Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation

Jalaj Bhandari , Daniel Russo , Raghav Singal

To cite this article:
Jalaj Bhandari , Daniel Russo , Raghav Singal (2021) A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation. Operations Research 69(3):950-973. https://doi.org/10.1287/opre.2020.2024

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 69, No. 3, May–June 2021, pp. 950–973
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation

Jalaj Bhandari,[a] Daniel Russo,[b] Raghav Singal[a]

[a] Operations Research, Columbia University, New York, New York 10027; [b] Graduate School of Business, Columbia University, New York, New York 10027

Contact: jb3618@columbia.edu (JB); djr2174@gsb.columbia.edu, https://orcid.org/0000-0001-5926-8624 (DR); rs3566@columbia.edu, https://orcid.org/0000-0001-9277-7383 (RS)

Received: October 31, 2018
Revised: July 23, 2019
Accepted: April 1, 2020
Published Online in Articles in Advance: March 19, 2021

Subject Classifications: dynamic programming/optimal control, decision analysis: sequential
Area of Review: Machine Learning and Data Science

https://doi.org/10.1287/opre.2020.2024

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(λ), and to Q-learning applied in high-dimensional optimal stopping problems.

Supplemental Material: The online appendices are available at https://doi.org/10.1287/opre.2020.2024.

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.

Although 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 are 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 nonasymptotic 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 here.

- *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 independent and identically distributed (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 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.

- *Nonasymptotic treatment with Markovian noise:* TD is usually applied online to a single Markovian data stream. However, to our knowledge, there has been no successful[1] nonasymptotic 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. 2009a, b; Liu et al. 2015; Dalal et al. 2018b; Lakshminarayanan and Szepesvári 2018; Touati et al. 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 TD 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(λ). 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 because of 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 nonlinear stochastic approximation. To our knowledge, no prior work has provided nonasymptotic guarantees for either TD(λ) or Q-learning with function approximation.

1.2. Related Literature

1.2.1. Nonasymptotic Analysis of TD(0). There has been very little nonasymptotic analysis of TD(0). To our knowledge, Korda and Prashanth (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[2] of $O(1/T^{\sigma})$. Unfortunately, although the analysis is technically nonasymptotic, the constant factors in the bound display a complex dependence on the problem instance and scale with some unusual quantities which can be very large in cases of practical interest.

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 approach dates back to the works of Ruppert (1988), Polyak and Juditsky (1992), and Györfi and Walk (1996), which shows 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 nonasymptotic 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. This is not because of any shortcoming in the techniques of Bach and Moulines (2013) and Lakshminarayanan and

Szepesvári (2018). Instead, under non-i.i.d. noise and a linear stochastic approximation algorithm applied with any constant step size, the averaged iterate might converge to the wrong limit as shown in a simple example by Györfi and Walk (1996).

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 8–10.

Although 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(λ) algorithm. 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 a 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(λ), in which it is equivalent to Monte Carlo policy evaluation (Sutton and Barto 1998).

1.2.2. Asymptotic Analysis of Stochastic Approximation. There is a well-developed theory around asymptotic analysis 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 Ordinary Differential Equation (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 because it completely circumvents issues with noise in the system and issues of step-size selection. However, 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).

In addition to such asymptotic techniques, the modern literature on first-order stochastic optimization also focuses heavily on nonasymptotic analysis (Bubeck 2015, Jain and Kar 2017, Bottou et al. 2018). One reason is that such asymptotic analysis necessarily focuses on a regime where step sizes are negligibly 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 interested 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 nonasymptotic 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.

1.2.3. 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 Szepesvári (2012), Prashanth et al. (2014), 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 because of significant computational advantages (Sutton and Barto 1998).

Gradient TD methods are another related class of algorithms. These were derived by Sutton et al. (2009a, b) to address the issue that TD can diverge in so-called *off-policy* settings, where data are 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. (2009a, b) 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. Alternatively, Macua et al. (2014) and Liu et al. (2015) propose to reformulate the original gradient TD optimization as a primal-dual saddle point problem and leverage convergence analysis from that literature to give a nonasymptotic 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(λ), and the Q-learning studied in this paper. Unlike gradient TD methods, they do not mimic noisy gradient steps with respect to any fixed objective.[3]

2. Problem Formulation

2.1. Markov Reward Process

We consider the problem of evaluating the value function V_μ of a given policy μ in a Markov decision process (MDP). We work in the on policy setting, where data are generated by applying the policy μ in the MDP. Because the policy μ is applied automatically to select actions, such problems are most naturally formulated as value function estimation in a Markov reward process (MRP). An MRP[4] 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.

The value function associated with this MRP, V_μ, 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_μ associates a value function $V : \mathcal{S} \to \mathbb{R}$ 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

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

Under this assumption, value functions are assured to exist and are the unique solution to Bellman's equation (Bertsekas 1995). We also assume that the Markov reward process induced by following the policy μ is ergodic with a unique stationary distribution π. For any two states s, s': $\pi(s') = \lim_{t \to \infty} \mathbb{P}(s_t = s' | s_0 = s)$.

Following common references (Bertsekas 1995, De Farias and Van Roy 2003, Dann et al. 2014), we will simplify the presentation by assuming the state space $\mathcal{S}$ is a finite set of size $n = |\mathcal{S}|$. 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, as long the uniform ergodicity condition stated in Assumption 1 continues to hold. For problems with general state space, even the core results in dynamic programming hold only under suitable technical conditions (Bertsekas and Shreve 1978).

