sup>.

## 2 Problem formulation

**Markov reward process.** We consider the problem of evaluating the value function  $V_\mu$  of a given policy  $\mu$  in a Markov decision process (MDP). We work in the *on policy* setting, where data is generated by applying the policy  $\mu$  in the MDP. Because the policy  $\mu$  is applied automatically to select actions, such problems are most naturally formulated as value function estimation in a Markov reward process (MRP). A MRP<sup>4</sup> comprises of  $(\mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma)$  [[Sutton and Barto, 1998](#)] where  $\mathcal{S}$  is the set of states,  $\mathcal{P}$  is the Markovian transition kernel,  $\mathcal{R}$  is a reward function, and  $\gamma < 1$  is the discount factor. For a discrete state-space  $\mathcal{S}$ ,  $\mathcal{P}(s'|s)$  specifies the probability of transitioning from a state  $s$  to another state  $s'$ . The reward function  $\mathcal{R}(s, s')$  associates a reward with each state transition. We denote by  $\mathcal{R}(s) = \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s) \mathcal{R}(s, s')$  the expected instantaneous reward generated from an initial state  $s$ .

---

<sup>3</sup>This can be formally verified for TD(0) with linear function approximation. If the TD step were a gradient with respect to a fixed objective, differentiating it should give the Hessian and hence a symmetric matrix. Instead, the matrix one attains is typically not a symmetric one.

<sup>4</sup>We avoid  $\mu$  from notation for simplicity.

The value function associated with this MRP,  $V_\mu$ , specifies the expected cumulative discounted future reward as a function of the state of the system. In particular,

$$V_\mu(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t) \mid s_0 = s \right],$$

where the expectation is over sequences of states generated according to the transition kernel  $\mathcal{P}$ . This value function obeys the Bellman equation  $T_\mu V_\mu = V_\mu$ , where the Bellman operator  $T_\mu$  associates a value function  $V : \mathcal{S} \rightarrow \mathbb{R}$  with another value function  $T_\mu V$  satisfying

$$(T_\mu V)(s) = \mathcal{R}(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s) V(s') \quad \forall s \in \mathcal{S}.$$

We assume rewards are bounded uniformly such that

$$|\mathcal{R}(s, s')| \leq r_{\max} \quad \forall s, s' \in \mathcal{S}.$$

Under this assumption, value functions are assured to exist and are the unique solution to Bellman's equation [Bertsekas, 2012]. We also assume that the Markov reward process induced by following the policy  $\mu$  is ergodic with a unique stationary distribution  $\pi$ . For any two states  $s, s'$ :  $\pi(s') = \lim_{t \rightarrow \infty} \mathbb{P}(s_t = s' \mid s_0 = s)$ .

Following common references [Bertsekas, 2012, Dann et al., 2014, De Farias and Van Roy, 2003], we will simplify the presentation by assuming the state space  $\mathcal{S}$  is a finite set of size  $n = |\mathcal{S}|$ . We elaborate on this choice in the remark below.

**Remark 1.** *Working with a finite state space allows for the use of compact matrix notation, which is the convention in work on linear value function approximation. It also avoids measure theoretic notation for conditional probability distributions. Our proofs extend in an obvious way to problems with countably infinite state-spaces. For problems with general state-space, even the core results in dynamic programming hold only under suitable technical conditions [Bertsekas and Shreve, 1978].*

**Value function approximation.** Given a fixed policy  $\mu$ , the problem is to efficiently estimate the corresponding value function  $V_\mu$  using only the observed rewards and state transitions. Unfortunately, due to the curse of dimensionality, most modern applications have intractably large state spaces, rendering exact value function learning hopeless. Instead, researchers resort to parametric approximations of the value function, for example by using a linear function approximator [Sutton and Barto, 1998] or a non-linear function approximation such as a neural network [Mnih et al., 2015]. In this work, we consider a linear function approximation architecture where the true value-to-go  $V_\mu(s)$  is approximated as

$$V_\mu(s) \approx V_\theta(s) = \phi(s)^\top \theta,$$

where  $\phi(s) \in \mathbb{R}^d$  is a fixed feature vector for state  $s$  and  $\theta \in \mathbb{R}^d$  is a parameter vector that is shared across states. When the state space is the finite set  $\mathcal{S} = \{s_1, \dots, s_n\}$ ,  $V_\theta \in \mathbb{R}^n$  can be expressed compactly as

$$V_\theta = \begin{bmatrix} \phi(s_1)^\top \\ \vdots \\ \phi(s_n)^\top \end{bmatrix} \theta = \begin{bmatrix} \phi_1(s_1) & \phi_k(s_1) & \phi_d(s_1) \\ \vdots & \vdots & \vdots \\ \phi_1(s_n) & \phi_k(s_n) & \phi_d(s_n) \end{bmatrix} \theta = \Phi \theta,$$

where  $\Phi \in \mathbb{R}^{n \times d}$  and  $\theta \in \mathbb{R}^d$ . We assume throughout that the  $d$  features vectors  $\{\phi_k\}_{k=1}^d$ , forming the columns of  $\Phi$  are linearly independent.

**Norms in value function and parameter space.** For a symmetric positive definite matrix  $A$ , define the inner product  $\langle x, y \rangle_A = x^\top A y$  and the associated norm  $\|x\|_A = \sqrt{x^\top A x}$ . If  $A$  is positive semi-definite rather than positive definite then  $\|\cdot\|_A$  is called a semi-norm. Let  $D = \text{diag}(\pi(s_1), \dots, \pi(s_n)) \in \mathbb{R}^{n \times n}$  denote the diagonal matrix whose elements are given by the entries of the stationary distribution  $\pi(\cdot)$ . Then, for two value functions  $V$  and  $V'$ ,

$$\|V - V'\|_D = \sqrt{\sum_{s \in \mathcal{S}} \pi(s) (V(s) - V'(s))^2},$$

measures the mean-square difference between the value predictions under  $V$  and  $V'$ , in steady-state. This suggests a natural norm on the space of parameter vectors. In particular, for any  $\theta, \theta' \in \mathbb{R}^d$ ,

$$\|V_\theta - V_{\theta'}\|_D = \sqrt{\sum_{s \in \mathcal{S}} \pi(s) (\phi(s)^\top (\theta - \theta'))^2} = \|\theta - \theta'\|_\Sigma$$

where

$$\Sigma := \Phi^\top D \Phi = \sum_{s \in \mathcal{S}} \pi(s) \phi(s) \phi(s)^\top$$

is the steady-state feature covariance matrix.

**Feature regularity.** We assume that any entirely redundant or irrelevant features have been removed, so  $\Sigma$  has full rank. Additionally, we also assume that  $\|\phi(s)\|_2^2 \leq 1$  for all  $s \in \mathcal{S}$ , which can be ensured through feature normalization. This also ensures that  $\Sigma$  exists. Let  $\omega > 0$  be the minimum eigenvalue of  $\Sigma$ . From our bound on the feature vectors, the maximum eigenvalue of  $\Sigma$  is less than  $1^5$ , so  $1/\omega$  bounds the condition number of the feature covariance matrix. The following lemma is an immediate consequence of our assumptions.

**Lemma 1** (Norm equivalence). *For all  $\theta \in \mathbb{R}^d$ ,  $\sqrt{\omega} \|\theta\|_2 \leq \|V_\theta\|_D \leq \|\theta\|_2$ .*

While we assume it has full rank, we will establish some finite time bounds that are independent of the conditioning of the feature covariance matrix.

### 3 Temporal difference learning

We consider the classic temporal difference learning algorithm [Sutton, 1988]. The algorithm starts with an initial parameter estimate  $\theta_0$  and at every time step  $t$ , it observes one data tuple  $O_t = (s_t, r_t = \mathcal{R}(s_t, s'_t), s'_t)$  consisting of the current state, the current reward and the next state reached by playing policy  $\mu$  in the current state. This tuple is used to define a loss function, which is taken to be the squared sample Bellman error. It then proceeds to compute the next iterate  $\theta_{t+1}$  by taking a gradient step. Some of our bounds guarantee accuracy of the average iterate, denoted by  $\bar{\theta}_t = t^{-1} \sum_{i=0}^{t-1} \theta_i$ . The version of TD presented in Algorithm 1 also makes online updates to the averaged iterate.

We present in Algorithm 1 the simplest variant of TD, which is known as TD(0). It is also worth highlighting that here we study online temporal difference learning, which makes incremental gradient-like updates to the parameter estimate based on the most recent data observations only. Such algorithms are widely used in practice, but harder to analyze than so-called batch TD methods like the LSTD algorithm of Bradtko and Barto [1996].

At time  $t$ , TD takes a step in the direction of the negative gradient  $g_t(\theta_t)$  evaluated at the current parameter. As a general function of  $\theta$  and the tuple  $O_t = (s_t, r_t, s'_t)$ , the negative gradient can be written as

$$g_t(\theta) = \left( r_t + \gamma \phi(s'_t)^\top \theta - \phi(s_t)^\top \theta \right) \phi(s_t). \quad (1)$$

---

<sup>5</sup>Let  $\lambda_{\max}(A) = \max_{\|x\|_2=1} x^\top A x$  denote the maximum eigenvalue of a symmetric positive-semidefinite matrix. Since this is a convex function,  $\lambda_{\max}(\Sigma) \leq \sum_{s \in \mathcal{S}} \pi(s) \lambda_{\max}(\phi(s) \phi(s)^\top) \leq \sum_{s \in \mathcal{S}} \pi(s) = 1$ .

---

**Algorithm 1:** TD(0) with linear function approximation

---

**Input :** initial guess  $\theta_0$ , step-size sequence  $\{\alpha_t\}_{t \in \mathbb{N}}$ .  
 Initialize:  $\bar{\theta}_0 \leftarrow \theta_0$ .  
**for**  $t = 0, 1, \dots$  **do**

Observe tuple:  $O_t = (s_t, r_t = \mathcal{R}(s_t, s'_t), s'_t)$   
 Define target:  $y_t = \mathcal{R}(s_t, s'_t) + \gamma V_{\theta_t}(s'_t)$  /\* sample Bellman operator \*/  
 Define loss function:  $\frac{1}{2}(y_t - V_{\theta_t}(s_t))^2$  /\* sample Bellman error squared \*/  
 Compute negative gradient:  $g_t(\theta_t) = -\frac{\partial}{\partial \theta} \frac{1}{2}(y_t - V_{\theta_t}(s_t))^2|_{\theta=\theta_t}$   
 Take a gradient step:  $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$  /\*  $\alpha_t$ : step-size \*/  
 Update averaged iterate:  $\bar{\theta}_{t+1} \leftarrow \left(\frac{t}{t+1}\right) \bar{\theta}_t + \left(\frac{1}{t+1}\right) \theta_{t+1}$  /\*  $\bar{\theta}_{t+1} = \frac{1}{t+1} \sum_{\ell=0}^t \theta_\ell$  \*/  
**end**

---

The long-run dynamics of TD are closely linked to the expected negative gradient step when the tuple  $O_t = (s_t, r_t, s'_t)$  follows its *steady-state* behavior:

$$\bar{g}(\theta) := \sum_{s, s' \in \mathcal{S}} \pi(s) \mathcal{P}(s'|s) (\mathcal{R}(s, s') + \gamma \phi(s')^\top \theta - \phi(s)^\top \theta) \phi(s) \quad \forall \theta \in \mathbb{R}^d.$$

This can be rewritten more compactly in several useful ways. One such way is,

$$\bar{g}(\theta) = \mathbb{E}[\phi r] + \mathbb{E}[\phi(\gamma \phi' - \phi)^\top] \theta, \quad (2)$$

where  $\phi = \phi(s)$  is the feature vector of a random initial state  $s \sim \pi$ ,  $\phi' = \phi(s')$  is the feature vector of a random next state drawn according to  $s' \sim \mathcal{P}(\cdot | s)$ , and  $r = \mathcal{R}(s, s')$ . In addition, since  $\sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s) (\mathcal{R}(s, s') + \gamma \phi(s')^\top \theta) = (T_\mu \Phi \theta)(s)$ , we can recognize that

$$\bar{g}(\theta) = \Phi^\top D(T_\mu \Phi \theta - \Phi \theta). \quad (3)$$

See [Tsitsiklis and Van Roy \[1997\]](#) for a derivation of this fact.

## 4 Asymptotic convergence of temporal difference learning

The main challenge in analyzing TD is that the gradient steps  $g_t(\theta)$  are not true stochastic gradients with respect to any fixed objective. The gradient step taken at time  $t$  pulls the value prediction  $V_{\theta_{t+1}}(s_t)$  closer to  $y_t$ , but  $y_t$  itself depends on  $V_{\theta_t}$ . So does this circular process converge? The key insight of [Tsitsiklis and Van Roy \[1997\]](#) was to interpret this as a stochastic approximation scheme for solving a fixed point equation known as the projected Bellman equation. Contraction properties together with general results from stochastic approximation theory can then be used to show convergence.

Should TD converge at all, it should be to a stationary point. Because the feature covariance matrix  $\Sigma$  is full rank there is a unique<sup>6</sup> vector  $\theta^*$  with  $\bar{g}(\theta^*) = 0$ . We briefly review results that offer insight into  $\theta^*$  and proofs of the asymptotic convergence of TD.

**Understanding the TD limit point.** [Tsitsiklis and Van Roy \[1997\]](#) give an interesting characterization of the limit point  $\theta^*$ . They show it is the unique solution to the *projected* Bellman equation

$$\Phi \theta = \Pi_D T_\mu \Phi \theta, \quad (4)$$

---

<sup>6</sup>This follows formally as a consequence of Lemma 3 in this paper.

where  $\Pi_D(\cdot)$  is the projection operator onto the subspace  $\{\Phi x \mid x \in \mathbb{R}^d\}$  spanned by these features in the inner product  $\langle \cdot, \cdot \rangle_D$ . To see why this is the case, note that by using  $\bar{g}(\theta^*) = 0$  along with Equation (3),

$$0 = x^\top \bar{g}(\theta^*) = \langle \Phi x, T_\mu \Phi \theta^* - \Phi \theta^* \rangle_D \quad \forall x \in \mathbb{R}^d.$$

That is, the Bellman error at  $\theta^*$ , given by  $(T_\mu \Phi \theta^* - \Phi \theta^*)$ , is orthogonal to the space spanned by the features in the inner product  $\langle \cdot, \cdot \rangle_D$ . By definition, this means  $\Pi_D(T_\mu \Phi \theta^* - \Phi \theta^*) = 0$  and hence  $\theta^*$  must satisfy the projected Bellman equation.

The following lemma shows the projected Bellman operator,  $\Pi_D T_\mu(\cdot)$  is a contraction, and so in principle, one could converge to the approximate value function  $\Phi \theta^*$  by repeatedly applying it. TD appears to serve a simple stochastic approximation scheme for solving the projected-Bellman fixed point equation.

**Lemma 2.** [[Tsitsiklis and Van Roy \[1997\]](#)]  $\Pi_D T_\mu(\cdot)$  is a contraction with respect to  $\|\cdot\|_D$  with modulus  $\gamma$ , that is,

$$\|\Pi_D T_\mu V_\theta - \Pi_D T_\mu V_{\theta'}\|_D \leq \gamma \|V_\theta - V_{\theta'}\|_D \quad \forall \theta, \theta' \in \mathbb{R}^d.$$

Finally, the limit of convergence comes with some competitive guarantees. From Lemma 2, a short argument shows

$$\|V_{\theta^*} - V_\mu\|_D \leq \frac{1}{\sqrt{1-\gamma^2}} \|\Pi_D V_\mu - V_\mu\|_D. \quad (5)$$

See Chapter 6 of [Bertsekas \[2012\]](#) for a proof. The left hand side of Equation (5) measures the root-mean-squared deviation between the value predictions of the limiting TD value function and the true value function. On the right hand side, the projected value function  $\Pi_D V_\mu$  minimizes root-mean-squared prediction errors among all value functions in the span of  $\Phi$ . If  $V_\mu$  actually falls within the span of the features, there is no approximation error at all and TD converges to the true value function.

**Asymptotic convergence via the ODE method.** Like many analyses in reinforcement learning, the convergence proof of [Tsitsiklis and Van Roy \[1997\]](#) appeals to a powerful technique from the stochastic approximation literature known as the “ODE method”. Under appropriate conditions, and assuming a decaying step-size sequence satisfying the Robbins-Monro conditions, this method establishes the asymptotic convergence of the stochastic recursion  $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$  as a consequence of the global asymptotic stability of the deterministic ODE:  $\dot{\theta}_t = \bar{g}(\theta_t)$ . The critical step in the proof of [Tsitsiklis and Van Roy \[1997\]](#) is to use the contraction properties of the Bellman operator to establish this ODE is globally asymptotically stable with the equilibrium point  $\theta^*$ .

The ODE method vastly simplifies convergence proofs. First, because the continuous dynamics can be easier to analyze than discretized ones, and more importantly, because it avoids dealing with stochastic noise in the problem. At the same time, by side-stepping these issues, the method offers little insight into the critical effect of step-size sequences, problem conditioning, and mixing time issues on algorithm performance.

## 5 Outline of analysis

The remainder of the paper focuses on a finite time analysis of TD. Broadly, we establish two types of finite time bounds. We first derive bounds that depend on the condition number of the feature covariance matrix. In that case, we state explicit bounds on the expected distance  $\mathbb{E} [\|\theta_T - \theta^*\|_2^2]$  of the iterate from the TD fixed-point,  $\theta^*$ . These mirror what one might expect from the literature on stochastic optimization of strongly convex functions: results showing that TD with constant step-sizes converges to within a radius of  $\theta^*$  at an exponential rate, and  $O(1/T)$  convergence rates with appropriate decaying step-sizes. Note that by Lemma 1,  $\|V_{\theta_T} - V_{\theta^*}\|_D^2 \leq \|\theta_T - \theta^*\|_2^2$ , so bounds on the distance of the iterate to the TD fixed point also imply bounds on the distance between value predictions.

These results establish fast rates of convergence, but only if the problem is well conditioned. The choice of step-sizes is also very sensitive to problem conditioning. Work on robust stochastic approximation

[Nemirovski et al., 2009] argues instead for the use of comparatively large step-sizes together with iterate averaging. Following the spirit of this work, we also give explicit bounds on  $\mathbb{E} [\|V_{\bar{\theta}_T} - V_{\theta^*}\|_D^2]$ , which measures the mean-squared gap between the predictions under the averaged-iterate  $\bar{\theta}_T$  and under the TD limit point  $\theta^*$ . These yield slower  $O(1/\sqrt{T})$  convergence rates, but both the bounds and step-sizes are completely independent of problem conditioning.

Our approach is to start by developing insights from simple, stylized settings, and then incrementally extend the analysis to more complex settings. The analysis is outlined below.

**Noiseless case:** Drawing inspiration from the ODE method discussed above, we start by analyzing the Euler discretization of the ODE  $\dot{\theta}_t = \bar{g}(\theta_t)$ , which is the deterministic recursion  $\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t)$ . We call this method “mean-path TD”. As motivation, the section first considers a fictitious gradient descent algorithm designed to converge to the TD fixed point. We then develop striking analogues for mean-path TD of the key properties underlying the convergence of gradient descent. Easy proofs then yield two bounds mirroring those given for gradient descent.

**Independent noise:** Section 7 studies TD under an i.i.d. observation model, where the data-tuples used by TD are drawn i.i.d. from the stationary distribution. The techniques used to analyze mean-path TD(0) extend easily to this setting, and the resulting bounds mirror standard guarantees for stochastic gradient descent.

**Markov noise:** In Section 8, we analyze TD in the more realistic setting where the data is collected from a single sample path of an ergodic Markov chain. This setting introduces significant challenges due to the highly dependent nature of the data. For tractability, we assume the Markov chain satisfies a certain uniform bound on the rate at which it mixes, and study a variant of TD that uses a projection step to ensure uniform boundedness of the iterates. In this case, our results essentially scale by a factor of the mixing time relative to the i.i.d. case.

**Extension to TD( $\lambda$ ):** In Section 9, we extend the analysis under the Markov noise to TD with eligibility traces, popularly known as TD( $\lambda$ ). Eligibility traces are known to often provide performance gains in practice, but theoretical analysis is more complex. Such analysis offers some insight into the subtle tradeoffs in the selection of the parameter  $\lambda \in [0, 1]$ .

**Approximate optimal stopping:** A final section extends our results to a class of high dimensional optimal stopping problems. We analyze Q-learning with linear function approximation. Building on observations of Tsitsiklis and Van Roy [1999], we show the key properties used in our analysis of TD continue to hold for Q-learning in this setting. The convergence bounds shown in Sections 7 and 8 therefore apply *without any modification*.

## 6 Analysis of mean-path TD

All practical applications of TD involve observation noise. However, a great deal of insight can be gained by investigating a natural deterministic analogue of the algorithm. Here we study the recursion

$$\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t) \quad t \in \mathbb{N}_0 = \{0, 1, 2, \dots\},$$

which is the Euler discretization of the ODE described in Section 4. We will refer to this iterative algorithm as *mean-path TD*. In this section, we develop key insights into the dynamics of mean-path TD that allow for a remarkably simple finite time analysis of its convergence. Later sections of the paper show how these ideas extend gracefully to analyses with observation noise.

The key to our approach is to develop properties of mean-path TD that closely mirror those of gradient descent on a particular quadratic loss function. To this end, in the next subsection, we review a simple analysis of gradient descent. In Subsection 6.2, we establish key properties of mean-path TD mirroring those used to analyze this gradient descent algorithm. Finally, Subsection 6.3 gives convergence rates of mean-path TD,

with proofs and rates mirroring those given for gradient descent except for a constant that depends on the discount factor,  $\gamma$ .

## 6.1 Gradient descent on a value function loss

Consider the cost function

$$f(\theta) = \frac{1}{2} \|V_{\theta^*} - V_\theta\|_D^2 = \frac{1}{2} \|\theta^* - \theta\|_\Sigma^2,$$

which measures the mean-squared gap between the value predictions under  $\theta$  and those under the stationary point of TD,  $\theta^*$ . Consider as well a hypothetical algorithm that performs gradient descent on  $f$ , iterating  $\theta_{t+1} = \theta_t - \alpha \nabla f(\theta_t)$  for all  $t \in \mathbb{N}_0$ . Of course, this algorithm is not implementable, as one does not know the limit point  $\theta^*$  of TD. However, reviewing an analysis of such an algorithm will offer great insights into our eventual analysis of TD.