2.2. Value Function Approximation

Given a fixed policy μ, the problem is to efficiently estimate the corresponding value function V_μ using only the observed rewards and state transitions. Unfortunately, because of 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 nonlinear 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 $\mathcal{S} = \{s_1, \ldots, s_n\}$, $V_\theta \in \mathbb{R}^n$ 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 Φ are linearly independent.

2.3. Norms in Value Function and Parameter Space

For a symmetric positive definite matrix A, define the inner product $\langle x, y\rangle_A = x^\top Ay$ and the associated norm $\|x\|_A = \sqrt{x^\top Ax}$. If A is positive semidefinite rather than positive definite then $\|\cdot\|_A$ is called a seminorm. Let $D = \mathrm{diag}(\pi(s_1), \ldots, \pi(s_n)) \in \mathbb{R}^{n\times n}$ 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 $\theta, \theta' \in \mathbb{R}^d$,

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

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.

2.4. Feature Regularity

We assume that the feature vectors are uniformly bounded, that is $\sup_{s\in\mathcal{S}} \|\phi(s)\|_2 < \infty$. For notational convenience, we also assume features are normalized so that $\|\phi(s)\|_2 \le 1$ for all $s \in \mathcal{S}$. This is without loss of generality because the TD algorithm is invariant to feature rescaling. Precisely, TD applied with feature mapping $\phi(\cdot)$ and initial parameter θ_0 produces an identical sequence of value functions to the TD algorithm with feature mapping $\tilde{\phi}(\cdot) = k\phi(\cdot)$ and initial parameter $\tilde{\theta}_0 = \theta_0/k$, for any scalar $k > 0$. All our results bound the mean-squared gap between value predictions. We also assume that any entirely redundant or irrelevant features have been removed, so Σ has full rank. Let $\omega > 0$ be the minimum eigenvalue of Σ. From our bound on the feature vectors, the maximum eigenvalue of Σ is less than 1, so $1/\omega$ bounds the condition number of the feature covariance matrix.[5] The following lemma is an immediate consequence of our assumptions.

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

One typical style of result in the study of strongly convex optimization gives fast rates of convergence in terms of the number of iterations T. However, these bounds degrade when ω is very small and generally require a priori knowledge of some good lower bound on ω. We give some results in that style, but also give results in the style of Nemirovski et al. (2009), where bounds and step sizes have no dependence on ω.

3. Temporal Difference Learning

We consider the classic temporal difference learning algorithm (Sutton 1988). The algorithm starts with an initial parameter estimate θ_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 μ in the current state. This tuple is used to define a loss function, which is taken to be the squared sample Bellman error. The algorithm then proceeds to compute the next iterate θ_{t+1} by making a gradient-like update. 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.

TD is not a true stochastic gradient method with respect to any fixed loss function, which makes its analysis challenging. The TD update can be written as $g_t(\theta) = (y_t - V_\theta(s_t))\frac{d}{d\theta}V_\theta(s_t)$, where $y_t = r_t + \gamma V_\theta(s_t')$ is sample based estimate of the Bellman update to V_{θ_t}. Then $g_t(\theta_t) = -\frac{\partial}{\partial\theta}\frac{1}{2}(y_t - V_\theta(s_t))^2|_{\theta=\theta_t}$ can be interpreted as the negative gradient of a certain squared loss function, but this calculation treats the target y_t as fixed and ignores its implicit dependence on θ_t. To emphasize the contrast with stochastic gradient methods, Sutton and Barto (1998) refer to TD as a *semigradient* method. Accordingly, we will refer to $g_t(\cdot)$ as *negative semigradient* throughout the paper.

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 semigradient 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 Bradtke and Barto (1996).

Algorithm 1 TD(0) with Linear Function Approximation

Input: initial guess θ_0, step-size sequence $\{\alpha_t\}_{t\in\mathbb{N}}$.
Initialize: $\bar{\theta}_0 \leftarrow \theta_0$.
for $t = 0, 1, \ldots$ **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 op` *
Define loss function:
$\frac{1}{2}(y_t - V_\theta(s_t))^2$; * `sample Bellman error` *
Compute negative semigradient:
$g_t(\theta_t) = -\frac{\partial}{\partial\theta}\frac{1}{2}(y_t - V_\theta(s_t))^2|_{\theta=\theta_t}$;
Take a semigradient step:
$\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$; * α_t:`step-size` *
Update averaged iterate:
$\bar{\theta}_{t+1} \leftarrow (\frac{t}{t+1})\bar{\theta}_t + (\frac{1}{t+1})\theta_t$; * $\bar{\theta}_{t+1} = \frac{1}{t+1}\sum_{\ell=0}^{t}\theta_\ell$ *
End

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

$$g_t(\theta) = \big(r_t + \gamma\phi(s_t')^\top\theta - \phi(s_t)^\top\theta\big)\phi(s_t). \tag{1}$$

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

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

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

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

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, because $\sum_{s'\in\mathcal{S}}\mathcal{P}(s'|s)(\mathcal{R}(s,s') + \gamma\phi(s')^\top\theta) = (T_\mu\Phi\theta)(s)$, we can recognize that

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

Tsitsiklis and Van Roy (1997) provides a derivation of this fact.

4. Asymptotic Convergence of Temporal Difference Learning

The main challenge in analyzing TD is that the semigradient steps $g_t(\theta)$ are not true stochastic gradients with respect to any fixed objective. The semigradient 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_{θ_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 Σ is full rank, there is a unique[6] vector θ^* with $\bar{g}(\theta^*) = 0$. We briefly review results that offer insight into θ^* and proofs of the asymptotic convergence of TD.

4.1. Understanding the TD Limit Point

Tsitsiklis and Van Roy (1997) give an interesting characterization of the limit point θ^*. They show it is the unique solution to the projected Bellman equation:

$$\Phi\theta = \Pi_D T_\mu\Phi\theta, \tag{4}$$

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 θ^*, 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 θ^* 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 as a simple stochastic approximation scheme for solving the projected-Bellman fixed point equation.

Lemma 2 (Tsitsiklis and Van Roy 1997). *The projected Bellman operator $\Pi_D T_\mu(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus γ, that is,*

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

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

$$\big\|V_{\theta^*} - V_\mu\big\|_D \le \frac{1}{\sqrt{1-\gamma^2}}\big\|\Pi_D V_\mu - V_\mu\big\|_D. \tag{5}$$

Chapter 6 in Bertsekas (1995) provides 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 Φ. If V_μ actually falls within the span of the

features, there is no approximation error at all and TD converges to the true value function.

4.2. 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 θ^*.

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 sidestepping 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 on $\mathbb{E}[\|V_{\bar{\theta}_T} - V_{\theta^*}\|_D^2]$, which measures the mean-squared gap between the value predictions under the averaged-iterate $\bar{\theta}_T$ and under the TD limit point θ^*. We first derive bounds that depend on the condition number of the feature covariance matrix. 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 V_{θ^*} at an exponential rate and $O(1/T)$ convergence rates with appropriate decaying step sizes.

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.[7] Following the spirit of this work, we also give explicit bounds on $\mathbb{E}[\|V_{\bar{\theta}_T} - V_{\theta^*}\|_D^2]$ with a slower $O(1/\sqrt{T})$ convergence rates, but importantly, 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 here.

Noiseless Case

Drawing inspiration from the ODE method discussed previously, 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 are collected from a single sample path of an ergodic Markov chain. This setting introduces significant challenges because of 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(λ)

In Section 9, we extend the analysis under the Markov noise to TD with eligibility traces, popularly known as TD(λ). Eligibility traces are known to often provide performance gains in practice, but theoretical analysis is more complex. Our analysis also 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

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

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 section, we review a simple analysis of gradient descent. In Section 6.2, we establish key properties of mean-path TD mirroring those used to analyze this gradient descent algorithm. Finally, Section 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, γ.

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 θ and those under the stationary point of TD, θ^*. 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 θ^* 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 θ_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 $(\theta^* - \theta_t)$. 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

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

In addition, one can show[8]

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

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

$$\|\theta^* - \theta_{t+1}\|_2^2 \le \|\theta^* - \theta_t\|_2^2 - \|V_{\theta^*} - V_{\theta_t}\|_D^2. \tag{8}$$

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

$$\|\theta^* - \theta_{t+1}\|_2^2 \le (1-\omega)\|\theta^* - \theta_t\|_2^2 \le \ldots \le (1-\omega)^{t+1}\|\theta^* - \theta_0\|_2^2. \tag{9}$$

Recall that ω denotes the minimum eigenvalue of Σ. This shows that error converges at a fast geometric rate. However, the rate of convergence degrades if the minimum eigenvalue ω 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 ω. 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

$$\left\|V_{\theta^*} - V_{\bar{\theta}_T}\right\|_D^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\|V_{\theta^*} - V_{\theta_t}\|_D^2 \le \frac{\|\theta^* - \theta_0\|_2^2}{T}. \tag{10}$$

6.2. Key Properties of Mean-Path TD

This section establishes analogues for mean-path TD of the key properties (6) and (7) used to analyze gradient descent. First, to characterize the semigradient 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 that

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

That is, the expected update of TD always forms a positive angle with $(\theta^* - \theta)$. Although 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, establishes that the expected updates of TD point in a descent direction of $\|\theta^* - \theta\|_2^2$, and do so more strongly when the gap between value functions under θ and θ^* is large. We will show that this more quantitative form of (11) allows for elegant finite time bounds on the performance of TD.

This lemma mirrors the property in Equation (6), but with a smaller constant of $(1 - \gamma)$. This reflects that expected TD must converge to θ^* by bootstrapping (Sutton 1988) and may follow a less direct path to θ^*

than the fictitious gradient descent method considered in the previous subsection. Recall that the limit point θ^* solves $\bar{g}(\theta^*) = 0$.

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

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

Proof of Lemma 3. We use the notation described in Equation (2). 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, ξ and ξ' are two correlated random variables with the same marginal distribution. By definition, $\mathbb{E}[\xi^2] = \|V_{\theta^*} - V_\theta\|_D^2$, because s is drawn from π.

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

$$\begin{aligned}\bar{g}(\theta) = \bar{g}(\theta) - \bar{g}(\theta^*) &= \mathbb{E}\left[\phi(\gamma\phi' - \phi)^\top(\theta - \theta^*)\right] \\ &= \mathbb{E}\left[\phi(\xi - \gamma\xi')\right]. \end{aligned} \tag{12}$$

Therefore,

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

Here we use the Cauchy-Schwartz inequality together with the fact that ξ and ξ' have the same marginal distribution to conclude that $\mathbb{E}[\xi\xi'] \leq \sqrt{\mathbb{E}[\xi^2]}\sqrt{\mathbb{E}[(\xi')^2]} = \mathbb{E}[\xi^2]$. □

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

Lemma 4. *For all* $\theta \in \mathcal{R}^d$, $\|\bar{g}(\theta)\|_2 \leq 2\|V_\theta - V_{\theta^*}\|_D$.