To start, a standard decomposition characterizes the evolution of the error at iterate  $\theta_t$ :

$$\|\theta^* - \theta_{t+1}\|_2^2 = \|\theta^* - \theta_t\|_2^2 + 2\alpha \nabla f(\theta_t)^\top (\theta^* - \theta_t) + \alpha^2 \|\nabla f(\theta_t)\|_2^2.$$

To use this decomposition, we need two things. First, some understanding of  $\nabla f(\theta_t)^\top (\theta^* - \theta_t)$ , capturing whether the gradient points in the direction of  $(\theta^* - \theta_t)$ . And second, we need an upper bound on the norm of the gradient  $\|\nabla f(\theta_t)\|_2^2$ . In this case,  $\nabla f(\theta) = \Sigma(\theta - \theta^*)$ , from which we conclude

$$\nabla f(\theta)^\top (\theta^* - \theta) = -\|\theta^* - \theta\|_\Sigma^2 = -\|V_{\theta^*} - V_\theta\|_D^2. \quad (6)$$

In addition, one can show<sup>7</sup>

$$\|\nabla f(\theta)\|_2 \leq \|V_{\theta^*} - V_\theta\|_D. \quad (7)$$

Now, using (6) and (7), we have that for step-size  $\alpha = 1$ ,

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - \|V_{\theta^*} - V_{\theta_t}\|_D^2. \quad (8)$$

The distance to  $\theta^*$  decreases in every step, and does so more rapidly if there is a large gap between the value predictions under  $\theta$  and  $\theta^*$ . Combining this with Lemma 1 gives

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq (1 - \omega) \|\theta^* - \theta_t\|_2^2 \leq \dots \leq (1 - \omega)^{t+1} \|\theta^* - \theta_0\|_2^2. \quad (9)$$

Recall that  $\omega$  denotes the minimum eigenvalue of  $\Sigma$ . This shows that error converges at a fast geometric rate. However the rate of convergence degrades if the minimum eigenvalue  $\omega$  is close to zero. Such a convergence rate is therefore only meaningful if the feature covariance matrix is well conditioned.

By working in the space of value functions and performing iterate averaging, one can also give a guarantee that is independent of  $\omega$ . Recall the notation  $\bar{\theta}_T = T^{-1} \sum_{t=0}^{T-1} \theta_t$  for the averaged iterate. A simple proof from (8) shows

$$\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \frac{\|\theta^* - \theta_0\|_2^2}{T}. \quad (10)$$

## 6.2 Key properties of mean-path TD

This subsection establishes analogues for mean-path TD of the key properties (6) and (7) used to analyze gradient descent. First, to characterize the gradient update, our analysis builds on Lemma 7 of [Tsitsiklis and Van Roy \[1997\]](#), which uses the contraction properties of the projected Bellman operator to conclude

$$\bar{g}(\theta)^\top (\theta^* - \theta) > 0 \quad \forall \theta \neq \theta^*. \quad (11)$$

---

<sup>7</sup>This can be seen from the fact that for any vector  $u$  with  $\|u\|_2 \leq 1$ ,

$$u^\top \nabla f(\theta) = \langle u, \theta - \theta^* \rangle_\Sigma \leq \|u\|_\Sigma \|\theta^* - \theta\|_\Sigma \leq \|\theta^* - \theta\|_\Sigma = \|V_{\theta^*} - V_\theta\|_D.$$

That is, the expected update of TD always forms a positive angle with  $(\theta^* - \theta)$ . Though only Equation (11) was stated in their lemma, Tsitsiklis and Van Roy [1997] actually reach a much stronger conclusion in their proof itself. This result, given in Lemma 3 below, establishes that the expected updates of TD point in a descent direction of  $\|\theta^* - \theta\|_D^2$ , and do so more strongly when the gap between value functions under  $\theta$  and  $\theta^*$  is large. We will show that this more quantitative form of (11) allows for elegant finite time-bounds on the performance of TD.

Note that this lemma mirrors the property in Equation (6), but with a smaller constant of  $(1 - \gamma)$ . This reflects that expected TD must converge to  $\theta^*$  by bootstrapping [Sutton, 1988] and may follow a less direct path to  $\theta^*$  than the fictitious gradient descent method considered in the previous subsection. Recall that the limit point  $\theta^*$  solves  $\bar{g}(\theta^*) = 0$ .

**Lemma 3.** *For any  $\theta \in \mathbb{R}^d$ ,*

$$(\theta^* - \theta)^\top \bar{g}(\theta) \geq (1 - \gamma) \|V_{\theta^*} - V_\theta\|_D^2.$$

*Proof.* We use the notation described in Equation (2) of Section 3. Consider a stationary sequence of states with random initial state  $s \sim \pi$  and subsequent state  $s'$ , which, conditioned on  $s$ , is drawn from  $\mathcal{P}(\cdot|s)$ . Set  $\phi = \phi(s)$ ,  $\phi' = \phi(s')$  and  $r = \mathcal{R}(s, s')$ . Define  $\xi = V_{\theta^*}(s) - V_\theta(s) = (\theta^* - \theta)^\top \phi$  and  $\xi' = V_{\theta^*}(s') - V_\theta(s') = (\theta^* - \theta)^\top \phi'$ . By stationarity,  $\xi$  and  $\xi'$  are two correlated random variables with the same same marginal distribution. By definition,  $\pi, \mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$  since  $s$  is drawn from  $\pi$ .

Using the expression for  $\bar{g}(\theta)$  in Equation (2),

$$\bar{g}(\theta) = \bar{g}(\theta) - \bar{g}(\theta^*) = \mathbb{E}[\phi(\gamma\phi' - \phi)^\top (\theta - \theta^*)] = \mathbb{E}[\phi(\xi - \gamma\xi')]. \quad (12)$$

Therefore

$$(\theta^* - \theta)^\top \bar{g}(\theta) = \mathbb{E}[\xi(\xi - \gamma\xi')] = \mathbb{E}[\xi^2] - \gamma \mathbb{E}[\xi'\xi] \geq (1 - \gamma) \mathbb{E}[\xi^2] = (1 - \gamma) \|V_{\theta^*} - V_\theta\|_D^2.$$

The inequality above uses Cauchy-Schwartz inequality together with the fact that  $\xi$  and  $\xi'$  have the same marginal distribution to conclude  $\mathbb{E}[\xi\xi'] \leq \sqrt{\mathbb{E}[\xi^2]} \sqrt{\mathbb{E}[(\xi')^2]} = \mathbb{E}[\xi^2]$ .  $\square$

Lemma 4 is the other key ingredient to our results. It upper bounds the norm of the expected negative gradient, providing an analogue of Equation (7).

**Lemma 4.**  $\|\bar{g}(\theta)\|_2 \leq 2 \|V_\theta - V_{\theta^*}\|_D \quad \forall \theta \in \mathbb{R}^d$ .

*Proof.* Beginning from (12) in the Proof of Lemma 3, we have

$$\|\bar{g}(\theta)\|_2 = \|\mathbb{E}[\phi(\xi - \gamma\xi')]\|_2 \leq \sqrt{\mathbb{E}[\|\phi\|_2^2]} \sqrt{\mathbb{E}[(\xi - \gamma\xi')^2]} \leq \sqrt{\mathbb{E}[\xi^2]} + \gamma \sqrt{\mathbb{E}[(\xi')^2]} = (1 + \gamma) \sqrt{\mathbb{E}[\xi^2]},$$

where the second inequality uses the assumption that  $\|\phi\|_2 \leq 1$  and the final equality uses that  $\xi$  and  $\xi'$  have the same marginal distribution. We conclude by recalling that  $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$  and  $1 + \gamma \leq 2$ .  $\square$

Lemmas 3 and 4 are quite powerful when used in conjunction. As in the analysis of gradient descent reviewed in the previous subsection, our analysis starts with a recursion for the error term,  $\|\theta_t - \theta^*\|^2$ . See Equation (13) in Theorem 1 below. Lemma 3 shows the first order term in this recursion reduces the error at each time step, while using the two lemmas in conjunction shows the first order term dominates a constant times the second order term. Precisely,

$$\bar{g}(\theta)^\top (\theta^* - \theta) \geq (1 - \gamma) \|V_{\theta^*} - V_\theta\|_D^2 \geq \frac{(1 - \gamma)}{4} \|\bar{g}(\theta)\|_2^2.$$

This leads immediately to conclusions like Equation (14), from which finite time convergence bounds follow.

It is also worth pointing out that as TD(0) is an instance of linear stochastic approximation, these two lemmas can be interpreted as statements about the eigenvalues of the matrix driving its behavior<sup>8</sup>.

<sup>8</sup>Recall from Section 3 that  $\bar{g}(\theta)$  is an affine function. That is, it can be written as  $A\theta - b$  for some  $A \in \mathbb{R}^{d \times d}$  and  $b \in \mathbb{R}^d$ . Lemma 3 shows that  $A \preceq -(1 - \gamma)\Sigma$ , i.e. that  $A + (1 - \gamma)\Sigma$  is negative definite. It is easy to show that  $\|\bar{g}(\theta)\|_2^2 = (\theta - \theta^*)^\top (A^\top A)(\theta - \theta^*)$ , so Lemma 4 shows that  $A^\top A \preceq \Sigma$ . Taking this perspective, the important part of these lemmas is that they allows us to understand TD in terms of feature covariance matrix  $\Sigma$  and the discount factor  $\gamma$  rather than the more mysterious matrix  $A$ .

### 6.3 Finite time analysis of mean-path TD

We now combine the insights of the previous subsection to establish convergence rates for mean-path TD. These mirror the bounds for gradient descent given in Equations (9) and (10), except for an additional dependence on the discount factor. The first result bounds the distance between the value function under an averaged iterate and under the TD stationary point. This gives a comparatively slow  $O(1/T)$  convergence rate, but does not depend at all on the conditioning of the feature covariance matrix. When this matrix is well conditioned, so the minimum eigenvalue  $\omega$  of  $\Sigma$  is not too small, the geometric convergence rate given in the second part of the theorem dominates. Note that by Lemma 1, bounds on  $\|\theta_t - \theta^*\|_2$  always imply bounds on  $\|V_{\theta_t} - V_{\theta^*}\|_D$ .

**Theorem 1.** Consider a sequence of parameters  $(\theta_0, \theta_1, \dots)$  obeying the recursion

$$\theta_{t+1} = \theta_t + \alpha \bar{g}(\theta_t) \quad t \in \mathbb{N}_0 = \{0, 1, 2, \dots\},$$

where  $\alpha = (1 - \gamma)/4$ . Then,

$$\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \leq \frac{4\|\theta^* - \theta_0\|_2^2}{T(1 - \gamma)^2}$$

and

$$\|\theta^* - \theta_T\|_2^2 \leq \exp \left\{ - \left( \frac{(1 - \gamma)^2 \omega}{4} \right) T \right\} \|\theta^* - \theta_0\|_2^2.$$

*Proof.* With probability 1, for every  $t \in \mathbb{N}_0$ , we have

$$\|\theta^* - \theta_{t+1}\|_2^2 = \|\theta^* - \theta_t\|_2^2 - 2\alpha(\theta^* - \theta_t)^\top \bar{g}(\theta_t) + \alpha^2 \|\bar{g}(\theta_t)\|_2^2. \quad (13)$$

Applying Lemmas 3 and 4 and using a constant step-size of  $\alpha = (1 - \gamma)/4$ , we get

$$\begin{aligned} \|\theta^* - \theta_{t+1}\|_2^2 &\leq \|\theta^* - \theta_t\|_2^2 - (2\alpha(1 - \gamma) - 4\alpha^2) \|V_{\theta^*} - V_{\theta_t}\|_D^2 \\ &= \|\theta^* - \theta_t\|_2^2 - \left( \frac{(1 - \gamma)^2}{4} \right) \|V_{\theta^*} - V_{\theta_t}\|_D^2. \end{aligned} \quad (14)$$

Then,

$$\left( \frac{(1 - \gamma)^2}{4} \right) \sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \sum_{t=0}^{T-1} (\|\theta^* - \theta_t\|_2^2 - \|\theta^* - \theta_{t+1}\|_2^2) \leq \|\theta^* - \theta_0\|_2^2.$$

Applying Jensen's inequality gives the first result:

$$\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2 \leq \frac{4\|\theta^* - \theta_0\|_2^2}{(1 - \gamma)^2 T}.$$

Now, returning to (14), and applying Lemma 1 implies

$$\begin{aligned} \|\theta^* - \theta_{t+1}\|_2^2 &\leq \|\theta^* - \theta_t\|_2^2 - \left( \frac{(1 - \gamma)^2}{4} \right) \omega \|\theta^* - \theta_t\|_2^2 = \left( 1 - \frac{\omega(1 - \gamma)^2}{4} \right) \|\theta^* - \theta_t\|_2^2 \\ &\leq \exp \left\{ - \frac{\omega(1 - \gamma)^2}{4} \right\} \|\theta^* - \theta_t\|_2^2, \end{aligned}$$

where the final inequality uses that  $\left( 1 - \frac{\omega(1 - \gamma)^2}{4} \right) \leq e^{-\frac{\omega(1 - \gamma)^2}{4}}$ . Repeating this inductively gives the desired result.  $\square$

## 7 Analysis for the i.i.d. observation model

This section studies TD under an i.i.d. observation model, and establishes three explicit guarantees that mirror standard finite time bounds available for SGD. Specifically, we study a model where the random tuples observed by the TD algorithm are sampled i.i.d. from the stationary distribution of the Markov reward process. This means that for all states  $s$  and  $s'$ ,

$$\mathbb{P}[(s_t, r_t, s'_t) = (s, \mathcal{R}(s, s'), s')] = \pi(s)\mathcal{P}(s'|s), \quad (15)$$

and the tuples  $\{(s_t, r_t, s'_t)\}_{t \in \mathbb{N}}$  are drawn independently across time. Note that the probabilities in Equation (15) correspond to a setting where the first state  $s_t$  is drawn from the stationary distribution, and then  $s'_t$  is drawn from  $\mathcal{P}(\cdot|s_t)$ . This model is widely used for analyzing RL algorithms. See for example Sutton et al. [2009b], Sutton et al. [2009a], Korda and La [2015], and Dalal et al. [2018a].

Theorem 2 follows from a unified analysis that combines the techniques of the previous section with typical arguments used in the SGD literature. All bounds depend on  $\sigma^2 = \mathbb{E}[\|g_t(\theta^*)\|_2^2] = \mathbb{E}[\|g_t(\theta^*) - \bar{g}(\theta^*)\|_2^2]$ , which roughly captures the variance of TD updates at the stationary point  $\theta^*$ . The bound in part (a) follows the spirit of work on so-called *robust stochastic approximation* [Nemirovski et al., 2009]. It applies to TD with iterate averaging and relatively large step-sizes. The result is a simple bound on the mean-squared gap between the value predictions under the averaged iterate and the TD fixed point. The main strength of this result is that the step-sizes and the bound do not depend at all on the condition number of the feature covariance matrix. Note that the requirement that  $\sqrt{T} \geq 8/(1-\gamma)$  is not critical; one can carry out analysis using the step-size  $\alpha_0 = \min\{(1-\gamma)/8, \sqrt{T}\}$ , but the bounds we attain only become meaningful in the case where  $T$  is sufficiently large, so we chose to simplify the exposition.

Parts (b) and (c) provide faster convergence rates in the case where the feature covariance matrix is well conditioned. Part (b) studies TD applied with a constant step-size, which is common in practice. In this case, the iterate  $\theta_t$  will never converge to the TD fixed point, but our results show the expected distance to  $\theta^*$  converges at an exponential rate below some level that depends on the choice of step-size. This is sometimes referred to as the rate at which the initial point  $\theta_0$  is “forgotten”. Bounds like this justify the common practice of starting with large step-sizes, and sometimes dividing the step-sizes in half once it appears error is no-longer decreasing. Part (c) attains an  $\mathcal{O}(1/T)$  convergence rate for a carefully chosen decaying step-size sequence. This step-size sequence requires knowledge of the minimum eigenvalue of the feature covariance matrix  $\Sigma$ , which plays a role similar to a strong convexity parameter in the optimization literature. In practice, this would need to be estimated, possibly by constructing a sample average approximation to the feature covariance matrix. The proof of part (c) closely follows an inductive argument presented in Bottou et al. [2018]. Recall that  $\bar{\theta}_T = T^{-1} \sum_{t=0}^{T-1} \theta_t$  denotes the averaged iterate.

**Theorem 2.** Suppose TD is applied under the i.i.d. observation model and set  $\sigma^2 = \mathbb{E}[\|g_t(\theta^*)\|_2^2]$ .

(a) For any  $T \geq (8/(1-\gamma))^2$  and a constant step-size sequence  $\alpha_0 = \dots = \alpha_T = \frac{1}{\sqrt{T}}$ ,

$$\mathbb{E}[\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2] \leq \frac{\|\theta^* - \theta_0\|_2^2 + 2\sigma^2}{\sqrt{T}(1-\gamma)}.$$

(b) For any constant step-size sequence  $\alpha_0 = \dots = \alpha_T \leq \omega(1-\gamma)/8$ ,

$$\mathbb{E}[\|\theta^* - \theta_T\|_2^2] \leq \left(e^{-\alpha_0(1-\gamma)\omega T}\right) \|\theta^* - \theta_0\|_2^2 + \alpha_0 \left(\frac{2\sigma^2}{(1-\gamma)\omega}\right).$$

(c) For a decaying step-size sequence  $\alpha_t = \frac{\beta}{\lambda+t}$  with  $\beta = \frac{2}{(1-\gamma)\omega}$  and  $\lambda = \frac{16}{(1-\gamma)^2\omega}$ ,

$$\mathbb{E}[\|\theta^* - \theta_T\|_2^2] \leq \frac{\nu}{\lambda+T} \quad \text{where} \quad \nu = \max \left\{ \frac{8\sigma^2}{(1-\gamma)^2\omega^2}, \frac{16\|\theta^* - \theta_0\|_2^2}{(1-\gamma)^2\omega} \right\}.$$

Our proof is able to directly leverage Lemma 3, but the analysis requires the following extension of Lemma 4 which gives an upper bound to the expected norm of the stochastic gradient.

**Lemma 5.** *For any fixed  $\theta \in \mathbb{R}^d$ ,  $\mathbb{E} [\|g_t(\theta)\|_2^2] \leq 2\sigma^2 + 8\|V_{\theta^*} - V_\theta\|_D^2$  where  $\sigma^2 = \mathbb{E} [\|g_t(\theta^*)\|_2^2]$ .*

*Proof.* For brevity of notation, set  $\phi = \phi(s_t)$  and  $\phi' = \phi(s'_t)$ . Define  $\xi = (\theta^* - \theta)^\top \phi$  and  $\xi' = (\theta^* - \theta)^\top \phi'$ . By stationarity,  $\xi$  and  $\xi'$  have the same marginal distribution and  $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$ , following the same argument as in Lemma 3. Using the formula for  $g_t(\theta)$  in Equation (1), we have

$$\begin{aligned}\mathbb{E} [\|g_t(\theta)\|_2^2] &\leq \mathbb{E} [(\|g_t(\theta^*)\|_2 + \|g_t(\theta) - g_t(\theta^*)\|_2)^2] \\ &\leq 2\mathbb{E} [\|g_t(\theta^*)\|_2^2] + 2\mathbb{E} [\|g_t(\theta) - g_t(\theta^*)\|_2^2] \\ &= 2\sigma^2 + 2\mathbb{E} [\|\phi(\phi - \gamma\phi')^\top (\theta^* - \theta)\|_2^2] \\ &= 2\sigma^2 + 2\mathbb{E} [\|\phi(\xi - \gamma\xi')\|_2^2] \\ &\leq 2\sigma^2 + 2\mathbb{E} [|\xi - \gamma\xi'|^2] \\ &\leq 2\sigma^2 + 4(\mathbb{E} [|\xi|^2] + \gamma^2\mathbb{E} [|\xi'|^2]) \\ &\leq 2\sigma^2 + 8\|V_{\theta^*} - V_\theta\|_D^2,\end{aligned}$$

where we used the assumption that  $\|\phi\|_2^2 \leq 1$ . The second inequality uses the basic algebraic identity  $(x + y)^2 \leq 2\max\{x, y\}^2 \leq 2x^2 + 2y^2$ , along with the monotonicity of expectation operators.  $\square$

Using this we give a proof of Theorem 2 below. Let us remark here on a consequence of the i.i.d noise model that considerably simplifies the proof. Until now, we have often developed properties of the TD updates  $g_t(\theta)$  applied to an arbitrary, but fixed, vector  $\theta \in \mathbb{R}^d$ . For example, we have given an expression for  $\bar{g}(\theta) := \mathbb{E}[g_t(\theta)]$ , where this expectation integrates over the random tuple  $O_t = (s_t, r_t, s'_t)$  influencing the TD update. In the i.i.d noise model, the current iterate,  $\theta_t$ , is independent of the tuple  $O_t$ , and so  $\mathbb{E}[g_t(\theta_t)|\theta_t] = \bar{g}(\theta_t)$ . In a similar manner, after conditioning on  $\theta_t$ , we can seamlessly apply Lemmas 3 and 5, as is done in inequality (16) of the proof below.

*Proof.* The TD algorithm updates the parameters as:  $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$ . Thus, for each  $t \in \mathbb{N}_0$ , we have,

$$\|\theta^* - \theta_{t+1}\|_2^2 = \|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^\top (\theta^* - \theta_t) + \alpha_t^2 \|g_t(\theta_t)\|_2^2.$$

Under the hypotheses of (a), (b) and (c), we have that  $\alpha_t \leq (1 - \gamma)/8$ . Taking expectations and applying Lemma 3 and Lemma 5 implies,

$$\begin{aligned}\mathbb{E} [\|\theta^* - \theta_{t+1}\|_2^2] &= \mathbb{E} [\|\theta^* - \theta_t\|_2^2] - 2\alpha_t \mathbb{E} [g_t(\theta_t)^\top (\theta^* - \theta_t)] + \alpha_t^2 \mathbb{E} [\|g_t(\theta_t)\|_2^2] \\ &= \mathbb{E} [\|\theta^* - \theta_t\|_2^2] - 2\alpha_t \mathbb{E} [\mathbb{E} [g_t(\theta_t)^\top (\theta^* - \theta_t) | \theta_t]] + \alpha_t^2 \mathbb{E} [\mathbb{E} [\|g_t(\theta_t)\|_2^2 | \theta_t]] \\ &\leq \mathbb{E} [\|\theta^* - \theta_t\|_2^2] - (2\alpha_t(1 - \gamma) - 8\alpha_t^2) \mathbb{E} [\|V_{\theta^*} - V_{\theta_t}\|_D^2] + 2\alpha_t^2 \sigma^2 \quad (16) \\ &\leq \mathbb{E} [\|\theta^* - \theta_t\|_2^2] - \alpha_t(1 - \gamma) \mathbb{E} [\|V_{\theta^*} - V_{\theta_t}\|_D^2] + 2\alpha_t^2 \sigma^2. \quad (17)\end{aligned}$$

The inequality (16) follows from Lemmas 3 and 5. The application of these lemmas uses that the random tuple  $O_t = (s_t, r_t, s'_t)$  influencing  $g_t(\cdot)$  is independent of the iterate,  $\theta_t$ .

**Part (a).** Consider a constant step-size of  $\alpha_T = \dots = \alpha_0 = 1/\sqrt{T}$ . Starting with Equation (17) and summing over  $t$  gives

$$\mathbb{E} \left[ \sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2}{\alpha_0(1 - \gamma)} + \frac{2\alpha_0 T \sigma^2}{(1 - \gamma)} = \frac{\sqrt{T} \|\theta^* - \theta_0\|_2^2}{(1 - \gamma)} + \frac{2\sqrt{T} \sigma^2}{(1 - \gamma)}.$$

We find

$$\mathbb{E} [\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2] \leq \frac{1}{T} \mathbb{E} \left[ \sum_{t=0}^{T-1} \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + 2\sigma^2}{\sqrt{T}(1 - \gamma)}.$$

**Part (b).** Consider a constant step-size of  $\alpha_0 \leq \omega(1 - \gamma)/8$ . Applying Lemma 1 to Equation (17) implies

$$\mathbb{E} [\|\theta^* - \theta_{t+1}\|_2^2] \leq (1 - \alpha_0(1 - \gamma)\omega) \mathbb{E} [\|\theta^* - \theta_t\|_2^2] + 2\alpha_0^2\sigma^2. \quad (18)$$

Iterating this inequality establishes that for any  $T \in \mathbb{N}_0$ ,

$$\mathbb{E} [\|\theta^* - \theta_T\|_2^2] \leq (1 - \alpha_0(1 - \gamma)\omega)^T \mathbb{E} [\|\theta^* - \theta_0\|_2^2] + 2\alpha_0^2\sigma^2 \sum_{t=0}^{\infty} (1 - \alpha_0(1 - \gamma)\omega)^t.$$

The result follows by solving the geometric series and using that  $(1 - \alpha_0(1 - \gamma)\omega) \leq e^{-\alpha_0(1-\gamma)\omega}$ .

**Part (c).** Note that by the definitions of  $\nu$ ,  $\lambda$  and  $\beta$ , we have

$$\nu = \max\{2\beta^2\sigma^2, \lambda\|\theta^* - \theta_0\|_2^2\}.$$

We then have  $\|\theta^* - \theta_0\|_2^2 \leq \frac{\nu}{\lambda}$  by the definition of  $\nu$ . Proceeding by induction, suppose  $\mathbb{E} [\|\theta^* - \theta_t\|_2^2] \leq \frac{\nu}{\lambda+t}$ . Then,

$$\begin{aligned} \mathbb{E} [\|\theta^* - \theta_{t+1}\|_2^2] &\leq (1 - \alpha_t(1 - \gamma)\omega) \mathbb{E} [\|\theta^* - \theta_t\|_2^2] + 2\alpha_t^2\sigma^2 \\ &\leq \left(1 - \frac{(1 - \gamma)\omega\beta}{\hat{t}}\right) \frac{\nu}{\hat{t}} + \frac{2\beta^2\sigma^2}{\hat{t}^2} \quad [\text{where } \hat{t} \equiv \lambda + t] \\ &= \left(\frac{\hat{t} - (1 - \gamma)\omega\beta}{\hat{t}^2}\right) \nu + \frac{2\beta^2\sigma^2}{\hat{t}^2} \\ &= \left(\frac{\hat{t} - 1}{\hat{t}^2}\right) \nu + \frac{2\beta^2\sigma^2 - ((1 - \gamma)\omega\beta - 1)\nu}{\hat{t}^2} \\ &= \left(\frac{\hat{t} - 1}{\hat{t}^2}\right) \nu + \frac{2\beta^2\sigma^2 - \nu}{\hat{t}^2} \quad [\text{using } \beta = \frac{2}{(1 - \gamma)\omega}] \\ &\leq \frac{\nu}{\hat{t} + 1}, \end{aligned}$$

where the final inequality uses that  $2\beta^2\sigma^2 - \nu \leq 0$ , which holds by the definition of  $\nu$  and the fact that  $\hat{t}^2 \geq (\hat{t} - 1)(\hat{t} + 1)$ .  $\square$

## 8 Analysis for the Markov chain observation model: Projected TD algorithm

In Section 7, we developed a method for analyzing TD under an i.i.d. sampling model in which tuples are drawn independently from the stationary distribution of the underlying MDP. But a more realistic setting is one in which the observed tuples used by TD are gathered from a single trajectory of the Markov chain. In particular, if for a given sample path the Markov chain visits states  $(s_0, s_1, \dots, s_t, \dots)$ , then these are processed into tuples  $O_t = (s_t, r_t = \mathcal{R}(s_t, s_{t+1}), s_{t+1})$  that are fed into the TD algorithm. Mathematical analysis is difficult since the tuples used by the algorithm can be highly correlated with each other. We outline the main challenges below.

**Challenges in the Markov chain noise model.** In the i.i.d. observation setting, our analysis relied heavily on a Martingale property of the noise sequence. This no longer holds in the Markov chain model due to strong dependencies between the noisy observations. To understand this, recall the expression of the negative gradient,

$$g_t(\theta) = \left(r_t + \gamma\phi(s_{t+1})^\top\theta - \phi(s_t)^\top\theta\right)\phi(s_t). \quad (19)$$

To make the statistical dependencies more transparent, we can overload notation to write this as  $g(\theta, O_t) \equiv g_t(\theta)$ , where  $O_t = (s_t, r_t, s_{t+1})$ . Assuming the sequence of states is stationary, we have defined the function  $\bar{g} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  by  $\bar{g}(\theta) = \mathbb{E}[g(\theta, O_t)]$ , where, since  $\theta$  is non-random, this expectation integrates over the marginal distribution of the tuple  $O_t$ . However,  $\mathbb{E}[g(\theta_t, O_t) \mid \theta_t = \theta] \neq \bar{g}(\theta)$  because  $\theta_t$  is a function of past tuples  $\{O_1, \dots, O_{t-1}\}$ , potentially introducing strong dependencies between  $\theta_t$  and  $O_t$ . Similarly, in general  $\mathbb{E}[g(\theta_t, O_t) - \bar{g}(\theta_t)] \neq 0$ , indicating bias in the algorithm's gradient evaluation. A related challenge arises in trying to control the norm of the gradient step,  $\mathbb{E}[\|g_t(\theta_t)\|_2^2]$ . Lemma 5 does not yield a bound due to coupling between the iterate  $\theta_t$  and the observation  $O_t$ .

Our analysis uses an information-theoretic technique to control for this coupling and explicitly account for the gradient basis. This technique may be of broader use in analyzing reinforcement learning and stochastic approximation algorithms. However, our analysis also requires some strong regularity conditions, as outlined below.

**Projected TD algorithm.** Our technique for controlling the gradient bias relies critically on a condition that, when step-sizes are small, the iterates  $(\theta_t)_{t \in \mathbb{N}_0}$  do not change too rapidly. This is the case as long as norms of the gradient steps do not explode. For tractability, we modify the TD algorithm itself by adding a projection step that ensures gradient norms are uniformly bounded across time. In particular, starting with an initial guess of  $\theta_0$  such that  $\|\theta_0\|_2 \leq R$ , we consider the Projected TD algorithm, which iterates

$$\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t)) \quad \forall t \in \mathbb{N}_0, \quad (20)$$

where

$$\Pi_{2,R}(\theta) = \arg \min_{\theta' : \|\theta'\|_2 \leq R} \|\theta - \theta'\|_2$$

is the projection operator onto a norm ball of radius  $R < \infty$ . The subscript 2 on the operator indicates that the projection is with respect the unweighted Euclidean norm. This should not be confused with the projection operator  $\Pi_D$  used earlier, which projects onto the subspace of approximate value functions with respect to a weighted norm.

One may wonder whether this projection step is practical. We note that, from a computational perspective, it only involves rescaling of the iterates, as  $\Pi_{2,R}(\theta) = R\theta/\|\theta\|_2$  if  $\|\theta\|_2 > R$  and is simply  $\theta$  otherwise. In addition, Subsection 8.2 suggests that by using a priori bounds on the value function, it should be possible to estimate a projection radius containing the TD fixed point. However, at this stage, we view this mainly as a tool that enables clean finite time analysis, rather than a practical algorithmic proposal.

It is worth mentioning that projection steps have a long history in the stochastic approximation literature, and many of the standard analyses for stochastic gradient descent rely on projections steps to control the norm of the gradient [Kushner, 2010, Lacoste-Julien et al., 2012, Bubeck, 2015, Nemirovski et al., 2009].

**Structural assumptions on the Markov reward process.** To control the statistical bias in the gradient updates—which is the main challenge under the Markov observation model—we assume that the Markov chain mixes at a uniform geometric rate, as stated below.

**Assumption 1.** *There are constants  $m > 0$  and  $\rho \in (0, 1)$  such that*

$$\sup_{s \in \mathcal{S}} d_{TV}(\mathbb{P}(s_t \in \cdot | s_0 = s), \pi) \leq m\rho^t \quad \forall t \in \mathbb{N}_0,$$

where  $d_{TV}(P, Q)$  denotes the total-variation distance between probability measures  $P$  and  $Q$ . In addition, the initial distribution of  $s_0$  is the steady-state distribution  $\pi$ , so  $(s_0, s_1, \dots)$  is a stationary sequence.

This uniform mixing assumption always holds for irreducible and aperiodic Markov chains [Levin and Peres, 2017]. We emphasize that the assumption that the chain begins in steady-state is not essential: given the uniform mixing assumption, we can always apply our analysis after the Markov chain has approximately reached

its steady-state. However, adding this assumption allows us to simplify many mathematical expressions. Another useful quantity for our analysis is the mixing time which we define as

$$\tau^{\text{mix}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}. \quad (21)$$

For interpreting the bounds, note that from Assumption 1,

$$\tau^{\text{mix}}(\epsilon) \sim \frac{\log(1/\epsilon)}{\log(1/\rho)} \quad \text{as } \epsilon \rightarrow 0.$$

We can therefore evaluate the mixing time at very small thresholds like  $\epsilon = 1/T$  while only contributing a logarithmic factor to the bounds.

**A bound on the norm of the gradient:** Before proceeding, we also state a bound on the euclidean norm of the gradient under TD(0) that follows from the uniform bound on rewards, along with feature normalization<sup>9</sup> and boundedness of the iterates through the projection step. Under projected TD(0) with projection radius  $R$ , this lemma implies that  $\|g_t(\theta_t)\|_2 \leq (r_{\max} + 2R)$ . This gradient bound plays an important role in our convergence bounds.

**Lemma 6.** *For all  $\theta \in \mathbb{R}^d$ ,  $\|g_t(\theta)\|_2 \leq r_{\max} + 2\|\theta\|_2$  with probability 1.*

*Proof.* Using the expression of  $g_t(\theta)$  in Equation (19), we have

$$\begin{aligned} \|g_t(\theta)\|_2 &\leq |r_t + (\gamma\phi(s'_t) - \phi(s_t))^\top \theta| \|\phi(s_t)\| \leq r_{\max} + \|\gamma\phi(s'_t) - \phi(s_t)\|_2 \|\theta\|_2 \\ &\leq r_{\max} + 2\|\theta\|. \end{aligned}$$

□

## 8.1 Finite time bounds

Following Section 7, we state several finite time bounds on the performance of the Projected TD algorithm. As before, in the spirit of robust stochastic approximation [Nemirovski et al., 2009], the bound in part (a) gives a comparatively slow convergence rate of  $\tilde{\mathcal{O}}(1/\sqrt{T})$ , but where the bound and step-size sequence are independent of the conditioning of the feature covariance matrix  $\Sigma$ . The bound in part (c) gives a faster convergence rate in terms of the number of samples  $T$ , but the bound and as well as the step-size sequence depend on the minimum eigenvalue  $\omega$  of  $\Sigma$ . Part (b) confirms that for sufficiently small step-sizes, the iterates converge at an exponential rate to within some radius of the TD fixed-point,  $\theta^*$ .

It is also instructive to compare the bounds for the Markov model vis-a-vis the i.i.d. model. One can see that in the case of part(b) for the Markov chain setting, a  $\mathcal{O}(G^2\tau^{\text{mix}}(\alpha_0))$  term controls the limiting error due to gradient noise. This scaling by the mixing time is intuitive, reflecting that roughly every cycle of  $\tau^{\text{mix}}(\cdot)$  observations provides as much information as a single independent sample from the stationary distribution. We can also imagine specializing the results to the case of Projected TD under the i.i.d. model, thereby eliminating all terms depending on the mixing time. We would attain bounds that mirror those in Theorem 2, except that the gradient noise term  $\sigma^2$  there would be replaced by  $G^2$ . This is a consequence using  $G$  as a uniform upper bound on the gradient norm in the proof, which is possible because of the projection step.

**Theorem 3.** *Suppose the Projected TD algorithm is applied with parameter  $R \geq \|\theta^*\|_2$  under the Markov chain observation model with Assumption 1. Set  $G = (r_{\max} + 2R)$ . Then the following claims hold.*

(a) *With a constant step-size sequence  $\alpha_0 = \dots = \alpha_T = 1/\sqrt{T}$ ,*

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2 \left( 9 + 12\tau^{\text{mix}}(1/\sqrt{T}) \right)}{2\sqrt{T}(1-\gamma)}.$$

---

<sup>9</sup>Recall that we assumed  $\|\phi(s)\|_2 \leq 1$  for all  $s \in \mathcal{S}$  and  $|\mathcal{R}(s, s')| \leq r_{\max}$  for all  $s, s' \in \mathcal{S}$

(b) With a constant step-size sequence  $\alpha_0 = \dots = \alpha_T < 1/(2\omega(1-\gamma))$ ,

$$\mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] \leq \left( e^{-2\alpha_0(1-\gamma)\omega T} \right) \|\theta^* - \theta_0\|_2^2 + \alpha_0 \left( \frac{G^2 (9 + 12\tau^{\text{mix}}(\alpha_0))}{2(1-\gamma)\omega} \right).$$

(c) With a decaying step-size sequence  $\alpha_t = 1/(\omega(t+1)(1-\gamma))$  for all  $t \in \mathbb{N}_0$ ,

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{G^2 (9 + 24\tau^{\text{mix}}(\alpha_T))}{T(1-\gamma)^2\omega} (1 + \log T),$$

**Remark 1:** The proof of part (c) also implies an  $\tilde{\mathcal{O}}(1/T)$  convergence rate for the iterate  $\theta_T$  itself; similar to the  $\mathcal{O}(1/T)$  convergence shown for the i.i.d. case, in part (c) of Theorem 2.

**Remark 2:** It is likely possible to eliminate the  $\log T$  term in the numerator of part (c) to get a  $\mathcal{O}(1/T)$  convergence rate. One approach is to use a different weighting of the iterates when averaging, as in Lacoste-Julien et al. [2012]. For brevity and simplicity, we do not pursue this direction.

## 8.2 Choice of the projection radius

We briefly comment on the choice of the projection radius,  $R$ . Note that Theorem 3 assumes that  $\|\theta^*\|_2 \leq R$ , so the TD limit point lies within the projected ball. How do we choose such an  $R$  when  $\theta^*$  is unknown? It turns out we can use Lemma 2, which relates the value function at the limit of convergence  $V_{\theta^*}$  to the true value function, to give a conservative upper bound. This is shown in the proof of the following lemma.

**Lemma 7.**  $\|\theta^*\|_{\Sigma} \leq \frac{2r_{\max}}{(1-\gamma)^{3/2}}$  and hence  $\|\theta^*\|_2 \leq \frac{2r_{\max}}{\sqrt{\omega(1-\gamma)^{3/2}}}$ .

*Proof.* Because rewards are uniformly bounded,  $|V_{\mu}(s)| \leq r_{\max}/(1-\gamma)$  for all  $s \in \mathcal{S}$ . Recall that  $V_{\mu}$  denotes the true value function of the Markov reward process. This implies that

$$\|V_{\mu}\|_D \leq \|V_{\mu}\|_{\infty} \leq \frac{r_{\max}}{(1-\gamma)}.$$

Lemma 2 along with simple matrix inequalities enable a simple upper bound on  $\|\theta^*\|_2$ . We have

$$\|V_{\theta^*} - V_{\mu}\|_D \leq \frac{1}{\sqrt{1-\gamma^2}} \|V_{\mu} - \Pi_D V_{\mu}\|_D \leq \frac{1}{\sqrt{1-\gamma^2}} \|V_{\mu}\|_D \leq \frac{1}{\sqrt{1-\gamma}} \|V_{\mu}\|_D,$$

where the penultimate inequality holds by the Pythagorean theorem. By the reverse triangle inequality we have  $||V_{\theta^*} - V_{\mu}||_D \leq ||V_{\theta^*} - V_{\mu}||_D$ . Thus,

$$\|V_{\theta^*}\|_D \leq \|V_{\theta^*} - V_{\mu}\|_D + \|V_{\mu}\|_D \leq \frac{2}{\sqrt{1-\gamma}} \|V_{\mu}\|_D \leq \frac{2}{\sqrt{1-\gamma}} \frac{r_{\max}}{(1-\gamma)}.$$

Recall from Section 2 we have,  $\|V_{\theta^*}\|_D = \|\theta^*\|_{\Sigma}$  which establishes first part of the claim. The second claim uses that  $\|\theta^*\|_{\Sigma} \geq \omega \|\theta^*\|_2$  which follows by Lemma 1.  $\square$

It is important to remark here that this bound is *problem dependent* as it depends on the minimum eigenvalue  $\omega$  of the steady-state feature covariance matrix  $\Sigma$ . We believe that estimating  $\omega$  online would make the projection step practical to implement.

### 8.3 Analysis

We now present the key analysis used to establish Theorem 3. Throughout, we assume the conditions of the theorem hold: we consider the Markov chain observation model with Assumption 1 and study the Projected TD algorithm applied with parameter  $R \geq \|\theta^*\|_2$  and some step-size sequence  $(\alpha_0, \dots, \alpha_T)$ .

We fix some notation throughout the scope of this subsection. Define the set  $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$ , so  $\theta_t \in \Theta_R$  for each  $t$  because of the algorithm's projection step. Set  $G = (r_{\max} + 2R)$ , so  $\|g_t(\theta)\|_2 \leq G$  for all  $\theta \in \Theta_R$  by Lemma 6. Finally, we set

$$\zeta_t(\theta) \equiv (g_t(\theta) - \bar{g}(\theta))^\top (\theta - \theta^*) \quad \forall \theta \in \Theta_R,$$

which can be thought of as the error in the evaluation of gradient-update under parameter  $\theta$  at time  $t$ .

Referring back to the analysis of the i.i.d. observation model, one can see that an error decomposition given in Equation (17) is the crucial component of the proof. The main objective in this section is to establish two key lemmas that yield a similar decomposition in the Markov chain observation model. The result can be stated cleanly in the case of a constant step-size. If  $\alpha_0 = \dots = \alpha_T = \alpha$ , we show

$$\begin{aligned} \mathbb{E} [\|\theta^* - \theta_{t+1}\|_2^2] &\leq \mathbb{E} [\|\theta^* - \theta_t\|_2^2] - 2\alpha(1-\gamma)\mathbb{E} [\|V_{\theta^*} - V_{\theta_t}\|_D^2] + \mathbb{E} [\alpha\zeta_t(\theta_t)] + \alpha^2 G^2 \\ &\leq \mathbb{E} [\|\theta^* - \theta_t\|_2^2] - 2\alpha(1-\gamma)\mathbb{E} [\|V_{\theta^*} - V_{\theta_t}\|_D^2] + \alpha^2 (5 + 6\tau^{\text{mix}}(\alpha)) G^2. \end{aligned} \quad (22)$$

The first inequality follows from Lemma 8. The second follows from Lemma 11, which in the case of a constant step-size  $\alpha$  shows  $\mathbb{E}[\alpha\zeta_t(\theta_t)] \leq G^2(4 + 6\tau^{\text{mix}}(\alpha))\alpha^2$ . Notice that bias in the gradient enters into the analysis as if by scaling the magnitude of the noise in gradient evaluations by a factor of the mixing time. From this decomposition, parts (a) and (b) of Theorem 3 follow by essentially copying the proof of Theorem 2. Similar, but messier, inequalities hold for any decaying step-size sequence, which allows us to establish part (c).

#### 8.3.1 Error decomposition under Projected TD

The next lemma establishes a recursion for the error under projected TD(0) that hold for each sample path.

**Lemma 8.** *With probability 1, for every  $t \in \mathbb{N}_0$ ,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.$$

*Proof.* From the projected TD(0) recursion in Equation (20), for any  $t \in \mathbb{N}_0$ ,

$$\begin{aligned} \|\theta^* - \theta_{t+1}\|_2^2 &= \|\theta^* - \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t))\|_2^2 \\ &= \|\Pi_{2,R}(\theta^*) - \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t))\|_2 \\ &\leq \|\theta^* - \theta_t - \alpha_t g_t(\theta_t)\|_2^2 \\ &= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^\top (\theta^* - \theta_t) + \alpha_t^2 \|g_t(\theta_t)\|_2^2 \\ &\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t g_t(\theta_t)^\top (\theta^* - \theta_t) + \alpha_t^2 G^2. \\ &= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t \bar{g}(\theta_t)^\top (\theta^* - \theta_t) + 2\alpha_t \zeta_t(\theta_t) + \alpha_t^2 G^2. \\ &\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t \zeta_t(\theta_t) + \alpha_t^2 G^2. \end{aligned}$$

The first inequality used that orthogonal projection operators onto a convex set are non-expansive<sup>10</sup>, the second used Lemma 6 together with the fact that  $\|\theta_t\|_2 \leq R$  due to projection, and the third used Lemma 3.  $\square$

By taking expectation of both sides, this inequality could be used to produce bounds in the same manner as in the previous section, except that in general  $\mathbb{E}[\zeta_t(\theta_t)] \neq 0$  due to bias in the gradient evaluations.

---

<sup>10</sup>Let  $\mathcal{P}_{\mathcal{C}}(x) = \arg \min_{x' \in \mathcal{C}} \|x' - x\|$  denote the projection operator onto a closed, non-empty, convex set  $\mathcal{C} \subset \mathbb{R}^d$ . Then  $\|\mathcal{P}_{\mathcal{C}}(x) - \mathcal{P}_{\mathcal{C}}(y)\| \leq \|x - y\|$  for all vectors  $x$  and  $y$ .