Proof of Lemma 4. Beginning from (12) in the proof of Lemma 3, we have

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

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

Lemmas 3 and 4 are quite powerful when used in conjunction. As in the analysis of gradient descent reviewed in the previous section, our analysis starts with a recursion for the error term, $\|\theta_t - \theta^*\|^2$. See Equation (13) in Theorem 1. 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 such as 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.[9]

6.3. Finite Time Analysis of Mean-Path TD

We now combine the insights of the previous section 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 ω of Σ is not too small, the geometric convergence rate given in the second part of the theorem dominates.

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

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

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

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

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

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

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

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

Then,

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

Applying Jensen's inequality gives the first result:

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

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

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

where the final inequality uses that $(1-\frac{\omega(1-\gamma)^2}{4}) \le e^{\frac{-\omega(1-\gamma)^2}{4}}$. Repeating this inductively and using that $\|V_{\theta^*}-V_{\theta_T}\|_D^2 \le \|\theta^*-\theta_T\|_2^2$ as shown in Lemma 1 gives the desired result. □

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',

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

and the tuples $\{(s_t, r_t, s_t')\}_{t\in\mathbb{N}}$ are drawn independently across time. 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. (2009a, b), Korda and Prashanth (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 θ^*. 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 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. The requirement that $\sqrt{T} \ge 8/(1-\gamma)$ 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 value function V_{θ_t} will never converge to the TD fixed point, but our results show the expected distance to V_{θ^*} 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 V_{θ_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 Σ, 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). The bound in part (c) is only meaningful when T is large relative to $1/\omega$ and $(1-\gamma)^{-1}$. We suspect this is because of fundamental challenges in applying TD to problems with poor conditioning or long time horizons, but it would be interesting to formally validate this.

We should note that step sizes were chosen to enable a convenient finite time analysis. Alternative choices may lead to stronger bounds and better practical performance. As in Bottou et al. (2018), our results in parts (b) and (c) could be modified so that the step sizes and final bound depend on some underestimate $\omega' < \omega$, of the true minimum eigenvalue ω. However, the challenge of setting such step sizes is one of the major reasons Nemirovski et al. (2009) advocate instead for results like those in part (a) of Theorem 2. It is also worth noting that our analysis in part (a) can be extended to decreasing step sizes of the form $\alpha_t = \min\{(1-\gamma)/8, 1/\sqrt{t}\}$, at the expense of slightly worse constants. Such extensions are common in the optimization literature. See, for example, corollary 3.2.8 of Duchi (2018). Recall that $\bar{\theta}_T = T^{-1}\sum_{t=0}^{T-1}\theta_t$ denotes the averaged iterate. We show the following result.

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* $\alpha_0 = \cdots = \alpha_T = \frac{1}{\sqrt{T}}$,

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

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

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

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

$$\mathbb{E}\left[\|V_{\theta^*} - V_{\theta_T}\|_D^2\right] \leq \frac{\nu}{\lambda + T},$$

where

$$\nu = \max\left\{\frac{8\sigma^2}{(1-\gamma)^2\omega^2}, \frac{16\|\theta^* - \theta_0\|_2^2}{(1-\gamma)^2\omega}\right\}.$$

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

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 of Lemma 5. 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, ξ and ξ' 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

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

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

Using this, we give a proof of Theorem 2. Let us remark here on a consequence of the i.i.d noise model that considerably simplifies the proof. Until now, we 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, θ_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 θ_t, we can seamlessly apply Lemmas 3 and 5, as is done in inequality (16) of the proof.

Proof of Theorem 2. 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 Lemmas 3 and 5 imply

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

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, θ_t.

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

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

We find

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

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

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

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

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

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

Proof of Part (c)**.** By the definitions of ν, λ and β, we have

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

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

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

where we use that $\beta = \frac{2}{(1-\gamma)\omega}$. The final inequality follows by using that uses that $2\beta^2\sigma^2 - \nu \le 0$, which holds by the definition of ν and the fact that $\hat{t}^2 \ge (\hat{t} - 1)(\hat{t} + 1)$. The final result follows by invoking the inequality $\|V_{\theta^*} - V_{\theta_T}\|_D^2 \le \|\theta^* - \theta_T\|_2^2$ as shown in Lemma 1. □

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 MRP. However, 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 $(s_0, s_1, \ldots s_t, \ldots)$, 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 because of strong dependencies between the noisy observations. To understand this, recall the expression of the TD update

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

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 $\bar{g} : \mathbb{R}^d \to \mathbb{R}^d$ by $\bar{g}(\theta) = \mathbb{E}[g(\theta, O_t)]$, where, because θ is nonrandom, this expectation integrates over the marginal distribution of the tuple O_t. However, $\mathbb{E}[g(\theta_t, O_t) \mid \theta_t = \theta] \ne \bar{g}(\theta)$ because θ_t is a function of past tuples $\{O_1, \ldots, O_{t-1}\}$, potentially introducing strong dependencies between θ_t and O_t. Similarly, in general $\mathbb{E}[g(\theta_t, O_t) - \bar{g}(\theta_t)] \ne 0$, indicating bias in the algorithm's semigradient evaluations. A related challenge arises in trying to control the norm of the semigradient step, $\mathbb{E}[\|g_t(\theta_t)\|_2^2]$. Lemma 5 does not yield a bound because of coupling between the iterate θ_t and the observation O_t.