### 8.3.2 Information-theoretic techniques for controlling the gradient bias

The uniform mixing condition in Assumption 1 can be used in conjunction with some information theoretic inequalities to control the magnitude of the gradient bias. This section presents a general lemma, which is the key to this analysis. We start by reviewing some important properties of information-measures.

**Information theory background.** The total-variation distance between two probability measures is a special case of the more general  $f$ -divergence defined as

$$d_f(P||Q) = \int f\left(\frac{dP}{dQ}\right) dQ,$$

where  $f$  is a convex function such that  $f(1) = 0$ . By choosing  $f(x) = |x - 1|/2$ , one recovers the total-variation distance. A choice of  $f(x) = x \log(x)$  yields the Kullback-Leibler divergence. This yields a generalization of the mutual information between two random variables  $X$  and  $Y$ . The  $f$ -information between  $X$  and  $Y$  is the  $f$ -divergence between their joint distribution and the product of their marginals:

$$I_f(X, Y) = d_f(\mathbb{P}(X = \cdot, Y = \cdot), \mathbb{P}(X = \cdot) \otimes \mathbb{P}(Y = \cdot)).$$

This measure satisfies several nice properties. By definition it is symmetric, so  $I_f(X, Y) = I_f(Y, X)$ . It can be expressed in terms of the expected divergence between conditional distributions:

$$I_f(X, Y) = \sum_x \mathbb{P}(X = x) d_f(\mathbb{P}(Y = \cdot | X = x), \mathbb{P}(Y = \cdot)). \quad (23)$$

Finally, it satisfies the following data-processing inequality. If  $X \rightarrow Y \rightarrow Z$  forms a Markov chain, then

$$I_f(X, Z) \leq I_f(X, Y).$$

Here, we use the notation  $X \rightarrow Y \rightarrow Z$ , which is standard in information theory and the study of graphical models, to indicate that the random variables  $Z$  and  $X$  are independent conditioned on  $Y$ . Note that by symmetry we also have  $I_f(X, Z) \leq I_f(Y, Z)$ . To use these results in conjunction with Assumption 1, we can specialize to total-variation distance ( $d_{\text{TV}}$ ) and total-variation mutual information ( $I_{\text{TV}}$ ) using  $f(x) = |x - 1|/2$ . The total-variation is especially useful for our purposes because of the following variational representation.

$$d_{\text{TV}}(P, Q) = \sup_{v: \|v\|_\infty \leq \frac{1}{2}} \left| \int v dP - \int v dQ \right|. \quad (24)$$

In particular, if  $P$  and  $Q$  are close in total-variation distance, then the expected value of any bounded function under  $P$  will be close to that under  $Q$ .

**Information theoretic control of coupling.** With this background in place, we are ready to establish a general lemma, which is central to our analysis. We use  $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$  to denote the supremum norm of a function  $f : \mathcal{X} \rightarrow \mathbb{R}$ .

**Lemma 9** (Control of coupling). *Consider two random variables  $X$  and  $Y$  such that*

$$X \rightarrow s_t \rightarrow s_{t+\tau} \rightarrow Y$$

*for some fixed  $t \in \{0, 1, 2, \dots\}$  and  $\tau > 0$ . Assume the Markov chain mixes uniformly, as stated in Assumption 1. Let  $X'$  and  $Y'$  denote independent copies drawn from the marginal distributions of  $X$  and  $Y$ , so  $\mathbb{P}(X' = \cdot, Y' = \cdot) = \mathbb{P}(X = \cdot) \otimes \mathbb{P}(Y = \cdot)$ . Then, for any bounded function  $v$ ,*

$$|\mathbb{E}[v(X, Y)] - \mathbb{E}[v(X', Y')]| \leq 2\|v\|_\infty(m\rho^\tau).$$

*Proof.* Let  $P = \mathbb{P}(X \in \cdot, Y \in \cdot)$  denote the joint distribution of  $X$  and  $Y$  and  $Q = \mathbb{P}(X \in \cdot) \otimes \mathbb{P}(Y \in \cdot)$  denote the product of the marginal distributions. Let  $h = \frac{v}{2\|v\|_\infty}$ , which is the function  $v$  rescaled to take values in  $[-1/2, 1/2]$ . Then, by Equation (24)

$$\mathbb{E}[h(X, Y)] - \mathbb{E}[h(X', Y')] = \int h dP - \int h dQ \leq d_{\text{TV}}(P, Q) = I_{\text{TV}}(X, Y),$$

where the last equality uses the definition of the total-variation mutual information,  $I_{\text{TV}}$ . Then,

$$\begin{aligned} I_{\text{TV}}(X, Y) &\leq I_{\text{TV}}(s_t, s_{t+\tau}) = \sum_{s \in \mathcal{S}} \mathbb{P}(s_t = s) d_{\text{TV}}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \mathbb{P}(s_{t+\tau} = \cdot)) \\ &\leq \sup_{s \in \mathcal{S}} d_{\text{TV}}(\mathbb{P}(s_{t+\tau} = \cdot | s_t = s), \pi) \\ &\leq m\rho^\tau, \end{aligned}$$

where the three steps follow, respectively, from the data-processing inequality, the property in Equation (23), the stationarity of the Markov chain, and the uniform mixing condition in Assumption 1. Combining these steps gives

$$|\mathbb{E}[v(X, Y)] - \mathbb{E}[v(X', Y')]| \leq 2\|v\|_\infty I_{\text{TV}}(X, Y) \leq 2\|v\|_\infty m\rho^\tau.$$

□

### 8.3.3 Bounding the gradient bias.

We are now ready to bound the expected gradient error  $\mathbb{E}[\zeta_t(\theta_t)]$ . First, we establish some basic regularity properties of the function  $\zeta_t(\cdot)$ .

**Lemma 10** (Gradient error is bounded and Lipschitz). *With probability 1,*

$$|\zeta_t(\theta)| \leq 2G^2 \quad \text{for all } \theta \in \Theta_R$$

and

$$|\zeta_t(\theta) - \zeta_t(\theta')| \leq 6G \|(\theta - \theta')\|_2 \quad \text{for all } \theta, \theta' \in \Theta_R.$$

*Proof.* The result follows from a straightforward application of the bounds  $\|g_t(\theta)\|_2 \leq G$  and  $\|\theta\|_2 \leq R \leq G/2$ , which hold for each  $\theta \in \Theta_R$ . A full derivation is given in Appendix A.3. □

We now use Lemmas 9 and 10 to establish a bound on the expected gradient error.

**Lemma 11** (Bound on gradient bias). *Consider a non-increasing step-size sequence,  $\alpha_0 \geq \alpha_1 \dots \geq \alpha_T$ . Fix any  $t < T$ , and set  $t^* \equiv \max\{0, t - \tau^{\text{mix}}(\alpha_T)\}$ . Then,*

$$\mathbb{E}[\zeta_t(\theta_t)] \leq G^2 (4 + 6\tau^{\text{mix}}(\alpha_T)) \alpha_{t^*}.$$

The following bound also holds:

$$\mathbb{E}[\zeta_t(\theta_t)] \leq 6G^2 \sum_{i=0}^{t-1} \alpha_i.$$

*Proof.* We break the proof down into three steps.

Step 1: Relate  $\zeta_t(\theta_t)$  and  $\zeta_t(\theta_{t-\tau})$ .

Note that for any  $i \in \mathbb{N}_0$ ,

$$\|\theta_{i+1} - \theta_i\|_2 = \|\Pi_{2,R}(\theta_i + \alpha_i g_i(\theta_i)) - \Pi_{2,R}(\theta_i)\|_2 \leq \|\theta_i + \alpha_i g_i(\theta_i) - \theta_i\|_2 = \alpha_i \|g_i(\theta_i)\|_2 \leq \alpha_i G.$$

Therefore,

$$\|\theta_t - \theta_{t-\tau}\|_2 \leq \sum_{i=t-\tau}^{t-1} \|\theta_{i+1} - \theta_i\|_2 \leq G \sum_{i=t-\tau}^{t-1} \alpha_i.$$

Applying Lemma 10, we conclude

$$\zeta_t(\theta_t) \leq \zeta_t(\theta_{t-\tau}) + 6G^2 \sum_{i=t-\tau}^{t-1} \alpha_i \quad \text{for all } \tau \in \{0, \dots, t\}. \quad (25)$$

Step 2: Bound  $\mathbb{E}[\zeta_t(\theta_{t-\tau})]$  using Lemma 9.

Recall that the gradient  $g_t(\theta)$  depends implicitly on the observed tuple  $O_t = (s_t, \mathcal{R}(s_t, s_{t+1}), s_{t+1})$ . Let us overload notation to make this statistical dependency more transparent. Put

$$g(\theta, O_t) := g_t(\theta) = \left( r_t + \gamma \phi(s_{t+1})^\top \theta - \phi(s_t)^\top \theta \right) \phi(s_t) \quad \theta \in \Theta_R$$

and

$$\zeta(\theta, O_t) := \zeta_t(\theta) = (g(\theta, O_t) - \bar{g}(\theta))^\top (\theta - \theta^*) \quad \theta \in \Theta_R.$$

We have defined  $\bar{g} : \Theta_R \rightarrow \mathbb{R}^d$  as  $\bar{g}(\theta) = \mathbb{E}[g(\theta, O_t)]$  for all  $\theta \in \Theta_R$ , where this expectation integrates over the marginal distribution of  $O_t$ . Then, by definition, for any fixed (non-random)  $\theta \in \Theta_R$ ,

$$\mathbb{E}[\zeta(\theta, O_t)] = (\mathbb{E}[g(\theta, O_t)] - \bar{g}(\theta))^\top (\theta - \theta^*) = 0.$$

Since  $\theta_0 \in \Theta_R$  is non-random, it follows immediately that

$$\mathbb{E}[\zeta(\theta_0, O_t)] = 0. \quad (26)$$

We use Lemma 10 to bound  $\mathbb{E}[\zeta_t(\theta_{t-\tau}, O_t)]$ . First, consider random variables  $\theta'_{t-\tau}$  and  $O'_t$  drawn independently from the marginal distributions of  $\theta_{t-\tau}$  and  $O_t$ , so  $\mathbb{P}(\theta'_{t-\tau} = \cdot, O'_t = \cdot) = \mathbb{P}(\theta_{t-\tau} = \cdot) \otimes \mathbb{P}(O_t = \cdot)$ . Then  $\mathbb{E}[\zeta(\theta'_{t-\tau}, O'_t)] = \mathbb{E}[\mathbb{E}[\zeta(\theta'_{t-\tau}, O'_t) | \theta'_{t-\tau}]] = 0$ . Since  $|\zeta(\theta, O_t)| \leq 2G^2$  for all  $\theta \in \Theta_R$  by Lemma 11 and  $\theta_{t-\tau} \rightarrow s_{t-\tau} \rightarrow s_t \rightarrow O_t$  forms a Markov chain, applying Lemma 10 gives

$$\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)] \leq 2(2G^2)(m\rho^\tau) = 4G^2m\rho^\tau. \quad (27)$$

Step 3: Combine terms.

The second claim follows immediately from Equation (25) together with Equation (26). We focus on establishing the first claim. Taking the expectation of Equation (25) implies

$$\mathbb{E}[\zeta_t(\theta_t)] \leq \mathbb{E}[\zeta_t(\theta_{t-\tau})] + 6G^2\tau\alpha_{t-\tau} \quad \forall \tau \in \{0, \dots, t\}.$$

For  $t \leq \tau^{\text{mix}}(\alpha_T)$ , choosing  $\tau = t$  gives

$$\mathbb{E}[\zeta_t(\theta_t)] \leq \underbrace{\mathbb{E}[\zeta_t(\theta_0)]}_{=0} + 6G^2t\alpha_0 \leq 6G^2\tau^{\text{mix}}(\alpha_T)\alpha_0.$$

For  $t > \tau^{\text{mix}}(\alpha_T)$ , choosing  $\tau = \tau_0 \equiv \tau^{\text{mix}}(\alpha_T)$  gives

$$\mathbb{E}[\zeta_t(\theta_t)] \leq 4G^2m\rho^{\tau_0} + 6G^2\tau_0\alpha_{t-\tau_0} \leq 4G^2\alpha_T + 6G^2\tau_0\alpha_{t-\tau} \leq G^2(4 + 6\tau_0)\alpha_{t-\tau_0}.$$

where the second inequality used that  $m\rho^{\tau_0} \leq \alpha_T$  by the definition of the mixing time  $\tau_0 \equiv \tau^{\text{mix}}(\alpha_T)$  and the second inequality uses that step-sizes are non-increasing.  $\square$

### 8.3.4 Completing the proof of Theorem 3

Combining Lemmas 8 and 10 gives the error decomposition in Equation 22 for the case of a constant step-size. As noted at the beginning of this subsection, from this decomposition, parts (a) and (b) of Theorem 3 can be established by essentially copying the proof of Theorem 2. For completeness, this is included in Appendix A. For part (c), we closely follow analysis of SGD with decaying step-sizes presented in Lacoste-Julien et al. [2012]. However, some headache is introduced because Lemma 11 includes terms of the form  $\alpha_{t-\tau^{\text{mix}}(\alpha_T)}$  instead of the typical  $\alpha_t$  terms present in analyses of SGD. A complete proof of part (c) is given in Appendix A as well.

## 9 Extension to TD with eligibility traces

This section extends our analysis to provide finite time guarantees for temporal difference learning *with eligibility traces*. We study a class of algorithms, denoted by  $\text{TD}(\lambda)$  and parameterized by  $\lambda \in [0, 1]$ , that contains as a special case the  $\text{TD}(0)$  algorithm studied in previous sections<sup>11</sup>. For  $\lambda > 0$ , the algorithm maintains an eligibility trace vector, which is a geometric weighted average of the negative gradients at all previously visited states, and makes parameter updates in the direction of the eligibility vector rather than the negative gradient. Eligibility traces sometimes provide substantial performance improvements in practice [Sutton and Barto, 1998]. Unfortunately, they also introduce subtle dependency issues that complicate theoretical analysis; to our knowledge, this section provides the *first* non-asymptotic analysis  $\text{TD}(\lambda)$ .

Our analysis focuses on the Markov chain observation model studied in the previous section and we mirror the technical assumptions used there. In particular, we assume that the Markov chain is stationary and mixes at a uniform geometric rate (Assumption 1). As before, for tractability, we study a projected variant of  $\text{TD}(\lambda)$ .

### 9.1 Projected $\text{TD}(\lambda)$ algorithm

$\text{TD}(\lambda)$  makes a simple, but highly consequential, modification to  $\text{TD}(0)$ . Pseudo-code for the algorithm is presented below in Algorithm 2. As with  $\text{TD}(0)$ , at each time-step  $t$  it observes a tuple  $(s_t, r_t = \mathcal{R}(s_t, s_{t+1}), s_{t+1})$  and computes the TD error  $\delta_t(\theta_t) = r_t + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t)$ . However, while  $\text{TD}(0)$  makes an update  $\theta_{t+1} = \theta_t + \alpha_t \delta_t(\theta_t) \phi(s_t)$  in the direction of the feature vector at the current state,  $\text{TD}(\lambda)$  makes the update  $\theta_{t+1} = \theta_t + \alpha_t \delta_t(\theta_t) z_{0:t}$ . The vector  $z_{0:t} = \sum_{k=0}^t (\gamma \lambda)^k \phi(s_{t-k})$  is called the eligibility trace which is updated incrementally as shown below in Algorithm 2. As the name suggests, the components of  $z_{0:t}$  roughly capture the extent to which each feature is eligible for receiving credit or blame for an observed TD error [Sutton and Barto, 1998, Seijen and Sutton, 2014].

---

**Algorithm 2:** Projected  $\text{TD}(\lambda)$  with linear function approximation

---

```

Input : radius  $R$ , initial guess  $\{\theta_0 : \|\theta_0\|_2 \leq R\}$ , and step-size sequence  $\{\alpha_t\}_{t \in \mathbb{N}}$ 
Initialize:  $\theta_0 \leftarrow \theta_0$ ,  $z_{-1} = 0$ ,  $\lambda \in [0, 1]$ .
for  $t = 0, 1, \dots$  do
    Observe tuple:  $O_t = (s_t, r_t, s_{t+1})$ 
    Get TD error:  $\delta_t(\theta_t) = r_t + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t)$  /* sample Bellman error */
    Update eligibility trace:  $z_{0:t} = (\gamma \lambda) z_{0:t-1} + \phi(s_t)$  /* Geometric weighting */
    Compute update direction:  $x_t(\theta_t, z_{0:t}) = \delta_t(\theta_t) z_{0:t}$ 
    Take a projected update step:  $\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))$  /*  $\alpha_t$ :step-size */
    Update averaged iterate:  $\bar{\theta}_{t+1} \leftarrow \left( \frac{t}{t+1} \right) \bar{\theta}_t + \left( \frac{1}{t+1} \right) \theta_{t+1}$  /*  $\bar{\theta}_{t+1} = \frac{1}{t+1} \sum_{\ell=1}^{t+1} \theta_\ell$  */
end

```

---

<sup>11</sup> $\text{TD}(0)$  corresponds to  $\lambda = 0$ .

Some new notation in Algorithm 2 should be highlighted. We use  $x_t(\theta, z_{0:t}) = \delta_t(\theta)z_{0:t}$  to denote the update to the parameter vector  $\theta$  at time  $t$ . This plays a role analogous to the negative gradient  $g_t(\theta)$  in TD(0).

## 9.2 Limiting behavior of TD( $\lambda$ )

We now review results on the asymptotic convergence of TD( $\lambda$ ) due to [Tsitsiklis and Van Roy \[1997\]](#). This provides the foundation of our finite time analysis and also offers insight into how the algorithm differs from TD(0).

Before giving any results, let us note that just as the true value function  $V_\mu(\cdot)$  is the unique solution to Bellman's fixed point equation  $V_\mu = T_\mu V_\mu$ , it is also the unique solution to a  $k$ -step Bellman equation  $V_\mu = T_\mu^{(k)}V_\mu$ . This can be written equivalently as

$$V_\mu(s) = \mathbb{E} \left[ \sum_{t=0}^k \gamma^t \mathcal{R}(s_t) + \gamma^{k+1} V(s_{k+1}) \mid s_0 = s \right] \quad \forall s \in S,$$

where the expectation is over states sampled when policy  $\mu$  is applied to the MDP. The asymptotic properties of TD( $\lambda$ ) are closely tied to a geometrically weighted version of the  $k$ -step Bellman equations described above. Define the averaged Bellman operator

$$(T_\mu^{(\lambda)} V)(s) = (1 - \lambda) \sum_{k=0}^{\infty} \lambda^k \mathbb{E} \left[ \sum_{t=0}^k \gamma^t \mathcal{R}(s_t) + \gamma^{k+1} V(s_{k+1}) \mid s_0 = s \right]. \quad (28)$$

One interesting interpretation of this equation is as a  $k$ -step Bellman equation, but where the horizon  $k$  itself is a random geometrically distributed random variable.

[Tsitsiklis and Van Roy \[1997\]](#) showed that under appropriate technical conditions, the approximate value function  $V_{\theta_t} = \Phi\theta_t$  estimated by TD( $\lambda$ ) converges almost surely to the unique solution,  $\theta^*$  of the projected fixed point equation

$$\Phi\theta = \Pi_D T_\mu^{(\lambda)} \Phi\theta.$$

TD( $\lambda$ ) is then interpreted as a stochastic approximation scheme for solving this fixed point equation. The existence and uniqueness of such a fixed point  $\Phi\theta^*$  is implied by the following lemma, which shows that  $\Pi_D T^\lambda(\cdot)$  is a contraction operator with respect to the steady-state weighted norm  $\|\cdot\|_D$ .

**Lemma 12.** [[Tsitsiklis and Van Roy \[1997\]](#)]  $\Pi_D T_\mu^{(\lambda)}(\cdot)$  is a contraction with respect to  $\|\cdot\|_D$  with modulus

$$\kappa = \frac{\gamma(1 - \lambda)}{1 - \gamma\lambda} \leq \gamma < 1.$$

As with TD(0), the limiting value function under TD( $\lambda$ ) comes with some competitive guarantees. A short argument using Lemma 12 shows

$$\|V_{\theta^*} - V_\mu\|_D \leq \frac{1}{\sqrt{1 - \kappa^2}} \|\Pi_D V_\mu - V_\mu\|_D. \quad (29)$$

See for example Chapter 6 of [Bertsekas \[2012\]](#) for a proof. It is important to note the distinction between the convergence guarantee results for TD( $\lambda$ ) and TD(0) in terms of the contraction factors. The contraction factor  $\kappa$  is always less than  $\gamma$ , the contraction factor under TD(0). In addition, as  $\lambda \rightarrow 1$ ,  $\kappa \rightarrow 0$  implying that the limit point of TD( $\lambda$ ) for large enough  $\lambda$  will be arbitrarily close to  $\Pi_D V_\mu$ , which minimizes the mean-square error in value predictions among all value functions representable by the features. This calculation suggests a choice of  $\lambda = 1$  will offer the best performance. However, the *rate of convergence* also depends on  $\lambda$ , and may degrade as  $\lambda$  grows. Disentangling such issues requires also a careful study of the statistical efficiency of TD( $\lambda$ ), which we undertake in the following subsection.

### 9.3 Finite time bounds for Projected TD( $\lambda$ )

Following Section 8, we establish three finite time bounds on the performance of the Projected TD( $\lambda$ ) algorithm. The first bound in part (a) does not depend on any special regularity of the problem instance but gives a comparatively slow convergence rate of  $\tilde{\mathcal{O}}(1/\sqrt{T})$ . It applies with the robust (problem independent) and aggressive step-size of  $1/\sqrt{T}$ . Part (b) illustrates the exponential rate of convergence to within some radius of the TD( $\lambda$ ) fixed-point for sufficiently small step-sizes. Part (c) attains an improved dependence on  $T$  of  $\tilde{\mathcal{O}}(1/T)$ , but the step-size sequence requires knowledge of the minimum eigenvalue  $\omega$  of  $\Sigma$ .

Compared to the results for TD(0), our bounds depend on a slightly different definition of the mixing time that takes into account the geometric weighting in the eligibility trace term. Define

$$\tau_\lambda^{\text{mix}}(\epsilon) = \max\{\tau^{\text{MC}}(\epsilon), \tau^{\text{Algo}}(\epsilon)\}, \quad (30)$$

where we denote  $\tau^{\text{MC}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}$  and  $\tau^{\text{Algo}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid (\gamma\lambda)^t \leq \epsilon\}$ . As we show next, this definition of mixing time enables compact bounds for convergence rates of TD( $\lambda$ ).