Our analysis uses an information-theoretic technique to control for this coupling and explicitly account for the semigradient bias. 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 later.

Projected TD Algorithm:

Our technique for controlling the semigradient 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 semigradient steps do not explode. For tractability, we modify the TD algorithm itself by adding a projection step that ensures semigradient norms are uniformly bounded across time. In particular, starting with an initial guess of θ_0 such that $\|\theta_0\|_2 \le R$, we consider the projected TD algorithm, which iterates

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

where

$$\Pi_{2,R}(\theta) = \operatorname*{argmin}_{\theta':\|\theta'\|_2\le R} \|\theta - \theta'\|_2$$

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 Π_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 $\Pi_{2,R}(\theta) = R\theta/\|\theta\|$ if $\|\theta\|_2 > R$ and is simply θ otherwise. In addition, Section 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 projection steps to control the norm of the gradient (Nemirovski et al. 2009, Kushner 2010, Lacoste-Julien et al. 2012, Bubeck 2015).

Structural Assumptions on the Markov Reward Process:

To control the statistical bias in the semigradient 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 here.

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

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

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

This uniform mixing assumption always holds for irreducible and aperiodic finite-state Markov chains (Levin and Peres 2017). Meyn and Tweedie (2012) and Roberts and Rosenthal (2004) provide a discussion on uniform ergodicity and relaxations of this concept in general state space Markov chains. We emphasize that assuming the Markov 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

$$\tau^{\mathrm{mix}}(\epsilon) = \min\{t \in \mathbb{N}_0 \mid m\rho^t \le \epsilon\}. \tag{21}$$

For interpreting the bounds, note that from Assumption 1:

$$\tau^{\mathrm{mix}}(\epsilon) \sim \frac{\log(m/\epsilon)}{\log(1/\rho)} \qquad \text{as } \epsilon \to 0.$$

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 Semigradient:

Before proceeding, we also state a bound on the Euclidean norm of the semigradient under TD(0) that follows from the uniform bound on rewards, along with feature normalization[10] and boundedness of the iterates through the projection step. Under projected TD(0) with projection radius R, this lemma implies that $\|g_t(\theta_t)\|_2 \le (r_{\max} + 2R)$. This semigradient bound plays an important role in our convergence bounds.

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

Proof of Lemma 6. Using the expression of $g_t(\theta)$ in Equation (19), we have

$$\begin{aligned}\|g_t(\theta)\|_2 &\le |r_t + (\gamma\phi(s_t') - \phi(s_t))^\top\theta|\,\|\phi(s_t)\|_2 \\ &\le r_{\max} + \|\gamma\phi(s_t') - \phi(s_t)\|_2\|\theta\|_2 \\ &\le r_{\max} + 2\|\theta\|_2. \quad \square\end{aligned}$$

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 Σ. The bound in part (c) gives a faster convergence rate in terms of the number of samples T, but the bound and the step-size sequence depend on the minimum eigenvalue ω of Σ.

Part (b) confirms that for sufficiently small step sizes, the value functions converge at an exponential rate to within some radius of the TD fixed point V_{θ^*}.

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 because of semigradient 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 semigradient noise term σ^2 there would be replaced by G^2. This is a consequence using G as a uniform upper bound on the semigradient norm in the proof, which is possible because of the projection step. Astute readers may notice the stepsize choices in parts (b) and (c) differ from those in parts (b) and (c) of Theorem 2. For each result, we have aimed for step-size choices that lead to the simplest proofs of strong finite time bounds. In Theorem 3, the projection step allowed us to give a simple proof without requiring as small a step size as in Theorem 2. This choice may reflect our analysis technique more than any fundamental differences between the problem settings.

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 $\alpha_0 = \cdots = \alpha_T = 1/\sqrt{T}$,*

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

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

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

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

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

There are two noteworthy points here. First, the proof of part (c) also implies an $\tilde{\mathcal{O}}(1/T)$ convergence rate for the value function V_{θ_T} itself; however, the bound degrades by a factor of ω. We refer the readers to Equation (EC.4) in Online Appendix B.2 for the complete result. Second, 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. 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 θ^* is unknown? It turns out we can use Lemma 2, which relates the value function at the limit of convergence V_{θ^*} to the true value function, to give a conservative upper bound. This is shown in the following lemma.

Lemma 7. *We have the following bounds on the TD limit point,*

$$\|\theta^*\|_\Sigma \leq \frac{2r_{\max}}{(1-\gamma)^{3/2}}$$

and hence

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

Proof of Lemma 7. See Online Appendix C for a detailed proof. □

It is important to remark here that this bound is problem dependent as it depends on the minimum eigenvalue ω of the steady-state feature covariance matrix Σ. We believe that estimating ω online would make the projection step practical to implement. We also remark that, although we have assumed for that feature vectors are bounded as $\|\phi(s)\|_2 \leq 1$, this is not required for the conclusion in Lemma 7. The required projection radius automatically reflects any scaling of the feature vectors through the minimum eigenvalue ω.

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 $(\alpha_0, \ldots, \alpha_T)$.

We fix some notation throughout the scope of 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. Set $G = (r_{\max} + 2R)$, so $\|g_t(\theta)\|_2 \leq G$ for all $\theta \in \Theta_R$ by Lemma 6. Finally, we set

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

which can be thought of as the error in the evaluation of semigradient-update under parameter θ 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 $\alpha_0 = \cdots = \alpha_T = \alpha$, we show

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

The first inequality follows from Lemma 8. The second follows from Lemma 11, which in the case of a constant step size α shows $\mathbb{E}[\alpha\zeta_t(\theta_t)] \le G^2(4 + 6\tau^{\text{mix}}(\alpha))\alpha^2$. Notice that bias in the semigradient enters into the analysis as if by scaling the magnitude of the noise in semigradient 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* $t \in \mathbb{N}_0$,

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

Proof of Lemma 8. From the projected TD(0) recursion in Equation (20), for any $t \in \mathbb{N}_0$,

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

The first inequality used that orthogonal projection operators onto a convex set are nonexpansive,[11] the second used Lemma 6 together with the fact that $\|\theta_t\|_2 \le R$ because of projection, and the third used Lemma 3. □

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 $\mathbb{E}[\zeta_t(\theta_t)] \ne 0$, because of bias in the semigradient evaluations.

8.3.2. Information–Theoretic Techniques for Controlling the Semigradient Bias. The uniform mixing condition in Assumption 1 can be used in conjunction with some information theoretic inequalities to control the magnitude of the semigradient 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:

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

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

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

Here, we use the notation $X \to Y \to Z$, 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. By symmetry we also have $I_f(X,Z) \le I_f(Y,Z)$. To use these results in conjunction with Assumption 1, we can

specialize to total variation distance (d_{TV}) and total-variation mutual information (I_{TV}) using $f(x) = |x-1|/2$. The total variation distance is especially useful for our purposes because of the following variational representation.

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

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 $f : \mathcal{X} \to \mathbb{R}$. The proof uses similar ideas to Russo and Zou (2019) and Xu and Raginsky (2017).

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

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

for some fixed $t \in \{0,1,2,\ldots\}$ 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,

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

Proof of Lemma 9. 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)

$$\begin{aligned}\mathbb{E}[h(X,Y)] - \mathbb{E}[h(X',Y')] &= \int h dP - \int h dQ \\ &\le d_{\mathrm{TV}}(P,Q) = I_{\mathrm{TV}}(X,Y),\end{aligned}$$

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

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

where the 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, we get

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

which gives us the desired result. □

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

Lemma 10 (Semigradient Error Is Bounded and Lipschitz)**.** *With probability 1,*

$$|\zeta_t(\theta)| \le 2G^2 \qquad \textit{for all } \theta \in \Theta_R$$

and

$$|\zeta_t(\theta) - \zeta_t(\theta')| \le 6G\|(\theta - \theta')\|_2 \qquad \textit{for all } \theta, \theta' \in \Theta_R.$$

Proof of Lemma 10. The result follows from a straightforward application of the bounds $\|g_t(\theta)\|_2 \le G$ and $\|\theta\|_2 \le R \le G/2$, which hold for each $\theta \in \Theta_R$. A full derivation is given in Online Appendix A.3. □

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

Lemma 11 (Bound on Semigradient Bias)**.** *Consider a nonincreasing step-size sequence, $\alpha_0 \ge \alpha_1 \ldots \ge \alpha_T$. Fix any $t < T$, and set $t^* \equiv \max\{0, t - \tau^{\mathrm{mix}}(\alpha_T)\}$. Then,*

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

The following bound also holds:

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

Proof of Lemma 11. 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 $i \in \mathbb{N}_0$,

$$\begin{aligned}\|\theta_{i+1} - \theta_i\|_2 &= \big\|\Pi_{2,R}\big(\theta_i + \alpha_i g_i(\theta_i)\big) - \Pi_{2,R}(\theta_i)\big\|_2 \\ &\le \|\theta_i + \alpha_i g_i(\theta_i) - \theta_i\|_2 = \alpha_i \|g_i(\theta_i)\|_2 \\ &\le \alpha_i G.\end{aligned}$$

Therefore,

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

Applying Lemma 10, we conclude

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

Step 2. Bound $\mathbb{E}[\zeta_t(\theta_{t-\tau})]$ using Lemma 9.
Recall that the semigradient $g_t(\theta)$ 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. For any $\theta \in \Theta_R$, put

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

and

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

We have defined $\bar{g} : \Theta_R \to \mathbb{R}^d$ 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 (nonrandom) $\theta \in \Theta_R$,

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

Because $\theta_0 \in \Theta_R$ is nonrandom, it follows immediately that

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

We use Lemma 9 to bound $\mathbb{E}[\zeta(\theta_{t-\tau}, O_t)]$. 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,

$$\mathbb{E}[\zeta(\theta'_{t-\tau}, O'_t)] = \mathbb{E}[\mathbb{E}[\zeta(\theta'_{t-\tau}, O'_t) \,|\, \theta'_{t-\tau}]] = 0.$$

Because $|\zeta(\theta, O_t)| \le 2G^2$ for all $\theta \in \Theta_R$ by Lemma 10 and $\theta_{t-\tau} \to s_{t-\tau} \to s_t \to O_t$ forms a Markov chain, applying Lemma 9 gives

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

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

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

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

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

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

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

The second inequality used that $m\rho^{\tau_0} \le \alpha_T$ by the definition of the mixing time $\tau_0 \equiv \tau^{\text{mix}}(\alpha_T)$ and the final inequality uses that step sizes are nonincreasing. □

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 section, 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 Online 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 α_t terms present in analyses of SGD. A complete proof of part (c) is given in Online 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 TD(λ) and parameterized by $\lambda \in [0, 1]$, that contains as a special case the TD(0) algorithm studied in previous sections.[12] For $\lambda > 0$, the algorithm maintains an eligibility trace vector, which is a geometric weighted average of the negative semigradients at all previously visited states, and makes parameter updates in the direction of the eligibility vector rather than the negative semigradient. 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 nonasymptotic analysis of TD(λ).