**Theorem 4.** Suppose the Projected TD( $\lambda$ ) algorithm is applied with parameter  $R \geq \|\theta^*\|_2$  under the Markov chain observation model with Assumption 1. Set  $B = \frac{(r_{\max}+2R)}{(1-\gamma\lambda)}$ . Then the following claims hold.

(a) With a constant step-size  $\alpha_t = \alpha_0 = 1/\sqrt{T}$ ,

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2 \left( 13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T}) \right)}{2\sqrt{T}(1-\kappa)}.$$

(b) With a constant step-size  $\alpha_t = \alpha_0 < 1/(2\omega(1-\kappa))$  and  $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$ ,

$$\mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] \leq \left( e^{-2\alpha_0(1-\kappa)\omega T} \right) \|\theta^* - \theta_0\|_2^2 + \alpha_0 \left( \frac{B^2 (13 + 24\tau_\lambda^{\text{mix}}(\alpha_0))}{2(1-\kappa)\omega} \right).$$

(c) With a decaying step-size  $\alpha_t = 1/(\omega(t+1)(1-\kappa))$ ,

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{B^2 (13 + 52\tau_\lambda^{\text{mix}}(\alpha_T))}{T(1-\kappa)^2\omega} (1 + \log T).$$

**Remark 2.** As was the case for TD(0), the proof of part (c) also implies a  $\tilde{\mathcal{O}}(1/T)$  convergence rate for the iterate  $\theta_T$  itself. Again, a different weighting of the iterates as shown in Lacoste-Julien et al. [2012] might enable us to eliminate the  $\log T$  term in the numerator of part (c) to give a  $\mathcal{O}(1/T)$  convergence rate. For brevity, we do not pursue this direction.

We now compare the bounds for TD( $\lambda$ ) with that of TD(0) ignoring the constant terms. First, let us look at the results for the constant step-size  $\alpha_t = 1/\sqrt{T}$  in part (a) of Theorems 3 and 4. Approximately, for the TD( $\lambda$ ) case, we have the term  $\frac{B^2}{\sqrt{T}(1-\kappa)}$  vis-a-vis the term  $\frac{G^2}{\sqrt{T}(1-\gamma)}$  for the TD(0) case. A simple argument below clarifies the relationship between these two.

$$\begin{aligned} \frac{B^2}{\sqrt{T}(1-\kappa)} &= \frac{(r_{\max}+2R)^2}{\sqrt{T}(1-\kappa)(1-\gamma\lambda)^2} = \frac{G^2}{\sqrt{T}(1-\kappa)(1-\gamma\lambda)^2} \\ &\geq \frac{G^2}{\sqrt{T}(1-\kappa)(1-\gamma\lambda)} = \frac{G^2}{\sqrt{T}(1-\gamma)}. \end{aligned}$$

As we will see later,  $B$  is an upper bound to the norm of  $x_t(\theta_t, z_{0:t})$ , the update direction for TD( $\lambda$ ). Correspondingly, from Section 8, we know that  $G$  is the upper bound on gradient norm,  $g_t(\theta_t)$  for TD(0). Intuitively, for TD( $\lambda$ ), the bound  $B$  is larger (due to the presence of the eligibility trace term) and more so as  $\lambda \rightarrow 1$ . This dominates any benefit (in terms of statistical efficiency) from a smaller contraction factor,  $\kappa$ .

However, for decaying step-sizes of  $\alpha_t = 1/(\omega(t+1)(1-\kappa))$ , the bounds are qualitatively the same. This follows as the terms that dominate part (c) of Theorems 3 and 4 are equal:

$$\frac{B^2}{T(1-\kappa)^2} = \frac{(r_{\max} + 2R)^2}{T(1-\kappa)^2(1-\gamma\lambda)^2} = \frac{G^2}{T(1-\kappa)^2(1-\gamma\lambda)^2} = \frac{G^2}{T(1-\gamma)^2}.$$

In conclusion, for constant step-sizes—which is often how TD algorithms as used in practice—our bounds establish a faster convergence rate for TD(0) than for TD( $\lambda$ ). Or equivalently, according to our bounds, more data is required to guarantee TD( $\lambda$ ) is close to its limit point. In this context, however, the trade-off we remarked on in Section 9.2 is noteworthy as the fixed point for TD( $\lambda$ ) comes with a better error guarantee.

## 10 Extension: Q-learning for high dimensional Optimal Stopping

So far, this paper has dealt with the problem of approximating the value function of a fixed policy in a computationally and statistically efficient manner. The Q-learning algorithm is one natural extension of temporal-difference learning to control problems, where the goal is to learn an effective policy from data. Although it is widely applied in reinforcement learning, in general Q-learning is unstable and its iterates may oscillate forever. An important exception to this was discovered by Tsitsiklis and Van Roy [1999], who showed that Q-learning converges asymptotically for optimal stopping problems. In this section, we show how the techniques developed in Sections 7 and 8 can be applied *in an identical manner* to give finite time bounds for Q-learning with linear function approximation applied to optimal-stopping problems with high dimensional state spaces. To avoid repetition, we only state key properties satisfied by Q-learning in this setting which establish exactly the same convergence bounds as shown in Theorems 2 and 3.

### 10.1 Problem formulation

The optimal stopping problem is that of determining the time to terminate a process to maximize cumulative expected rewards accrued. Problems of this nature arise naturally in many settings, most notably in the pricing of financial derivatives [Andersen and Broadie, 2004, Haugh and Kogan, 2004, Desai et al., 2012]. We first give a brief formulation for a class of optimal stopping problems. A more detailed exposition can be found in Tsitsiklis and Van Roy [1999], or Chapter 5 of the thesis work of Van Roy [1998].

Consider a discrete-time Markov chain  $\{s_t\}_{t \geq 0}$  with finite state space  $\mathcal{S}$  and unique stationary distribution  $\pi$ . At each time  $t$ , the decision-maker observes the state  $s_t$  and decides whether to stop or continue. Let  $\gamma \in [0, 1)$  denote the discount factor and let  $u(\cdot)$  and  $U(\cdot)$  denote the reward functions associated with *continuation* and *termination* decisions respectively. Let the stopping time  $\tau$  denote the (random) time at which the decision-maker stops. The expected total discounted reward from initial state  $s$  associated with the stopping time  $\tau$  is

$$\mathbb{E} \left[ \sum_{t=0}^{\tau-1} \gamma^t u(s_t) + \gamma^\tau U(s_\tau) \mid s_0 = s \right], \quad (31)$$

where  $U(s_\tau)$  is defined to be zero for  $\tau = \infty$ . We seek an optimal stopping policy, which determines when to stop as a function of the observed states so as to maximize (31).

For any Markov decision process, the optimal state-action value function  $Q^* : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  specifies the expected value to go from choosing an action  $a \in \mathcal{A}$  in a state  $s \in \mathcal{S}$  and following the optimal policy in subsequent states. In optimal stopping problems, there are only two possible actions at every time step: whether to *terminate* or to *continue*. The value of stopping in state  $s$  is just  $U(s)$ , which allows us to simplify notation by only representing the continuation value.

For the remainder of this section, we let  $Q^* : \mathcal{S} \rightarrow \mathbb{R}$  denote the continuation-value function. It can be shown that  $Q^*$  is the unique solution to the Bellman equation  $Q^* = FQ^*$ , where the Bellman operator is given by

$$FQ(s) = u(s) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s) \max \{U(s'), Q(s')\}.$$

Given the optimal continuation values  $Q^*(\cdot)$ , the optimal stopping time is simply

$$\tau^* = \min \{t \mid U(s_t) \geq Q^*(s_t)\}. \quad (32)$$

## 10.2 Q-Learning for high dimensional Optimal Stopping

In principle, one could generate the optimal stopping time using Equation (32) by applying exact dynamic programming algorithms to compute the optimal Q-function. However, such methods are only implementable for small state spaces. To scale to high dimensional state spaces, we consider a feature-based approximation of the optimal continuation value function,  $Q^*$ . We focus on linear function approximation, where  $Q^*(s)$  is approximated as

$$Q^*(s) \approx Q_\theta(s) = \phi(s)^\top \theta,$$

where  $\phi(s) \in \mathbb{R}^d$  is a fixed feature vector for state  $s$  and  $\theta \in \mathbb{R}^d$  is a parameter vector that is shared across states. As shown in Section 2, for a finite state space,  $\mathcal{S} = \{s_1, \dots, s_n\}$ ,  $Q_\theta \in \mathbb{R}^n$  can be expressed compactly as  $Q_\theta = \Phi\theta$ , where  $\Phi \in \mathbb{R}^{n \times d}$  and  $\theta \in \mathbb{R}^d$ . We also assume that the  $d$  features vectors  $\{\phi_k\}_{k=1}^d$ , forming the columns of  $\Phi$  are linearly independent.

We consider the Q-learning approximation scheme in Algorithm 3. The algorithm starts with an initial parameter estimate of  $\theta_0$  and observes a data tuple  $O_t = (s_t, u(s_t), s'_t)$ . This is used to compute the target  $y_t = u(s_t) + \gamma \max\{U(s'_t), Q_{\theta_t}(s'_t)\}$ , which is a sampled version of the  $F(\cdot)$  operator applied to the current Q-function. The next iterate,  $\theta_{t+1}$ , is computed by taking a gradient step with respect to a loss function measuring the distance between  $y_t$  and predicted value-to-go. An important feature of this method is that problem data is generated by the exploratory policy that chooses to continue at all time-steps.

---

### Algorithm 3: Q-Learning for Optimal Stopping problems.

---

```

Input : initial guess  $\theta_0$ , step-size sequence  $\{\alpha_t\}_{t \in \mathbb{N}}$  and radius  $R$ .
Initialize:  $\bar{\theta}_0 \leftarrow \theta_0$ .
for  $t = 0, 1, \dots$  do
    Observe tuple:  $O_t = (s_t, u(s_t), s'_t)$ 
    Define target:  $y_t = u(s_t) + \gamma \max\{U(s'_t), Q_{\theta_t}(s'_t)\}$  /* sample Bellman operator */
    Define loss function:  $\frac{1}{2}(y_t - Q_\theta(s_t))^2$  /* sample Bellman error */
    Compute negative gradient:  $g_t(\theta_t) = -\frac{\partial}{\partial \theta} \frac{1}{2}(y_t - Q_\theta(s_t))^2|_{\theta=\theta_t}$ 
    Take a gradient step:  $\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$  /*  $\alpha_t$ :step-size */
    Update averaged iterate:  $\bar{\theta}_{t+1} \leftarrow \left(\frac{t}{t+1}\right) \bar{\theta}_t + \left(\frac{1}{t+1}\right) \theta_{t+1}$  /*  $\bar{\theta}_{t+1} = \frac{1}{t+1} \sum_{\ell=1}^{t+1} \theta_\ell$  */
end

```

---

## 10.3 Asymptotic guarantees

Similar to the asymptotic results for TD algorithms, [Tsitsiklis and Van Roy \[1999\]](#) show that the variant of Q-learning detailed above in Algorithm 3 converges to the unique solution,  $\theta^*$ , of the projected Bellman equation,

$$\Phi\theta = \Pi_D F\Phi\theta.$$

This results crucially relies on the fact that the projected Bellman operator  $\Pi_D F(\cdot)$  is a contraction with respect to  $\|\cdot\|_D$  with modulus  $\gamma$ . The analogous result for our study of TD(0) was stated in Lemma 2. [Tsitsiklis and Van Roy \[1999\]](#) also give error bounds for the limit of convergence with respect to  $Q^*$ , the optimal Q-function. In particular, it can be shown that

$$\|\Phi\theta^* - Q^*\|_D \leq \frac{1}{\sqrt{1-\gamma^2}} \|\Pi_D Q^* - Q^*\|_D$$

where the left hand side measures the error between the estimated and the optimal Q-function which is upper bounded by the *representational power* of the linear approximation architecture, as given on the right hand side. In particular, if  $Q^*$  can be represented as a linear combination of the feature vectors then there is no approximation error and the algorithm converges to the optimal Q-function. Finally, one can ask whether the stopping times suggested by this approximate continuation value function,  $\Phi\theta^*$ , are effective. Let  $\tilde{\mu}$  be the policy that stops at the first time  $t$  when

$$U(s_t) \geq (\Phi\theta^*)(s_t).$$

Then, for an initial state  $s_0$  drawn from the stationary distribution  $\pi$ ,

$$\mathbb{E}[V^*(s_0)] - \mathbb{E}[V_{\tilde{\mu}}(s_0)] \leq \frac{2}{(1-\gamma)\sqrt{1-\gamma^2}} \|\Pi_D Q^* - Q^*\|_D,$$

where  $V^*$  and  $V_{\tilde{\mu}}$  denote the value functions corresponding, respectively, to the optimal stopping policy the approximate stopping policy  $\mu$ . Again, this error guarantee depends on the choice of feature representation.

## 10.4 Finite time analysis

In this section, we show how our results in Sections 7, 8 for TD(0) and its projected counterpart can be extended, without any modification, to give convergence bounds for the Q-function approximation algorithm described above. To this effect, we highlight that key lemmas that enable our analysis in Sections 7 and 8 also hold in this setting. The contraction property of the  $F(\cdot)$  operator will be crucial to our arguments here. Convergence rates for an i.i.d. noise model, mirroring those established for TD(0) in Theorem 2, can be shown for Algorithm 3. Results for the Markov chain sampling model, mirroring those established for TD(0) in Theorem 3, can be shown for a projected variant of Algorithm 3.

First, we give mathematical expressions for the negative gradient. As a general function of  $\theta$  and tuple  $O_t = (s_t, u(s_t), s'_t)$ , the negative gradient can be written as

$$g_t(\theta) = \left( u(s_t) + \gamma \max \{U(s'_t), \phi(s'_t)^\top \theta\} - \phi(s_t)^\top \theta \right) \phi(s_t). \quad (33)$$

The negative expected gradient, when the tuple  $(s_t, u(s_t), s'_t)$  follows its steady-state behavior, can be written as

$$\bar{g}(\theta) = \sum_{s, s' \in \mathcal{S}} \pi(s) \mathcal{P}(s'|s) \left( u(s) + \gamma \max \{U(s'), \phi(s')^\top \theta\} - \phi(s)^\top \theta \right) \phi(s).$$

Additionally, using  $\sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s) (u(s) + \gamma \max \{U(s'), \phi(s')^\top \theta\}) = (F\Phi\theta)(s)$ , it is easy to show

$$\bar{g}(\theta) = \Phi^\top D(F\Phi\theta - \Phi\theta).$$

Note the close similarity of this expression with its counterparts for TD learning (see Section 3 and Appendix B); the only difference is that the appropriate Bellman operator(s) for TD learning,  $T_\mu(\cdot)$ , has been replaced with the appropriate Bellman operator  $F(\cdot)$  for this optimal stopping problem.

### 10.4.1 Analysis with i.i.d. noise

In this section, we show how to analyze the Q-learning algorithm under an i.i.d. observation model, where the random tuples observed by the algorithm are sampled i.i.d. from the stationary distribution of the Markov process. All our ideas follow the presentation in Section 7, a careful understanding of which reveals that Lemmas 3 and 5 form the backbone of our results. Recall that Lemma 3 establishes how, at any iterate  $\theta$ , TD updates point in the descent direction of  $\|\theta^* - \theta\|_2^2$ . Lemma 5 bounds the expected norm of the stochastic gradient, thus giving a control over system noise.

In Lemmas 13 and 14, given below, we show how exactly the same results also hold for the Q-function approximation algorithm under the i.i.d. sampling model. With these two key lemmas, convergence bounds shown in Theorem 2 follows by repeating the analysis in Section 7.

**Lemma 13.** [Tsitsiklis and Van Roy [1999]] Let  $V_{\theta^*}$  be the unique fixed point of  $\Pi_D F(\cdot)$ , i.e.  $V_{\theta^*} = \Pi_D F V_{\theta^*}$ . Then, for any  $\theta \in \mathbb{R}^d$ ,

$$(\theta^* - \theta)^\top \bar{g}(\theta) \geq (1 - \gamma) \|V_{\theta^*} - V_\theta\|_D^2.$$

*Proof.* This property is a consequence of the fact that  $\Pi_D F(\cdot)$  is a contraction with respect to  $\|\cdot\|_D$  with modulus  $\gamma$ . It was established by Tsitsiklis and Van Roy [1999] in the process of proving their Lemma 8. For completeness, we provide a standalone proof in Appendix C.  $\square$

**Lemma 14.** We use notation and proof strategy mirroring the proof of Lemma 5. For any fixed  $\theta \in \mathbb{R}^d$ ,  $\mathbb{E}\|g_t(\theta)\|_2^2 \leq 2\sigma^2 + 8\|V_\theta - V_{\theta^*}\|_D^2$  where  $\sigma^2 = \mathbb{E}\|g_t(\theta^*)\|_2^2$ .

*Proof.* For brevity of notation, set  $\phi = \phi(s_t)$ ,  $\phi' = \phi(s'_t)$  and  $U' = U(s')$ . Define  $\xi = (\theta^* - \theta)^\top \phi$  and  $\xi' = (\theta^* - \theta)^\top \phi'$ . By stationarity  $\xi$  and  $\xi'$  have the same marginal distribution and  $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$ . Using the formula for  $g_t(\theta)$  in Equation (33), we have

$$\begin{aligned} \mathbb{E}\|g_t(\theta)\|_2^2 &\leq 2\mathbb{E}\|g_t(\theta^*)\|_2^2 + 2\mathbb{E}\|g_t(\theta) - g_t(\theta^*)\|_2^2 \\ &= 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(\phi^\top(\theta^* - \theta) - \gamma\left[\max(U', \phi'^\top\theta^*) - \max(U', \phi^\top\theta)\right]\right)\right\|_2^2\right] \\ &\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(|\phi^\top(\theta^* - \theta)| + \gamma\left|\max(U', \phi'^\top\theta^*) - \max(U', \phi^\top\theta)\right|\right)\right\|_2^2\right] \\ &\leq 2\sigma^2 + 2\mathbb{E}\left[\left\|\phi\left(|\phi^\top(\theta^* - \theta)| + \gamma|\phi'^\top(\theta^* - \theta)|\right)\right\|_2^2\right] \\ &\leq 2\sigma^2 + 2\mathbb{E}[|\xi + \gamma\xi'|^2] \\ &\leq 2\sigma^2 + 4(\mathbb{E}|\xi|^2 + \gamma^2\mathbb{E}|\xi'|^2) \\ &= 2\sigma^2 + 4(1 - \gamma^2)\|V_\theta - V_{\theta^*}\|_D^2 \leq 2\sigma^2 + 8\|V_\theta - V_{\theta^*}\|_D^2, \end{aligned} \tag{34}$$

where we used the assumption that features are normalized so that  $\|\phi\|_2^2 \leq 1$  almost surely. Additionally, in going to Equation (34), we used that  $|\max(c_1, c_3) - \max(c_2, c_3)| \leq |c_1 - c_2|$  for any scalars  $c_1, c_2$  and  $c_3$ .  $\square$

#### 10.4.2 Analysis under the Markov chain model

Analogous to Section 8, we analyze a projected variant of Algorithm 3 under the Markov chain sampling model. Let  $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$ . Starting with an initial guess of  $\theta_0 \in \Theta_R$ , the algorithm updates to the next iterate by taking a gradient step followed by projection onto  $\Theta_R$ , so iterates satisfy the stochastic recursion  $\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t g_t(\theta_t))$ . We make the similar structural assumptions to those in Section 8. In particular, assume the feature vectors and the continuation, termination rewards to be uniformly bounded, with  $\|\phi(s)\|_2 \leq 1$  and  $\max\{|u(s)|, |U(s)|\} \leq r_{\max}$  for all  $s \in \mathcal{S}$ . We assume  $r_{\max} \leq R$ , which can always be ensured by rescaling rewards or the projection radius.

We first show a uniform bound on the gradient norm.

**Lemma 15.** Define  $G = (r_{\max} + 2R)$ . With probability 1,  $\|g_t(\theta)\|_2 \leq G$  for all  $\theta \in \Theta_R$ .

*Proof.* We start with the mathematical expression for the stochastic gradient,

$$g_t(\theta) = \left(u(s_t) + \gamma \max\{U(s'_t), \phi(s'_t)^\top \theta\} - \phi(s_t)^\top \theta\right) \phi(s_t).$$

As  $r_{\max} \leq R$ , we have:  $\max\{U(s'_t), \phi(s'_t)^\top \theta\} \leq \max\{U(s'_t), \|\phi(s'_t)\|_2 \|\theta\|_2\} \leq R$ . Then,

$$\begin{aligned} \|g_t(\theta)\|_2^2 &= (u(s_t) + \gamma \max\{U(s'_t), \phi(s'_t)^\top \theta\} - \phi(s_t)^\top \theta)^2 \|\phi(s_t)\|^2 \\ &\leq (r_{\max} + \gamma R - \|\phi(s_t)\|_2)^2 \\ &\leq (r_{\max} + \gamma R + \|\phi(s_t)\|_2 \|\theta\|_2)^2 \leq (r_{\max} + 2R)^2 = G^2. \end{aligned}$$

We used here that the basis vectors are normalized,  $\|\phi(s_t)\|_2 \leq 1$  for all  $t$ .  $\square$

If we assume the Markov process  $(s_0, s_1, \dots)$  satisfies Assumption 1, then Lemma 15 paves the way to show exactly the same convergence bounds as given in Theorem 3. For this, we refer the readers to Section 8 and Appendix A, where we show all the key lemmas and a detailed proof of Theorem 3. One can mirror the same proof, using Lemmas 13 and 15 in place of Lemmas 3 and 11, which apply to TD(0). In particular, note that we can use Lemma 15 along with some basic algebraic inequalities to show the gradient bias,  $\zeta_t(\theta)$ , to be Lipschitz and bounded. This, along with the information-theoretic arguments of Lemma 9 enables the exact same upper bound on the gradient bias as shown in Lemma 11. Combining these with standard proof techniques for SGD [Lacoste-Julien et al., 2012, Nemirovski et al., 2009] shows the convergence bounds for Q-learning.

## 11 Conclusions

In this paper we provide a simple finite time analysis of a foundational and widely used algorithm known as temporal difference learning. Although asymptotic convergence guarantees for the TD method were previously known, characterizing its data efficiency stands as an important open problem. Our work makes a substantial advance in this direction by providing a number of explicit finite time bounds for TD, including in the much more complicated case where data is generated from a single trajectory of a Markov chain. Our analysis inherits the simplicity of and elegance enjoyed by SGD analysis and can gracefully extend to different variants of TD, for example TD learning with eligibility traces ( $\text{TD}(\lambda)$ ) and Q-function approximation for optimal stopping problems. Owing to the close connection with SGD, we believe that optimization researchers can further build on our techniques to develop principled improvements to TD.

There are a number of research directions one can take to extend our work. First, we use a projection step for analysis under the Markov chain model, a choice we borrowed from the optimization literature to simplify our analysis. It will be interesting to find alternative ways to add regularity to the TD algorithm and establish similar convergence results; we think analysis without the projection step is possible if one can show that the iterates remain bounded under additional regularity conditions. Second, the  $\tilde{\mathcal{O}}(1/T)$  convergence rate we showed used step-sizes which crucially depends on the minimum eigenvalue  $\omega$  of the feature covariance matrix, which would need to be estimated from samples. While such results are common in optimization for strongly convex functions, very recently Lakshminarayanan and Szepesvári [2018] showed TD(0) with iterate averaging and *universal constant step-sizes* can attain an  $\tilde{\mathcal{O}}(1/T)$  convergence rate in the i.i.d. sampling model. Extending our analysis for problem independent, robust step-size choices is a research direction worth pursuing.

## References

- Leif Andersen and Mark Broadie. Primal-dual simulation algorithm for pricing multidimensional american options. *Management Science*, 50(9):1222–1234, 2004.
- András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.
- Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate  $o(1/n)$ . In *Advances in neural information processing systems 26*, pages 773–781, 2013.
- Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pages 30–37. 1995.
- Albert Benveniste, Michel Métivier, and Pierre Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

- Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 2. Athena Scientific, 2012.
- Dimitri P Bertsekas and Steven E Shreve. Stochastic optimal control: the discrete time case. 1978.
- Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.
- Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.
- Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.
- Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1-3):33–57, 1996.
- Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.
- Gal Dalal, Balzs Sznyi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td(0) with function approximation, 2018a. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16392>.
- Gal Dalal, Gugan Thoppe, Balzs Sznyi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Proceedings of the 31st Conference On Learning Theory*, pages 1199–1233, 2018b.
- Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, 15(1):809–883, 2014.
- Daniela Pucci De Farias and Benjamin Van Roy. The linear programming approach to approximate dynamic programming. *Operations research*, 51(6):850–865, 2003.
- Vijay V Desai, Vivek F Farias, and Ciamac C Moallemi. Pathwise optimization for optimal stopping problems. *Management Science*, 58(12):2292–2308, 2012.
- Adithya M Devraj and Sean P Meyn. Zap q-learning. In *Advances in Neural Information Processing Systems 30*, pages 2235–2244, 2017.
- Mohammad Ghavamzadeh, Alessandro Lazaric, Odalric Maillard, and Rémi Munos. LSTD with Random Projections. In *Advances in Neural Information Processing Systems 23*, pages 721–729, 2010.
- David A Goldberg and Yilun Chen. Beating the curse of dimensionality in options pricing and optimal stopping. *arXiv preprint arXiv:1807.02227*, 2018.
- László Györfi and Harro Walk. On the averaged stochastic approximation for linear regression. *SIAM Journal on Control and Optimization*, 34(1):31–61, 1996.
- Martin B Haugh and Leonid Kogan. Pricing american options: A duality approach. *Operations Research*, 52(2):258–270, 2004.
- Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems 7*, pages 703–710, 1994.
- Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Foundations and Trends® in Machine Learning*, 10(3-4):142–336, 2017.
- Vijay R Konda. *Actor-Critic Algorithms*. PhD thesis, Massachusetts Institute of Technology, 2002.
- Nathaniel Korda and Prashanth La. On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 626–634, 2015.
- Harold Kushner. Stochastic approximation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):87–96, 2010.

- Harold Kushner and Gang G Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.
- Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an  $o(1/t)$  convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.
- Chandrashekhar Lakshminarayanan and Csaba Szepesvári. Finite Time Bounds for Temporal Difference Learning with Function Approximation: Problems with some “state-of-the-art” results, 2017. URL <https://sites.ualberta.ca/~szepesva/papers/TD-issues17.pdf>.
- Chandrashekhar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.
- Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of LSTD. In *Proceedings of the 27th International Conference on Machine Learning*, pages 615–622, 2010.
- David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.
- Bo Liu, Mohammad Liu, Ji ]and Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 504–513, 2015.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
- 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.
- Bernardo Á Pires and Csaba Szepesvári. Statistical linear estimation with penalized estimators: An application to reinforcement learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pages 1755–1762, 2012.
- La Prashanth, Nathaniel Korda, and Rémi Munos. Fast LSTD using stochastic approximation: Finite time analysis and application to traffic control. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 66–81, 2013.
- Robert E Schapire and Manfred K Warmuth. On the worst-case analysis of temporal-difference learning algorithms. *Machine Learning*, 22(1-3):95–121, 1996.
- Harm Seijen and Richard S Sutton. True online td (lambda). In *Proceedings of the 31st International Conference on Machine Learning*, pages 692–700, 2014.
- Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.
- Richard S Sutton, Hamid R Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, pages 993–1000, 2009a.
- Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent  $o(n)$  temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems 21*, pages 1609–1616, 2009b.
- Ahmed Touati, Pierre-Luc Bacon, Doina Precup, and Pascal Vincent. Convergent TREE BACKUP and RETRACE with function approximation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4962–4971, 2018.

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

John N Tsitsiklis and Benjamin Van Roy. Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840 – 1851, 1999.

Stephen Tu and Benjamin Recht. Least-squares temporal difference learning for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5012–5021, 2018.

Benjamin Van Roy. *Learning and value function approximation in complex decision processes*. PhD thesis, Massachusetts Institute of Technology, 1998.

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

## A Analysis of Projected TD(0) under the Markov chain sampling model

In this section, we complete the proof of Theorem 3. The first subsection restates the theorem, as well as the two key lemmas from Section 8 that underly the proof. The second subsection contains a proof of Theorem 3. Finally, Subsection A.3 contains the proof of a technical result, Lemma 10, which was omitted from the main text but we need for the proof.

### A.1 Restatement of the theorem and key lemmas from the main text

**Theorem 3.** *Suppose the Projected TD algorithm is applied with parameter  $R \geq \|\theta^*\|_2$  under the Markov chain observation model with Assumption 1. Set  $G = (r_{\max} + 2R)$ . Then the following claims hold.*

(a) *With a constant step-size sequence  $\alpha_0 = \dots = \alpha_T = 1/\sqrt{T}$ ,*

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2 \left( 9 + 12\tau^{\text{mix}}(1/\sqrt{T}) \right)}{2\sqrt{T}(1-\gamma)}.$$

(b) *With a constant step-size sequence  $\alpha_0 = \dots = \alpha_T < 1/(2\omega(1-\gamma))$ ,*

$$\mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] \leq \left( e^{-2\alpha_0(1-\gamma)\omega T} \right) \|\theta^* - \theta_0\|_2^2 + \alpha_0 \left( \frac{G^2 (9 + 12\tau^{\text{mix}}(\alpha_0))}{2(1-\gamma)\omega} \right).$$

(c) *With a decaying step-size sequence  $\alpha_t = 1/(\omega(t+1)(1-\gamma))$  for all  $t \in \mathbb{N}_0$ ,*

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{G^2 (9 + 24\tau^{\text{mix}}(\alpha_T))}{T(1-\gamma)^2\omega} (1 + \log T),$$

The key to our proof is the following lemmas, which were establish in Section 8. Recall the definition of the gradient error  $\zeta_t(\theta) \equiv (g_t(\theta) - \bar{g}(\theta))^{\top} (\theta - \theta^*)$ .

**Lemma 8.** *With probability 1, for every  $t \in \mathbb{N}_0$ ,*

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1-\gamma)\|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t\zeta_t(\theta_t) + \alpha_t^2 G^2.$$

**Lemma 11** (Bound on gradient bias). *Consider a non-increasing step-size sequence,  $\alpha_0 \geq \alpha_1 \dots \geq \alpha_T$ . Fix any  $t < T$ , and set  $t^* \equiv \max\{0, t - \tau^{\text{mix}}(\alpha_T)\}$ . Then,*

$$\mathbb{E} [\zeta_t(\theta_t)] \leq G^2 (4 + 6\tau^{\text{mix}}(\alpha_T)) \alpha_{t^*}.$$

*The following bound also holds:*

$$\mathbb{E} [\zeta_t(\theta_t)] \leq 6G^2 \sum_{i=0}^{t-1} \alpha_i.$$

### A.2 Proof of Theorem 3.

We now complete the proof of Theorem 3. The proof directly uses Lemma 8 and Lemma 11.

*Proof.* From Lemma 8, we have

$$\mathbb{E} \left[ \|\theta^* - \theta_{t+1}\|_2^2 \right] \leq \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] - 2\alpha_t(1-\gamma)\mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] + 2\alpha_t\mathbb{E} [\zeta_t(\theta_t)] + \alpha_t^2 G^2. \quad (35)$$

**Proof of part (a):** We first show the analysis for a constant step-size and iterate averaging. Considering  $\alpha_t = \alpha_0 = 1/\sqrt{T}$  in Equation (35), rearranging terms and summing from  $t = 0$  to  $t = T - 1$ , we get

$$2\alpha_0(1-\gamma) \sum_{t=0}^{T-1} \mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] \leq \sum_{t=0}^{T-1} \left( \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] - \mathbb{E} \left[ \|\theta^* - \theta_{t+1}\|_2^2 \right] \right) + G^2 + 2\alpha_0 \sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t)].$$

Using Lemma 11 (in which  $\alpha_{t^*} = \alpha_0$  in this case) and simplifying, we find

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] &\leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2}{2\alpha_0(1-\gamma)} + \frac{T \cdot 2G^2(2 + 3\tau^{\text{mix}}(1/\sqrt{T}))\alpha_0}{(1-\gamma)} \\ &= \frac{\sqrt{T} (\|\theta^* - \theta_0\|_2^2 + G^2)}{2(1-\gamma)} + \frac{\sqrt{T} \cdot 2G^2(2 + 3\tau^{\text{mix}}(1/\sqrt{T}))}{(1-\gamma)}. \end{aligned}$$

This gives us our desired result,

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + G^2 (9 + 12\tau^{\text{mix}}(1/\sqrt{T}))}{2\sqrt{T}(1-\gamma)}.$$

**Proof of part (b):** The proof is analogous to part (b) of Theorem 2. Consider a constant step-size of  $\alpha_0 < 1/(2\omega(1-\gamma))$ . Starting with Equation (35) and applying Lemma 1, which showed  $\|V_{\theta^*} - V_\theta\|_D^2 \geq \omega\|\theta^* - \theta\|_2^2$  for all  $\theta$ , we get

$$\begin{aligned} \mathbb{E} \left[ \|\theta^* - \theta_{t+1}\|_2^2 \right] &\leq (1 - 2\alpha_0(1-\gamma)\omega) \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] + \alpha_0^2 G^2 + 2\alpha_0 \mathbb{E} [\zeta_t(\theta_t)] \\ &\leq (1 - 2\alpha_0(1-\gamma)\omega) \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] + \alpha_0^2 G^2 (9 + 12\tau^{\text{mix}}(\alpha_0)), \end{aligned}$$

where we used Lemma 11 to go to the second inequality. Iterating over this inequality gives us our final result. For any  $T \in \mathbb{N}_0$ ,

$$\begin{aligned} \mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] &\leq (1 - 2\alpha_0(1-\gamma)\omega)^T \|\theta^* - \theta_0\|_2^2 + \alpha_0^2 G^2 (9 + 12\tau^{\text{mix}}(\alpha_0)) \sum_{t=0}^{\infty} (1 - 2\alpha_0(1-\gamma)\omega)^t \\ &\leq \left( e^{-2\alpha_0(1-\gamma)\omega T} \right) \|\theta^* - \theta_0\|_2^2 + \frac{\alpha_0 G^2 (9 + 12\tau^{\text{mix}}(\alpha_0))}{2(1-\gamma)\omega}. \end{aligned}$$

The second inequality above follows by solving the geometric series and then using the fact that  $(1 - 2\alpha_0(1-\gamma)\omega) \leq e^{-2\alpha_0(1-\gamma)\omega}$ .

**Proof of part (c):** We now show the analysis for a linearly decaying step-size using Equation (35) as our starting point. We again use Lemma 1, which showed  $\|V_{\theta^*} - V_\theta\|_D^2 \geq \omega\|\theta^* - \theta\|_2^2$  for all  $\theta$ , to get,

$$\begin{aligned} \mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] &\leq \frac{1}{(1-\gamma)\alpha_t} \left( (1 - (1-\gamma)\omega\alpha_t) \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] - \mathbb{E} \left[ \|\theta^* - \theta_{t+1}\|_2^2 \right] + \alpha_t^2 G^2 \right) + \\ &\quad \frac{2}{(1-\gamma)} \mathbb{E} [\zeta_t(\theta_t)]. \end{aligned}$$

Consider a decaying step-size  $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$ , simplify and sum from  $t = 0$  to  $T - 1$  to get

$$\sum_{t=0}^{T-1} \mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] \leq \underbrace{-\omega T \mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right]}_{<0} + \frac{G^2}{\omega(1-\gamma)^2} \sum_{t=0}^{T-1} \frac{1}{t+1} + \frac{2}{(1-\gamma)} \sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t)]. \tag{36}$$

To simplify notation, for the remainder of the proof put  $\tau = \tau^{\text{mix}}(\alpha_T)$ . We can decompose the sum of gradient errors as

$$\sum_{t=0}^{T-1} \mathbb{E}[\zeta_t(\theta_t)] = \sum_{t=0}^{\tau} \mathbb{E}[\zeta_t(\theta_t)] + \sum_{t=\tau+1}^{T-1} \mathbb{E}[\zeta_t(\theta_t)]. \quad (37)$$

We will upper bound each term. In each case we use that, since  $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$ ,

$$\sum_{t=0}^{T-1} \alpha_t = \frac{1}{\omega(1-\gamma)} \sum_{t=0}^{T-1} \frac{1}{(t+1)} \leq \frac{1 + \log T}{\omega(1-\gamma)}.$$

Combining this with Lemma 11 gives,

$$\sum_{t=0}^{\tau} \mathbb{E}[\zeta_t(\theta_t)] \leq \sum_{t=0}^{\tau} \left( 6G^2 \sum_{i=0}^{t-1} \alpha_i \right) \leq \tau \left( 6G^2 \sum_{i=0}^{T-1} \alpha_i \right) \leq \frac{6G^2 \tau}{\omega(1-\gamma)} (1 + \log T).$$

Similarly, using Lemma 11, we have

$$\sum_{t=\tau+1}^{T-1} \mathbb{E}[\zeta_t(\theta_t)] \leq 2G^2 (2 + 3\tau) \sum_{t=\tau+1}^{T-1} \alpha_{t-\tau} \leq 2G^2 (2 + 3\tau) \sum_{t=1}^{T-1} \alpha_t \leq \frac{2G^2 (2 + 3\tau)}{\omega(1-\gamma)} (1 + \log T).$$

Combining the two parts, we get

$$\sum_{t=0}^{T-1} \mathbb{E}[\zeta_t(\theta_t)] \leq \frac{4G^2 (1 + 3\tau)}{\omega(1-\gamma)} (1 + \log T).$$

Using this in conjunction with Equation (36) we give final result.

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left[ \|V_{\theta_t} - V_{\theta^*}\|_D^2 \right] \leq \frac{G^2}{\omega T (1-\gamma)^2} (1 + \log T) + \frac{2}{T(1-\gamma)} \sum_{t=0}^{T-1} \mathbb{E}[\zeta_t(\theta_t)].$$

Simplifying and substituting  $\tau = \tau^{\text{mix}}(\alpha_T)$ , we get

$$\begin{aligned} \mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] &\leq \frac{G^2}{\omega T (1-\gamma)^2} (1 + \log T) + \frac{8G^2 (1 + 3\tau^{\text{mix}}(\alpha_T))}{\omega T (1-\gamma)^2} (1 + \log T) \\ &\leq \frac{G^2 (9 + 24\tau^{\text{mix}}(\alpha_T))}{\omega T (1-\gamma)^2} (1 + \log T). \end{aligned}$$

Additionally, Equation (36) also gives us a convergence rate of  $\mathcal{O}(\log T/T)$  for the iterate  $\theta_T$  itself:

$$\mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] \leq \frac{G^2 (9 + 24\tau^{\text{mix}}(\alpha_T))}{\omega^2 T (1-\gamma)^2} (1 + \log T).$$

□

### A.3 Proof of Lemma 10

**Lemma 10** (Gradient error is bounded and Lipschitz). *With probability 1,*

$$|\zeta_t(\theta)| \leq 2G^2 \quad \text{for all } \theta \in \Theta_R$$

and

$$|\zeta_t(\theta) - \zeta_t(\theta')| \leq 6G \left\| (\theta - \theta') \right\|_2 \quad \text{for all } \theta, \theta' \in \Theta_R.$$

*Proof.* The first claim follows from a simple argument using Lemma 6.

$$|\zeta_t(\theta)| = \left| (g_t(\theta) - \bar{g}(\theta))^{\top} (\theta - \theta^*) \right| \leq (\|g_t(\theta)\|_2 + \|\bar{g}(\theta)\|_2) (\|\theta\|_2 + \|\theta^*\|_2) \leq 4GR \leq 2G^2,$$

where the first inequality follows from the triangle inequality and the Cauchy-Schwartz inequality, and the final inequality uses that  $R \leq G/2$  by definition of  $G = r_{\max} + 2R$ .

To establish the second claim, consider the following inequality for any vectors  $(a_1, b_1, a_2, b_2)$ :

$$|a_1^{\top} b_1 - a_2^{\top} b_2| = |a_1^{\top} (b_1 - b_2) + b_2^{\top} (a_1 - a_2)| \leq \|a_1\| \|b_1 - b_2\| + \|b_2\| \|a_1 - a_2\|.$$

This follows as a direct application of Cauchy-Schwartz. It implies that for any  $\theta, \theta' \in \Theta_R$ ,

$$\begin{aligned} |\zeta_t(\theta) - \zeta_t(\theta')| &= \left| (g_t(\theta) - \bar{g}(\theta))^{\top} (\theta - \theta^*) - (g_t(\theta') - \bar{g}(\theta'))^{\top} (\theta' - \theta^*) \right| \\ &\leq \|g_t(\theta) - \bar{g}(\theta)\|_2 \|\theta - \theta'\|_2 + \|\theta' - \theta^*\|_2 \| (g_t(\theta) - \bar{g}(\theta)) - (g_t(\theta') - \bar{g}(\theta')) \|_2 \\ &\leq 2G \|\theta - \theta'\|_2 + 2R (\|g_t(\theta) - g_t(\theta')\|_2 + \|\bar{g}(\theta) - \bar{g}(\theta')\|_2) \\ &\leq 2G \|\theta - \theta'\|_2 + 8R \|\theta - \theta'\|_2 \\ &\leq 6G \|\theta - \theta'\|_2. \end{aligned}$$

where we used that  $R \leq G/2$  by the definition of  $G$ . We also used that both  $g_t(\cdot)$  and  $\bar{g}(\cdot)$  are 2-Lipschitz functions which is easy to see. Starting with  $g_t(\theta) = (r_t + \gamma\phi(s'_t)^{\top}\theta - \phi(s_t)^{\top}\theta)\phi(s_t)$ , consider

$$\begin{aligned} \|g_t(\theta) - g_t(\theta')\|_2 &= \left\| \phi(s_t) (\gamma\phi(s'_t) - \phi(s_t))^{\top} (\theta - \theta') \right\|_2 \\ &\leq \|\phi(s_t)\|_2 \|(\gamma\phi(s'_t) - \phi(s_t))\|_2 \|(\theta - \theta')\|_2 \\ &\leq 2\|(\theta - \theta')\|_2. \end{aligned}$$

Similarly, following Equation (2), we have  $\|\bar{g}(\theta) - \bar{g}(\theta')\|_2 = \left\| \mathbb{E}[\phi(\gamma\phi' - \phi)]^{\top} (\theta - \theta') \right\|_2$ , where  $\phi = \phi(s)$  is the feature vector of a random initial state  $s \sim \pi$ ,  $\phi' = \phi(s')$  is the feature vector of a random next state drawn according to  $s' \sim \mathcal{P}(\cdot | s)$ . Therefore,

$$\|\bar{g}(\theta) - \bar{g}(\theta')\|_2 \leq \left\| \phi(\gamma\phi' - \phi)^{\top} (\theta - \theta') \right\| \leq 2\|(\theta - \theta')\|_2.$$

□

## B Analysis of Projected TD( $\lambda$ ) under Markov chain observation model

In this section, we give a detailed proof of the convergence bounds presented in Theorem 4. Subsection B.1 details our proof strategy along with key lemmas which come together in Subsection B.2 to establish the results. We begin by providing mathematical expressions for TD( $\lambda$ ) updates.

**Stationary distribution of TD( $\lambda$ ) updates:** Recall that the projected TD( $\lambda$ ) update at time  $t$  is given by:

$$\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))$$

where  $\Pi_{2,R}(\cdot)$  denotes the projection operator onto a norm ball of radius  $R < \infty$  and  $x_t(\theta_t, z_{0:t})$  is the update direction. Let us now give explicit mathematical expressions for  $x_t(\theta, z_{0:t})$  and its steady-state mean  $\bar{x}(\theta)$ . Note that these are analogous to the expressions for the negative gradient  $g_t(\theta)$  and its steady-state expectation  $\bar{g}(\theta)$  for TD(0). At time  $t$ , as a general function of (non-random)  $\theta$  and the tuple  $O_t = (s_t, r_t, s'_t)$  along with the eligibility trace term  $z_{0:t}$ , we have

$$x_t(\theta, z_{0:t}) = (r_t + \gamma\phi(s'_t)^\top \theta - \phi(s_t)^\top \theta) z_{0:t} = \delta_t(\theta) z_{0:t} \quad \forall \theta \in \mathbb{R}^d.$$

The asymptotic convergence of TD( $\lambda$ ) is closely related to the expected value of  $x_t(\theta, z_{0:t})$  under the steady-state behavior of  $(O_t, z_{0:t})$ ,

$$\bar{x}(\theta) = \lim_{t \rightarrow \infty} \mathbb{E}[\delta_t(\theta) z_{0:t}].$$

Rather than take this limit, it will be helpful in our analysis to think of an equivalent *backward view* by constructing a stationary process with mean  $\bar{x}(\theta)$ . Consider a stationary sequence of states  $(\dots, s_{-1}, s_0, s_1, \dots)$  and set  $z_{-\infty:t} = \sum_{k=0}^{\infty} (\gamma\lambda)^k \phi(s_{t-k})$ . Then the sequence  $(x_0(\theta, z_{-\infty:0}), x_1(\theta, z_{-\infty:1}), \dots)$  is stationary, and we have

$$\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta) z_{-\infty:t}]. \tag{38}$$

It should be emphasized that  $\bar{x}(\theta)$  and the states  $(\dots, s_{-2}, s_{-1})$  are introduced only for the purposes of our analysis and are never used by the algorithm itself. However, this turns out to be quite useful as it is easy to show (Van Roy, 1998) that

$$\bar{x}(\theta) = \Phi^\top D \left( T_\mu^{(\lambda)} \Phi \theta - \Phi \theta \right), \tag{39}$$

where  $\Phi$  is the feature matrix and  $(T_\mu^{(\lambda)} \Phi \theta - \Phi \theta)$  denotes the Bellman error defined with respect to the Bellman operator  $T_\mu^{(\lambda)}(\cdot)$ , corresponding to a policy  $\mu$ . Careful readers will notice the stark similarity between Equation (39) and Equation (3). Exploiting the property that  $\Pi_D T_\mu^\lambda(\cdot)$  is also a contraction operator, one can easily show a result equivalent to Lemma 3, thus quantifying the progress we make by taking steps in the direction of  $\bar{x}(\theta)$ . The rest of our proof essentially shows how to control for the observation noise, i.e. the fact that we use  $x_t(\theta, z_{0:t})$  rather than  $\bar{x}(\theta)$  to make updates. To remind the readers of the results, we first restate Theorem 4 below.