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 TD(λ).

9.1. Projected TD(λ) Algorithm

TD(λ) makes a simple, but a highly consequential, modification to TD(0). The pseudo-code for this algorithm is presented below in Algorithm 2. As with TD(0), it observes a tuple $O_t = (s_t, r_t = \mathcal{R}(s_t, s_{t+1}), s_{t+1})$ at each time-step t 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 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, TD(λ) 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 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 TD(λ) with Linear Function Approximation

Input: radius R, initial guess $\{\theta_0 : \|\theta_0\|_2 \le R\}$, and step-size sequence $\{\alpha_t\}_{t\in\mathbb{N}}$

Initialize: $\bar{\theta}_0 \leftarrow \theta_0$, $z_{-1} = 0$, $\lambda \in [0,1]$.

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

Observe tuple: $O_t = (s_t, r_t, s_{t+1})$;

Get TD error:
$\delta_t(\theta_t) = r_t + \gamma V_{\theta_t}(s_{t+1}) - V_{\theta_t}(s_t)$;

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:
$\theta_{t+1} = \Pi_{2,R}(\theta_t + \alpha_t x_t(\theta_t, z_{0:t}))$; * α_t:`step-size` *

Update averaged iterate:
$\bar{\theta}_{t+1} \leftarrow (\frac{t}{t+1})\bar{\theta}_t + (\frac{1}{t+1})\theta_{t+1}$; * $\bar{\theta}_{t+1} = \frac{1}{t+1}\sum_{\ell=1}^{t+1}\theta_\ell$ *

end

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 θ at time t. This plays a role analogous to the negative semigradient $g_t(\theta)$ in TD(0).

9.2. Limiting Behavior of TD(λ)

We now review results on the asymptotic convergence of TD(λ) from 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 μ is applied to the MDP. The asymptotic properties of TD(λ) are closely tied to a geometrically weighted version of the k-step Bellman equations described above. Define the averaged Bellman operator

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

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(λ) converges almost surely to the unique solution, θ^* of the projected fixed point equation

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

TD(λ) is then interpreted as a stochastic approximation scheme for solving this fixed point equation. The existence and uniqueness of such a fixed point 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 $\|\cdot\|_D$. Throughout this section, we let θ^* denote the unique fixed point of the projected TD(λ) operator as shown previously.

Lemma 12 (Tsitsiklis and Van Roy 1997)**.** *The projected TD(λ) operator $\Pi_D T_\mu^{(\lambda)}(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus*

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

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

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

Chapter 6 of Bertsekas (1995) provides a proof. It is important to note the distinction between the convergence results for TD(λ) and TD(0) in terms of the contraction factors. The contraction factor κ is always less than γ, the contraction factor under TD(0). In addition, as $\lambda \to 1$, $\kappa \to 0$ implying that the limit point of TD(λ) for large enough λ 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 λ, and may degrade as λ grows. Disentangling such issues requires also a careful study of the statistical efficiency of TD(λ), which we undertake in the following section.

9.3. Finite Time Bounds for Projected TD(λ)

Following Section 8, we establish three finite time bounds on the performance of the projected TD(λ) 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) shows an exponential rate of convergence to within some radius of the TD(λ) fixed-point under a sufficiently small step size. 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 ω of Σ.

Compared with 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

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

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

Theorem 4. *Suppose the projected TD(λ) algorithm is applied with parameter $R \ge \|\theta^*\|_2$ 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 $\alpha_t = \alpha_0 = 1/\sqrt{T}$,*

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

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

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

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

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

As was the case for TD(0), the proof of part (c) also implies an $\tilde{\mathcal{O}}(1/T)$ convergence rate for the value function V_{θ_T} itself; however, the bound degrades by a factor of ω. We refer the readers to Equation (EC.4) in Online Appendix B.2 for the complete result. 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(λ) with that of TD(0) ignoring the constant terms. It should be emphasized that these are only upper bounds, so differences could be because of looseness of the analysis rather than true differences in statistical performance. First, let us look at the results for the robust step size $\alpha_t = 1/\sqrt{T}$ in part (a) of Theorems 3 and 4. Approximately, for the TD(λ) 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:

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

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(λ). Correspondingly, from Section 8, we know that G is the upper bound on semigradient norm, $g_t(\theta_t)$ for TD(0). Intuitively, for TD(λ), the bound B is larger (because of the presence of the eligibility trace term) and more so as $\lambda \to 1$. This calculation reveals that our bounds give a slower rate of convergence for TD(λ) than for TD(0). This means more data are required for our bound to guarantee TD(λ) is close to its limit point. In this context, however, the tradeoff we remarked on in Section 9.2 is noteworthy as the fixed point for TD(λ) comes with a better error guarantee.

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

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

It is unclear whether the difference between the two step-size regimes is an artifact of our analysis technique.

10. Extension: Q-Learning for High-Dimensional Optimal Stopping

Thus 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 π. 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 τ denote the (random) time at which the decision maker stops. The expected total discounted reward from initial state s associated with the stopping time τ is

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

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 to maximize (31).