**Theorem 4.** Suppose the Projected TD( $\lambda$ ) algorithm is applied with parameter  $R \geq \|\theta^*\|_2$  under the Markov chain observation model with Assumption 1. Set  $B = \frac{(r_{\max} + 2R)}{(1-\gamma\lambda)}$ . Then the following claims hold.

(a) With a constant step-size  $\alpha_t = \alpha_0 = 1/\sqrt{T}$ ,

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2 \right] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2 \left( 13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T}) \right)}{2\sqrt{T}(1-\kappa)}.$$

(b) With a constant step-size  $\alpha_t = \alpha_0 < 1/(2\omega(1-\kappa))$  and  $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$ ,

$$\mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] \leq \left( e^{-2\alpha_0(1-\kappa)\omega T} \right) \|\theta^* - \theta_0\|_2^2 + \alpha_0 \left( \frac{B^2 (13 + 24\tau_\lambda^{\text{mix}}(\alpha_0))}{2(1-\kappa)\omega} \right).$$

(c) With a decaying step-size  $\alpha_t = 1/(\omega(t+1)(1-\kappa))$ ,

$$\mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_T}\|_D^2 \right] \leq \frac{B^2 (13 + 52\tau_\lambda^{\text{mix}}(\alpha_T))}{T(1-\kappa)^2 \omega} (1 + \log T).$$

## B.1 Proof strategy and key lemmas

We now describe our proof strategy and give key lemmas used to establish Theorem 4. Throughout, we consider the Markov chain observation model with Assumption 1 and study the Projected TD ( $\lambda$ ) algorithm applied with parameter  $R \geq \|\theta^*\|_2$  and step-size sequence  $(\alpha_0, \dots, \alpha_T)$ . To simplify our exposition, we introduce some notation below.

**Notation:** We specify the notation used throughout this section. Define the set  $\Theta_R = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq R\}$ , so  $\theta_t \in \Theta_R$  for each  $t$  because of the algorithm's projection step. Next, we generically define  $z_{l:t} = \sum_{k=0}^{t-l} (\gamma\lambda)^k \phi(s_{t-k})$  for any lower limit  $l \leq t$ . Thus,  $z_{l:t}$  denotes the eligibility trace as a function of the states  $(s_l, \dots, s_t)$ . Next, we define  $\zeta_t(\theta, z_{l:t})$  as a general function of  $\theta$  and  $z_{l:t}$ ,

$$\zeta_t(\theta, z_{l:t}) = (\delta_t(\theta)z_{l:t} - \bar{x}(\theta))^\top (\theta - \theta^*). \quad (40)$$

Here, the subscript  $t$  in  $\zeta_t$  encodes the dependence on the tuple  $O_t = (s_t, r_t, s'_t)$  which is used to compute the Bellman error,  $\delta_t(\cdot)$  at time  $t$ . Finally, we set  $B := (r_{\max} + 2R)/(1 - \gamma\lambda)$  which implies  $B > R/2$ , a fact we use many times in our proofs to simplify constant terms. As a reminder, note that our bounds depend on the mixing time, which we defined in Section 9 as

$$\tau_\lambda^{\text{mix}}(\epsilon) = \max\{\tau^{\text{MC}}(\epsilon), \tau^{\text{Algo}}(\epsilon)\},$$

where  $\tau^{\text{MC}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \leq \epsilon\}$  and  $\tau^{\text{Algo}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid (\gamma\lambda)^t \leq \epsilon\}$ .

**Proof outline:** The analysis for TD( $\lambda$ ) can be broadly divided into three parts and closely mimics the steps used to prove TD(0) results.

- As a first step, we do an error decomposition, similar to the result shown in Lemma 8. This is enabled by two key lemmas, which are analogues of Lemma 3 and Lemma 6 for Projected TD(0). The first one spells out a clear relationship of how the updates following  $\bar{x}(\theta)$  point in the descent direction of  $\|\theta^* - \theta\|_2^2$  while the second one upper bounds the norm of the update direction,  $x_t(\theta, z_{0:t})$ , by the constant  $B$  (as defined above).
- The error decomposition that we obtain from Step 1 can be stated as:

$$\mathbb{E}[\|\theta^* - \theta_{t+1}\|_2^2] \leq \mathbb{E}[\|\theta^* - \theta_t\|_2^2] - 2\alpha_t(1-\kappa)\mathbb{E}[\|V_{\theta^*} - V_{\theta_t}\|_D^2] + 2\alpha_t\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] + \alpha_t^2 B^2.$$

In the second step, we establish an upper bound on the bias term,  $\mathbb{E}[\zeta_t(\theta_t, z_{0:t})]$ , which is the main challenge in our proof. Recall that the dependent nature of the state transitions may result in strong coupling between the tuples  $O_{t-1}$  and  $O_t$  under the Markov chain observation model. Therefore, this bias in update direction can potentially be non-zero. Presence of the eligibility trace term,  $z_{0:t}$ , which is a function of the entire history of states,  $(s_0, \dots, s_t)$ , further complicates the analysis by introducing subtle dependencies.

To control for this, we use information-theoretic techniques shown in Lemma 9 which exploit the geometric ergodicity of the MDP, along with the geometric weighting of state features in the eligibility trace term. Our result essentially shows that the bias scales the noise in update direction by a factor of the mixing time. Mathematically, for a constant step-size  $\alpha$ , we show that  $\mathbb{E}[\alpha\zeta_t(\theta_t, z_{0:t})] \approx B^2(6 + 12\tau_\lambda^{\text{mix}}(\alpha))\alpha^2$ . We show a similar result for decaying step-sizes as well.

3. In the final step, we combine the error decomposition from Step 1 and the bound on the bias from Step 2, to establish finite time bounds on the performance of Projected TD( $\lambda$ ) for different step-size choices. We closely mimic the analysis of Nemirovski et al. (2009) for a constant, aggressive step-size of  $(1/\sqrt{T})$  and the proof ideas of Lacoste-Julien et al. (2012) for decaying step-sizes.

### B.1.1 Error decomposition under Projected TD( $\lambda$ )

We first prove Lemmas 16 and 17 which enable the error decomposition shown in Lemma 18.

**Lemma 16.** [Tsitsiklis and Van Roy (1997)] Let  $V_{\theta^*}$  be the unique fixed point of  $\Pi_D T_\mu^{(\lambda)}(\cdot)$  i.e.  $V_{\theta^*} = \Pi T_\mu^{(\lambda)} V_{\theta^*}$ . For any  $\theta \in \mathbb{R}^d$ ,

$$(\theta^* - \theta)^\top \bar{x}(\theta) \geq (1 - \kappa) \|V_{\theta^*} - V_\theta\|_D^2.$$

*Proof.* We use the definition of  $\bar{x}(\theta) = \langle \Phi^\top, T_\mu^{(\lambda)} \Phi \theta - \Phi \theta \rangle_D$  as shown in Equation (39) along with the fact that  $\Pi_D T_\mu^{(\lambda)}(\cdot)$  is a contraction with respect to  $\|\cdot\|_D$  with modulus  $\kappa$ . See Appendix C for a complete proof.  $\square$

**Lemma 17.** For all  $\theta \in \Theta_R$ ,  $\|x_t(\theta, z_{0:t})\|_2 \leq B$  with probability 1. Additionally,  $\|\bar{x}(\theta)\|_2 \leq B$ .

*Proof.* See Subsection B.3 for a complete proof.  $\square$

The above two lemmas can be easily combined to establish a recursion for the error under projected TD( $\lambda$ ) that holds for each sample path.

**Lemma 18.** With probability 1, for every  $t \in \mathbb{N}_0$ ,

$$\|\theta^* - \theta_{t+1}\|_2^2 \leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1 - \kappa) \|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t \zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 B^2.$$

*Proof.* The Projected TD( $\lambda$ ) algorithm updates the parameter as:  $\theta_{t+1} = \Pi_{2,R}[\theta_t + \alpha_t x_t(\theta_t, z_{0:t})]$   $\forall t \in \mathbb{N}_0$ . This implies,

$$\begin{aligned} \|\theta^* - \theta_{t+1}\|_2^2 &= \|\theta^* - \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))\|_2^2 \\ &= \|\Pi_{2,R}(\theta^*) - \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))\|_2^2 \\ &\leq \|\theta^* - \theta_t - \alpha_t x_t(\theta_t, z_{0:t})\|_2^2 \\ &= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t x_t(\theta_t, z_{0:t})^\top (\theta^* - \theta_t) + \alpha_t^2 \|x_t(\theta_t, z_{0:t})\|_2^2 \\ &\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t x_t(\theta_t, z_{0:t})^\top (\theta^* - \theta_t) + \alpha_t^2 B^2 \\ &= \|\theta^* - \theta_t\|_2^2 - 2\alpha_t \bar{x}(\theta_t)^\top (\theta^* - \theta_t) + 2\alpha_t \zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 G^2. \\ &\leq \|\theta^* - \theta_t\|_2^2 - 2\alpha_t(1 - \kappa) \|V_{\theta^*} - V_{\theta_t}\|_D^2 + 2\alpha_t \zeta_t(\theta_t, z_{0:t}) + \alpha_t^2 B^2. \end{aligned}$$

The first inequality used that orthogonal projection operators onto a convex set are non-expansive, the second used Lemma 17 together with the fact  $\|\theta_t\|_2 \leq R$  due to projection, and the third used Lemma 16. Note that we used  $\zeta_t(\theta_t, z_{0:t})$  to simplify the notation for the error in the update direction. Recall the definition of the error function from Equation (40) which implies,

$$\zeta_t(\theta_t, z_{0:t}) = (\delta_t(\theta_t) z_{0:t} - \bar{x}(\theta_t))^\top (\theta_t - \theta^*) = (x_t(\theta_t, z_{0:t}) - \bar{x}(\theta_t))^\top (\theta_t - \theta^*).$$

$\square$

### B.1.2 Upper bound on the bias in update direction.

We give an upper bound on the expected error in the update direction,  $\mathbb{E}[\zeta_t(\theta_t, z_{0:t})]$ , which as explained above, is the key challenge for our analysis. For this, we first establish some basic regularity properties of the error function  $\zeta_t(\cdot, \cdot)$  in Lemma 19 below. In particular, part (a) shows boundedness, part (b) shows that it is Lipschitz in the first argument and part (c) bounds the error due to truncation of the eligibility trace. Recall that  $z_{l:t}$  denotes the eligibility trace as a function of the states  $(s_l, \dots, s_t)$ .

**Lemma 19.** *Consider any  $l \leq t$  and any  $\theta, \theta' \in \Theta_R$ . With probability 1,*

- (a)  $|\zeta_t(\theta, z_{l:t})| \leq 2B^2$ .
- (b)  $|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| \leq 6B\|(\theta - \theta')\|_2$ .
- (c) *The following two bounds also hold,*

$$\begin{aligned} |\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| &\leq B^2(\gamma\lambda)^\tau \text{ for all } \tau \leq t, \\ |\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| &\leq B^2(\gamma\lambda)^t. \end{aligned}$$

*Proof.* We essentially use the uniform bound on  $x_t(\theta, z_{0:t})$  and  $\bar{x}(\theta)$  as stated in Lemma 17 to show this result. See Subsection B.3 for a detailed proof.  $\square$

Lemma 19 can be combined with Lemma 9 to give an upper bound on the bias term,  $\mathbb{E}[\zeta_t(\theta_t, z_{0:t})]$ , as shown below.

**Lemma 20.** *Consider a non-increasing step-size sequence,  $\alpha_0 \geq \alpha_1 \dots \geq \alpha_T$ . Then the following hold.*

- (a) *For  $2\tau_\lambda^{\text{mix}}(\alpha_T) < t \leq T$ ,*

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T))\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)}.$$

- (b) *For  $0 \leq t \leq 2\tau_\lambda^{\text{mix}}(\alpha_T)$ ,*

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2(1 + 2\tau_\lambda^{\text{mix}}(\alpha_T))\alpha_0 + B^2(\gamma\lambda)^t.$$

- (c) *For all  $t \in \mathbb{N}_0$ ,*

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t$$

*Proof.* We proceed in two cases below. Throughout the proof, results from Lemma 19 are applied using the fact that  $\theta_t \in \Theta_R$ , because of the algorithm's projection step.

**Case (a):** Let  $t > 2\tau$  and consider the following decomposition for all  $\tau \in \{0, 1, \dots, t/2\}$ . We show an upper bound on each of the three terms separately.

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq |\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| + |\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]|.$$

Step 1: Use regularity properties of the error function to bound first two terms.

We relate  $\zeta_t(\theta_t, z_{0:t})$  and  $\zeta_t(\theta_{t-\tau}, z_{0:t})$  using the Lipschitz property shown in part (b) of Lemma 19 to get,

$$|\zeta_t(\theta_t, z_{0:t}) - \zeta_t(\theta_{t-2\tau}, z_{0:t})| = 6B\|\theta_t - \theta_{t-2\tau}\|_2 \leq 6B^2 \sum_{i=t-2\tau}^{t-1} \alpha_i. \quad (41)$$

Taking expectations on both sides gives us the desired bound on the first term. The last inequality used the norm bound on update direction as shown in Lemma 17 to simplify,

$$\|\theta_t - \theta_{t-2\tau}\|_2 \leq \sum_{i=t-2\tau}^{t-1} \|\Pi_{2,R}(\theta_{i+1} + \alpha_i x_i(\theta_i, z_{0:i})) - \theta_i\|_2 \leq \sum_{i=t-2\tau}^{t-1} \alpha_i \|x_i(\theta_i, z_{0:i})\|_2 \leq B \sum_{i=t-2\tau}^{t-1} \alpha_i.$$

Similarly, by part (c) of Lemma 19, we have a bound on the second term.

$$|\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| \leq B^2(\gamma\lambda)^\tau. \quad (42)$$

Step 2: Use information-theoretic arguments to upper bound  $\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]$ .

We will essentially use Lemma 9 to upper bound  $\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]$ . We first introduce some notation to highlight subtle dependency issues. Note that  $\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})$  is a function of  $(\theta_{t-2\tau}, s_{t-\tau}, \dots, s_{t-1}, O_t)$ . To simplify, let  $Y_{t-\tau:t} = (s_{t-\tau}, \dots, s_{t-1}, O_t)$ . Define,

$$f(\theta_{t-2\tau}, Y_{t-\tau:t}) := \zeta_t(\theta_{t-2\tau}, z_{t-\tau:t}).$$

Consider random variables  $\theta'_{t-2\tau}$  and  $Y'_{t-\tau:t}$  drawn independently from the marginal distributions of  $\theta_{t-2\tau}$  and  $Y_{t-\tau:t}$ , so  $\mathbb{P}(\theta'_{t-2\tau} = \cdot, Y'_{t-\tau:t} = \cdot) = \mathbb{P}(\theta_{t-2\tau} = \cdot) \otimes \mathbb{P}(Y_{t-\tau:t} = \cdot)$ . By Lemma 19 we have that  $|f(\theta, Y_{t-\tau:t})| \leq 2B^2$  for all  $\theta \in \Theta_R$  with probability 1. As

$$\theta_{t-2\tau} \rightarrow s_{t-2\tau} \rightarrow s_{t-\tau} \rightarrow s_t \rightarrow O_t$$

form a Markov chain, a direct application of Lemma 9 gives us:

$$|\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})] - \mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]| \leq 4B^2m\rho^\tau. \quad (43)$$

We also have the following bound for all fixed  $\theta \in \Theta_R$ . Using  $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta)z_{-\infty:t}]$ , we get

$$\mathbb{E}[f(\theta, Y_{t-\tau:t})] = (\mathbb{E}[\delta_t(\theta)z_{t-\tau:t}] - \bar{x}(\theta))^\top (\theta - \theta^*) \leq |(\delta_t(\theta)z_{-\infty:t})^\top (\theta - \theta^*)| \leq B^2(\gamma\lambda)^\tau$$

Combining the above with Equation (43), we get

$$\begin{aligned} |\mathbb{E}[\zeta_t(\theta_{t-2\tau}, z_{t-\tau:t})]| &= |\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})]| \\ &\leq |\mathbb{E}[f(\theta_{t-2\tau}, Y_{t-\tau:t})] - \mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]| + |\mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})]| \\ &\leq 4B^2m\rho^\tau + |\mathbb{E}[\mathbb{E}[f(\theta'_{t-2\tau}, Y'_{t-\tau:t})|\theta'_{t-2\tau}]]| \\ &\leq 4B^2m\rho^\tau + B^2(\gamma\lambda)^\tau. \end{aligned} \quad (44)$$

Step 3. Combine terms to show part (a) of our claim.

Taking  $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$  and combining Equations (41), (42) and (44) establishes the first claim.

$$\begin{aligned} \mathbb{E}[\zeta_t(\theta_t, z_{0:t})] &\leq 6B^2 \sum_{i=t-2\tau}^{t-1} \alpha_i + 4B^2m\rho^\tau + 2B^2(\gamma\lambda)^\tau \leq 12B^2\tau_\lambda^{\text{mix}}(\alpha_T)\alpha_{t-2\tau}^{\text{mix}}(\alpha_T) + 6B^2\alpha_T \\ &\leq 6B^2(1 + 2\tau^{\text{mix}}(\alpha_T))\alpha_{t-2\tau}^{\text{mix}}(\alpha_T). \end{aligned}$$

Here we used that letting  $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$  implies:  $\max\{m\rho^\tau, (\gamma\lambda)^\tau\} \leq \alpha_T$ . Two additional facts which we also use follow from a non-increasing step-size sequence,  $\sum_{i=t-2\tau}^{t-1} \alpha_i \leq 2\tau\alpha_{t-2\tau}$  and  $\alpha_T \leq \alpha_{t-2\tau}$ .

**Case (b):** Consider the following decomposition for all  $t \in \mathbb{N}_0$ ,

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq |\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]|.$$

Step 1: Use regularity properties of the error function to upper bound the first two terms.

Using parts (b), (c) of Lemma 19 and following the arguments shown in Step 1, 2 of case (a) above, we get

$$|\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{0:t})]| + |\mathbb{E}[\zeta_t(\theta_0, z_{0:t})] - \mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})]| \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t. \quad (45)$$

Step 2: Characterizing  $\mathbb{E}[\zeta_t(\theta, z_{-\infty:t})]$  for any fixed (non-random)  $\theta$ .

Recall the definition of  $\bar{x}(\theta)$  from Equation (38). For any fixed (non-random)  $\theta$ , we have  $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta)z_{-\infty:t}]$ . Therefore,

$$\mathbb{E}[\zeta_t(\theta_0, z_{-\infty:t})] = (\mathbb{E}[\delta_t(\theta_0)z_{-\infty:t}] - \bar{x}(\theta_0))^\top (\theta_0 - \theta^*) = 0. \quad (46)$$

Step 3. Combine terms to show parts (b), (c) of our claim.

Combining Equations (45) and (46) establishes part (c) which states,

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2 \sum_{i=0}^{t-1} \alpha_i + B^2(\gamma\lambda)^t \quad \forall t \in \mathbb{N}_0.$$

We establish part (b) by using that the step-size sequence is non-increasing which implies:  $\sum_{i=0}^{t-1} \alpha_i \leq t\alpha_0$ . For all  $t \leq 2\tau_\lambda^{\text{mix}}(\alpha_T)$ , we have the following loose upper bound.

$$\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] \leq 6B^2 t\alpha_0 + B^2(\gamma\lambda)^t \leq 6B^2 (1 + 2\tau_\lambda^{\text{mix}}(\alpha_T)) \alpha_0 + B^2(\gamma\lambda)^t.$$

□

## B.2 Proof of Theorem 4

In this subsection, we establish convergence bounds for Projected TD( $\lambda$ ) as stated in Theorem 4 using Lemmas 18 and 20. From Lemma 18 we have,

$$\mathbb{E}[\|\theta^* - \theta_{t+1}\|_2^2] \leq \mathbb{E}[\|\theta^* - \theta_t\|_2^2] - 2\alpha_t(1 - \kappa)\mathbb{E}[\|V_{\theta^*} - V_{\theta_t}\|_D^2] + 2\alpha_t\mathbb{E}[\zeta_t(\theta_t, z_{0:t})] + \alpha_t^2 B^2. \quad (47)$$

Equation (47) will be used as a starting point for analyzing different step-size choices.

**Proof of part (a):** Fix a constant step-size of  $\alpha_0 = \dots = \alpha_t = 1/\sqrt{T}$  in Equation (47), rearrange terms and sum from  $t = 0$  to  $t = T - 1$ , we get

$$2\alpha_0(1 - \kappa) \sum_{t=0}^{T-1} \mathbb{E}[\|V_{\theta^*} - V_{\theta_t}\|_D^2] \leq \sum_{t=0}^{T-1} \left( \mathbb{E}[\|\theta^* - \theta_t\|_2^2] - \mathbb{E}[\|\theta^* - \theta_{t+1}\|_2^2] \right) + B^2 + 2\alpha_0 \sum_{t=0}^{T-1} \mathbb{E}[\zeta_t(\theta_t, z_{0:t})].$$

Using Lemma 20 where  $\alpha_{t-2\tau_\lambda^{\text{mix}}(\alpha_T)} = \alpha_0$  along with the fact that  $(\gamma\lambda) < 1$ , we simplify to get

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[\|V_{\theta^*} - V_{\theta_t}\|_D^2] &\leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2}{2\alpha_0(1 - \kappa)} + \frac{T \cdot 6B^2(1 + 2\tau_\lambda^{\text{mix}}(1/\sqrt{T}))\alpha_0}{(1 - \kappa)} + \frac{1}{(1 - \kappa)} \sum_{t=0}^{2\tau_\lambda^{\text{mix}}(1/\sqrt{T})} B^2(\gamma\lambda)^t \\ &\leq \frac{\sqrt{T}(\|\theta^* - \theta_0\|_2^2 + B^2)}{2(1 - \kappa)} + \frac{\sqrt{T} \cdot 6B^2(1 + 2\tau_\lambda^{\text{mix}}(1/\sqrt{T}))}{(1 - \kappa)} + \frac{2B^2\tau_\lambda^{\text{mix}}(1/\sqrt{T})}{(1 - \kappa)}. \end{aligned}$$

Adding these terms, we conclude

$$\mathbb{E}[\|V_{\theta^*} - V_{\theta_T}\|_D^2] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[\|V_{\theta^*} - V_{\theta_t}\|_D^2] \leq \frac{\|\theta^* - \theta_0\|_2^2 + B^2(13 + 28\tau_\lambda^{\text{mix}}(1/\sqrt{T}))}{2\sqrt{T}(1 - \kappa)}.$$

**Proof of part (b):** For a constant step-size of  $\alpha_0 < 1/(2\omega(1-\kappa))$ , we show that the expected distance between the iterate  $\theta_T$  and the TD( $\lambda$ ) limit point,  $\theta^*$  converges at an exponential rate below some level that depends on the choice of step-size and  $\lambda$ . Starting with Equation (47) and applying Lemma 1 which shows that  $\|V_{\theta^*} - V_\theta\|_D^2 \geq w\|\theta^* - \theta\|_2^2$  for any  $\theta$ , we have that for all  $t > 2\tau_\lambda^{\text{mix}}(\alpha_0)$ ,

$$\begin{aligned}\mathbb{E} \left[ \|\theta^* - \theta_{t+1}\|_2^2 \right] &\leq (1 - 2\alpha_0(1-\kappa)\omega) \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] + \alpha_0^2 B^2 + 2\alpha_0 \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] \\ &\leq (1 - 2\alpha_0(1-\kappa)\omega) \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] + \alpha_0^2 B^2 (13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)),\end{aligned}$$

where we used part (a) of Lemma 20 for the second inequality. Iterating over it gives us our final result. For any  $T > 2\tau_\lambda^{\text{mix}}(\alpha_0)$ ,

$$\begin{aligned}\mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] &\leq (1 - 2\alpha_0(1-\kappa)\omega)^T \|\theta^* - \theta_0\|_2^2 + B^2 \left( \alpha_0^2 (13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)) \right) \sum_{t=0}^{\infty} (1 - 2\alpha_0(1-\kappa)\omega)^t \\ &\leq \left( e^{-2\alpha_0(1-\kappa)\omega T} \right) \|\theta^* - \theta_0\|_2^2 + \frac{B^2 \left( \alpha_0 (13 + 24\tau_\lambda^{\text{mix}}(\alpha_0)) \right)}{2(1-\kappa)\omega}.\end{aligned}$$

The second inequality follows by solving the geometric series and using that  $(1 - 2\alpha_0(1-\kappa)\omega) \leq e^{-2\alpha_0(1-\kappa)\omega}$ .

**Proof of part (c):** Consider a decaying step-size of  $\alpha_t = 1/(\omega(t+1)(1-\kappa))$ . We start with Equation (47) and use Lemma 1 which showed  $\mathbb{E} \left[ \|V_{\theta^*} - V_\theta\|_D^2 \right] \geq w\mathbb{E} \left[ \|\theta^* - \theta\|_2^2 \right]$  for all  $\theta$  to get,

$$\begin{aligned}\mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] &\leq \frac{1}{(1-\kappa)\alpha_t} \left( (1 - (1-\kappa)\omega\alpha_t) \mathbb{E} \left[ \|\theta^* - \theta_t\|_2^2 \right] - \mathbb{E} \left[ \|\theta^* - \theta_{t+1}\|_2^2 \right] + \alpha_t^2 B^2 \right) + \\ &\quad \frac{2}{(1-\kappa)} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})].\end{aligned}$$

Substituting  $\alpha_t = \frac{1}{\omega(t+1)(1-\kappa)}$ , simplify and sum from  $t = 0$  to  $T - 1$  to get,

$$\begin{aligned}\sum_{t=0}^{T-1} \mathbb{E} \left[ \|V_{\theta^*} - V_{\theta_t}\|_D^2 \right] &\leq -\omega T \mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right] + \frac{B^2}{\omega(1-\kappa)^2} \sum_{t=0}^{T-1} \frac{1}{t+1} + \frac{2}{(1-\kappa)} \sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] \\ &\leq \underbrace{-\omega T \mathbb{E} \left[ \|\theta^* - \theta_T\|_2^2 \right]}_{<0} + \frac{B^2(1 + \log T)}{\omega(1-\kappa)^2} + \frac{2}{(1-\kappa)} \sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})], \quad (48)\end{aligned}$$

where we used that  $\sum_{t=0}^{T-1} \frac{1}{t+1} \leq (1 + \log T)$ . To simplify notation, we put  $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$  for the remainder of the proof. We use Lemma 20 to upper bound the total bias,  $\sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})]$  which can be decomposed as:

$$\sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] = \sum_{t=0}^{2\tau} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] + \sum_{t=2\tau+1}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})]. \quad (49)$$

First, note that for a decaying step-size  $\alpha_t = \frac{1}{\omega(t+1)(1-\gamma)}$  we have

$$\sum_{t=0}^{T-1} \alpha_t = \frac{1}{\omega(1-\gamma)} \sum_{t=0}^{T-1} \frac{1}{(t+1)} \leq \frac{1 + \log T}{\omega(1-\gamma)}.$$

We will combine this with Lemma 20 to upper bound each term separately. First,

$$\begin{aligned} \sum_{t=0}^{2\tau} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] &\leq \sum_{t=0}^{2\tau} \left( 6B^2 \sum_{i=0}^{t-1} \alpha_i \right) + \sum_{t=0}^{2\tau} B^2 (\gamma\lambda)^t \\ &\leq \frac{6B^2}{\omega(1-\kappa)} \sum_{t=0}^{2\tau} \sum_{i=0}^{T-1} \frac{1}{(i+1)} + 2B^2\tau \leq \frac{14B^2\tau}{\omega(1-\kappa)} (1 + \log T), \end{aligned}$$

where we used the fact that  $\omega, \kappa, (\gamma\lambda) < 1$ . Similarly,

$$\sum_{t=2\tau+1}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] \leq 6B^2(1+2\tau) \sum_{t=2\tau+1}^{T-1} \alpha_{t-2\tau} \leq 6B^2(1+2\tau) \sum_{t=0}^{T-1} \alpha_t \leq \frac{6B^2(1+2\tau)}{\omega(1-\kappa)} (1 + \log T).$$

Combining the two parts, we get

$$\sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})] \leq \frac{B^2(6+26\tau)}{\omega(1-\kappa)} (1 + \log T).$$

Using this in conjunction with Equation (48) we get our final result,

$$\mathbb{E} [\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2] \leq \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|V_{\theta_t} - V_{\theta^*}\|_D^2] \leq \frac{B^2}{\omega T(1-\kappa)^2} (1 + \log T) + \frac{2}{T(1-\kappa)} \sum_{t=0}^{T-1} \mathbb{E} [\zeta_t(\theta_t, z_{0:t})].$$

Simplifying and putting back  $\tau = \tau_\lambda^{\text{mix}}(\alpha_T)$ , we get

$$\begin{aligned} \mathbb{E} [\|V_{\theta^*} - V_{\bar{\theta}_T}\|_D^2] &\leq \frac{B^2}{\omega T(1-\kappa)^2} (1 + \log T) + \frac{2B^2(6+26\tau_\lambda^{\text{mix}}(\alpha_T))}{\omega T(1-\kappa)^2} (1 + \log T) \\ &\leq \frac{B^2(13+52\tau_\lambda^{\text{mix}}(\alpha_T))}{\omega T(1-\kappa)^2} (1 + \log T). \end{aligned}$$

Additionally, Equation (48) implies a convergence rate of  $\mathcal{O}(\log T/T)$  for the iterate  $\theta_T$  itself:

$$\mathbb{E} [\|\theta^* - \theta_T\|_2^2] \leq \frac{B^2(13+52\tau_\lambda^{\text{mix}}(\alpha_T))}{\omega^2 T(1-\kappa)^2} (1 + \log T).$$

### B.3 Proof of supporting lemmas.

In this subsection, we provide standalone proofs of Lemma 17 and 19 used above.

**Lemma 17.** *For all  $\theta \in \Theta_R$ ,  $\|x_t(\theta, z_{0:t})\|_2 \leq B$  with probability 1. Additionally,  $\|\bar{x}(\theta)\|_2 \leq B$ .*

*Proof.* We start with the mathematical expression for  $x_t(\theta, z_{0:t})$ .

$$x_t(\theta, z_{0:t}) = \delta_t(\theta) z_{0:t} \Rightarrow \|x_t(\theta, z_{0:t})\|_2 = |\delta_t(\theta)| \|z_{0:t}\|_2.$$

We give an upper bound on both  $|\delta_t(\theta)|$  and  $\|z_{0:t}\|_2$ . Starting with the definition of  $\delta_t(\theta)$  and using that  $\|\phi(s_t)\|_2 \leq 1 \forall t$  along with  $\|\theta\|_2 \leq R$ , we get

$$|\delta_t(\theta)| = |r_t + \gamma\phi(s'_t)^\top \theta - \phi(s_t)^\top \theta| \leq r_{\max} + \|\phi(s'_t)\|_2 \|\theta\|_2 + \|\phi(s_t)\|_2 \|\theta\|_2 \leq (r_{\max} + 2R).$$

Next,

$$\|z_{0:t}\|_2^2 = \left\| \sum_{k=0}^t (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2^2 \leq \left( \sum_{k=0}^t (\gamma\lambda)^k \right)^2 \leq \left( \sum_{k=0}^{\infty} (\gamma\lambda)^k \right)^2 = \frac{1}{(1-\gamma\lambda)^2}.$$

Combining these two implies the first part of our claim.

$$\|x_t(\theta, z_{0:t})\|_2 = |\delta_t(\theta)| \|z_{0:t}\|_2 \leq \frac{(r_{\max} + 2R)}{(1 - \gamma\lambda)} = B.$$

Note that we can easily show an upper bound  $\|\delta_t(\theta)z_{l:t}\|_2 \leq B$  for any pair  $(\theta, z_{l:t})$  with  $l \leq t$ . Consider,

$$\begin{aligned} \|z_{l:t}\|_2^2 &\leq \|z_{-\infty:t}\|_2^2 \leq \left( \sum_{k=0}^{\infty} (\gamma\lambda)^k \right)^2 = \frac{1}{(1 - \gamma\lambda)^2} \\ \Rightarrow \|\delta_t(\theta)z_{l:t}\|_2 &= |\delta_t(\theta)| \|z_{l:t}\|_2 \leq \frac{(r_{\max} + 2R)}{(1 - \gamma\lambda)} = B. \end{aligned}$$

Taking  $l \rightarrow -\infty$  implies that  $\|\delta_t(\theta)z_{-\infty:t}\|_2 \leq B$ . As  $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta)z_{-\infty:t}]$ , we also have a uniform norm bound on the expected updates,  $\|\bar{x}(\theta)\|_2 \leq B$ , as claimed.  $\square$

**Lemma 19.** Consider any  $l \leq t$  and any  $\theta, \theta' \in \Theta_R$ . With probability 1,

- (a)  $|\zeta_t(\theta, z_{l:t})| \leq 2B^2$ .
- (b)  $|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| \leq 6B\|(\theta - \theta')\|_2$ .
- (c) The following two bounds also hold,

$$\begin{aligned} |\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| &\leq B^2(\gamma\lambda)^\tau \text{ for all } \tau \leq t, \\ |\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| &\leq B^2(\gamma\lambda)^t. \end{aligned}$$

*Proof.* Throughout, we use the assumption that basis vectors are normalized i.e.  $\|\phi(s_t)\|_2 \leq 1 \forall t$ .

**Part (a):** We show a uniform norm bound on  $\zeta_t(\theta, z_{l:t}) \forall \theta \in \Theta_R$ . First consider the following:

$$\begin{aligned} \|\delta_t(\theta)z_{l:t}\|_2 &= |\delta_t(\theta)| \|z_{l:t}\|_2 \leq |r_t + \gamma\phi(s'_t)^\top \theta - \phi(s_t)^\top \theta| \left\| \sum_{k=0}^{t-l} (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2 \\ &\leq |r_t + \|\phi(s'_t)\|_2 \|\theta\|_2 + \|\phi(s_t)\|_2 \|\theta\|_2| \left\| \sum_{k=0}^{\infty} (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2 \\ &\leq \frac{(r_{\max} + 2R)}{(1 - \gamma\lambda)} = B. \end{aligned}$$

Using this along with the fact that  $\|\theta - \theta^*\|_2 \leq 2R \leq B$  and  $\|\bar{x}(\theta)\|_2 \leq B$  for all  $\theta \in \Theta_R$ , we get

$$\begin{aligned} |\zeta_t(\theta, z_{l:t})| &= \left| (\delta_t(\theta)z_{l:t} - \bar{x}(\theta))^\top (\theta - \theta^*) \right| \leq \|\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\|_2 \|(\theta - \theta^*)\|_2 \\ &\leq (\|\delta_t(\theta)z_{l:t}\|_2 + \|\bar{x}(\theta)\|_2) \|(\theta - \theta^*)\|_2 \\ &\leq 2B\|(\theta - \theta^*)\|_2 \leq 2B^2. \end{aligned}$$

**Part (b):** To show that  $\zeta_t(\cdot, z_{l:t})$  is  $L$ -Lipschitz, consider the following inequality for any four vectors  $(a_1, b_1, a_2, b_2)$ , which follows as a direct application of Cauchy-Schwartz.

$$|a_1^\top b_1 - a_2^\top b_2| = |a_1^\top (b_1 - b_2) + b_2^\top (a_1 - a_2)| \leq \|a_1\|_2 \|b_1 - b_2\|_2 + \|b_2\|_2 \|a_1 - a_2\|_2.$$

This implies,

$$\begin{aligned}
|\zeta_t(\theta, z_{l:t}) - \zeta_t(\theta', z_{l:t})| &= \left| (\delta_t(\theta)z_{l:t} - \bar{x}(\theta))^\top (\theta - \theta^*) - (\delta_t(\theta')z_{l:t} - \bar{x}(\theta'))^\top (\theta' - \theta^*) \right| \\
&\leq \|\delta_t(\theta)z_{l:t} - \bar{x}(\theta)\|_2 \|\theta - \theta'\|_2 + \|\theta' - \theta^*\|_2 \|(\delta_t(\theta)z_{l:t} - \bar{x}(\theta)) - (\delta_t(\theta')z_{l:t} - \bar{x}(\theta'))\|_2 \\
&\leq 2B\|\theta - \theta'\|_2 + 2R \left[ \|z_{l:t}(\delta_t(\theta) - \delta_t(\theta'))\|_2 + \|\bar{x}(\theta) - \bar{x}(\theta')\|_2 \right] \\
&\leq 2B\|\theta - \theta'\|_2 + \frac{8R}{(1-\gamma\lambda)} \|\theta - \theta'\|_2 \\
&\leq 6B\|\theta - \theta'\|_2,
\end{aligned}$$

where the last inequality follows as  $\frac{R}{1-\gamma\lambda} \leq B/2$  by definition. In the penultimate inequality, we used that  $\|z_{l:t}(\delta_t(\theta) - \delta_t(\theta'))\|_2 \leq \frac{2}{(1-\gamma\lambda)} \|\theta - \theta'\|_2$  which is easy to prove. Consider,

$$\begin{aligned}
\|z_{l:t}(\delta_t(\theta) - \delta_t(\theta'))\|_2 &\leq \|z_{l:t}\|_2 |(\delta_t(\theta) - \delta_t(\theta'))| \\
&\leq \left\| \sum_{k=0}^{\infty} (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2 |(\delta_t(\theta) - \delta_t(\theta'))| \\
&\leq \frac{1}{(1-\gamma\lambda)} |(\gamma\phi(s'_t) - \phi(s_t))^\top (\theta - \theta')| \\
&\leq \frac{(\|\phi(s'_t)\|_2 + \|\phi(s_t)\|_2)}{(1-\gamma\lambda)} \|\theta - \theta'\|_2 \leq \frac{2}{(1-\gamma\lambda)} \|\theta - \theta'\|_2.
\end{aligned}$$

As  $\bar{x}(\theta) = \mathbb{E}[\delta_t(\theta)z_{-\infty:t}]$ , this also implies that  $\|\bar{x}(\theta) - \bar{x}(\theta')\|_2 \leq \frac{2}{(1-\gamma\lambda)} \|\theta - \theta'\|_2$  which completes the proof.

**Part (c):** To show that  $|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| \leq B^2(\gamma\lambda)^\tau$  for all  $\theta \in \Theta_R$  and  $\tau \leq t$ , we use that  $\|\theta - \theta^*\|_2 \leq 2R \leq B$ .

$$\begin{aligned}
|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{t-\tau:t})| &= \left| (\delta_t(\theta)z_{0:t} - \delta_t(\theta)z_{t-\tau:t})^\top (\theta - \theta^*) \right| \\
&\leq |\delta_t(\theta)| \|z_{0:t} - z_{t-\tau:t}\|_2 \|\theta - \theta^*\|_2 \\
&\leq |r_t + \gamma\phi(s'_t)^\top \theta - \phi(s_t)^\top \theta| \left\| \sum_{k=\tau}^{\infty} (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2 B \\
&\leq |r_t + 2\|\theta\|_2| \cdot \frac{(\gamma\lambda)^\tau}{(1-\gamma\lambda)} \cdot B \\
&\leq B \frac{(r_{\max} + 2R)}{(1-\gamma\lambda)} (\gamma\lambda)^\tau = B^2(\gamma\lambda)^\tau.
\end{aligned}$$

Similarly,

$$\begin{aligned}
|\zeta_t(\theta, z_{0:t}) - \zeta_t(\theta, z_{-\infty:t})| &\leq |\delta_t(\theta)(z_{0:t} - z_{-\infty:t})^\top (\theta - \theta^*)| \\
&\leq |\delta_t(\theta)| \|z_{0:t} - z_{-\infty:t}\|_2 \|\theta - \theta^*\|_2 \\
&\leq |(r_t + \gamma\phi(s'_t)^\top \theta - \phi(s_t)^\top \theta)| \left\| \sum_{k=t}^{\infty} (\gamma\lambda)^k \phi(s_{t-k}) \right\|_2 B \\
&\leq B \frac{(r_{\max} + 2R)}{(1-\gamma\lambda)} (\gamma\lambda)^t \leq B^2(\gamma\lambda)^t.
\end{aligned}$$

□

## C Proofs of Lemmas 13 and 16

In this section, we give a combined proof of Lemmas 13 and 16 which quantify the progress of the expected updates towards the limit point  $\theta^*$  for TD( $\lambda$ ) and the Q-function approximation algorithm. These lemmas can be stated more generally as shown below, instead of using the Bellman operators  $F(\cdot)$  and  $T^{(\lambda)}(\cdot)$ .

**Lemma 21.** *Let  $\Pi_D H(\cdot)$  be a contraction with respect to  $\|\cdot\|_D$  with modulus  $\gamma$  and let  $V_{\theta^*}$  be the unique fixed point of  $\Pi_D H(\cdot)$ , i.e.  $V_{\theta^*} = \Pi_D H V_{\theta^*}$ . Define  $\bar{g}(\theta) = \Phi^\top D(H\Phi\theta - \Phi\theta)$  for all  $\theta \in \mathbb{R}^d$  to be the expected update. Then,*

$$(\theta^* - \theta)^\top \bar{g}(\theta) \geq (1 - \gamma) \|V_{\theta^*} - V_\theta\|_D^2.$$

*Proof.* We have

$$\begin{aligned} (\theta^* - \theta)^\top \bar{g}(\theta) &= (\theta^* - \theta)^\top \Phi^\top D(H\Phi\theta - \Phi\theta) \\ &= \langle \Phi(\theta^* - \theta), (H\Phi\theta - \Phi\theta) \rangle_D \\ &= \langle \Pi_D \Phi(\theta^* - \theta), (H\Phi\theta - \Phi\theta) \rangle_D \end{aligned} \tag{50}$$

$$= \langle \Phi(\theta^* - \theta), \Pi_D (H\Phi\theta - \Phi\theta) \rangle_D \tag{51}$$

$$\begin{aligned} &= \langle \Phi(\theta^* - \theta), \Pi_D H\Phi\theta - \Phi\theta \rangle_D \\ &= \langle \Phi(\theta^* - \theta), \Pi_D H\Phi\theta - \Phi\theta^* + \Phi\theta^* - \Phi\theta \rangle_D \\ &= \|\Phi(\theta^* - \theta)\|_D^2 - \langle \Phi(\theta^* - \theta), \Phi\theta^* - \Pi_D H\Phi\theta \rangle_D \\ &\geq \|\Phi(\theta^* - \theta)\|_D^2 - \|\Phi(\theta^* - \theta)\|_D \cdot \|\Pi_D H\Phi\theta - \Phi\theta^*\|_D \\ &\geq \|\Phi(\theta^* - \theta)\|_D^2 - \gamma \cdot \|\Phi(\theta^* - \theta)\|_D^2 \end{aligned} \tag{52}$$

$$= (1 - \gamma) \cdot \|\Phi(\theta^* - \theta)\|_D^2 = (1 - \gamma) \cdot \|V_{\theta^*} - V_\theta\|_D^2,$$

where in going to Equation (50), we used that  $\forall \mathbf{x} \in \text{Span}(\Phi)$ , we have  $\Pi_D \mathbf{x} = \mathbf{x}$ . In Equation (51), we used that the projection matrix  $\Pi_D$  is symmetric. In going to Equation (52), we used that that  $\Pi_D H(\cdot)$  is a contraction operator with modulus  $\gamma$  with  $\Phi\theta^*$  as its fixed point, which implies that  $\|\Pi_D H\Phi\theta - \Phi\theta^*\|_D = \|\Pi_D H\Phi\theta - \Pi_D H\Phi\theta^*\|_D \leq \gamma \|\Phi\theta - \Phi\theta^*\|_D$ .  $\square$