For any Markov decision process, the optimal state-action value function $Q^* : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ 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 $Q^* : \mathcal{S} \to \mathbb{R}$ denote the optimal 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 given by

$$\tau^* = \min\{t \,|\, U(s_t) \geq Q^*(s_t)\}. \tag{32}$$

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 continual value 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, $\mathcal{S} = \{s_1, \ldots, s_n\}$, $Q_\theta \in \mathbb{R}^n$ 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 feature vectors $\{\phi_k\}_{k=1}^d$, forming the columns of Φ are linearly independent.

We consider the Q-learning approximation scheme in Algorithm 3. The algorithm starts with an initial parameter estimate of θ_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, θ_{t+1}, is computed by taking a semigradient 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 are generated by the exploratory policy that chooses to continue at all time steps.

Algorithm 3 Q-Learning for Optimal Stopping Problems.

Input: initial guess θ_0, step-size sequence $\{\alpha_t\}_{t\in\mathbb{N}}$ and radius R.

Initialize: $\bar{\theta}_0 \leftarrow \theta_0$.

for $t = 0, 1, \ldots$ **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 op` *

Define loss function:

$\frac{1}{2}(y_t - Q_\theta(s_t))^2$; * `sample Bellman op` *

Compute negative semigradient:

$g_t(\theta_t) = -\frac{\partial}{\partial\theta}\frac{1}{2}(y_t - Q_\theta(s_t))^2|_{\theta=\theta_t}$;

Take a semigradient step:

$\theta_{t+1} = \theta_t + \alpha_t g_t(\theta_t)$; * α_t:`step-size` *

Update averaged iterate:

$\bar{\theta}_{t+1} \leftarrow (\frac{t}{t+1})\bar{\theta}_t + (\frac{1}{t+1})\theta_t$; * $\bar{\theta}_{t+1} = \frac{1}{t+1}\sum_{\ell=1}^{t}\theta_\ell$ *

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 in Algorithm 3 converges to the unique solution, θ^*, 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 γ. 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

$$\left\|\Phi\theta^* - Q^*\right\|_D \le \frac{1}{\sqrt{1-\gamma^2}}\left\|\Pi_D Q^* - Q^*\right\|_D,$$

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

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

Then, for an initial state s_0 drawn from the stationary distribution π,

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

where V^* and $V_{\tilde{\mu}}$ denote the value functions corresponding, respectively, to the optimal stopping policy and the approximate stopping policy $\tilde{\mu}$. 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 and 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 previously. To this effect, we highlight that the 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 semigradient. As a general function of θ and tuple $O_t = (s_t, u(s_t), s_t')$, the negative semigradient can be written as

$$g_t(\theta) = \big(u(s_t) + \gamma \max\{U(s_t'), \phi(s_t')^\top\theta\} - \phi(s_t)^\top\theta\big)\phi(s_t). \tag{33}$$

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

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

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 Online 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 θ, TD updates point in the descent direction of $\|\theta^* - \theta\|_2^2$. Lemma 5 bounds the expected norm of the stochastic semigradient, thus giving a control over system noise.

In Lemmas 13 and 14, we state exactly the same results for the Q-function approximation algorithm under the i.i.d. sampling model. With these two key lemmas, convergence bounds shown in Theorem 2 follow by repeating the analysis in Section 7. Recall that Q_{θ^*} denotes the unique fixed point of $\Pi_D F(\cdot)$, that is, $Q_{\theta^*} = \Pi_D F Q_{\theta^*}$.

Lemma 13 (Tsitsiklis and Van Roy 1999)**.** *For any* $\theta \in \mathbb{R}^d$,

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

Proof of Lemma 13. This property is a consequence of the fact that $\Pi_D F(\cdot)$ is a contraction with respect to $\|\cdot\|_D$ with modulus γ. 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 Online Appendix C. □

Lemma 14. *For any fixed* $\theta \in \mathbb{R}^d$, $\mathbb{E}[\|g_t(\theta)\|_2^2] \leq 2\sigma^2 + 8\|Q_\theta - Q_{\theta^*}\|_D^2$, *where* $\sigma^2 = \mathbb{E}[\|g_t(\theta^*)\|_2^2]$.

Proof of Lemma 14. See Online Appendix C for a detailed proof. □

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 semigradient step followed by projection onto Θ_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 state a uniform bound on the semigradient norm.

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

Proof of Lemma 15. See Online Appendix C for a detailed proof. □

If we assume the Markov process $(s_0, s_1, \ldots)$ 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 Online 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 semigradient 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 semigradient bias as shown in Lemma 11. Combining these with standard proof techniques for SGD (Nemirovski et al. 2009, Lacoste-Julien et al. 2012) 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 are generated from a single trajectory of a Markov chain. Our analysis inherits the simplicity and elegance enjoyed by SGD analysis and can gracefully extend to different variants of TD, for example, TD learning with eligibility traces (TD(λ)) 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 that crucially depend on the minimum eigenvalue ω of the feature covariance matrix, which would need to be estimated from samples. Although 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.

Acknowledgments

The authors thank the anonymous referees for feedback and stimulating exchanges and Garud Iyengar for pointing to some references early on in the project.

Endnotes

[1] This was previously attempted by Korda and Prashanth (2015), but critical errors were shown by Lakshminarayanan and Szepesvári (2017).

[2] 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.

[3] This can be formally verified for TD(0) with linear function approximation. If the TD step was 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.

[4] We avoid μ from notation for simplicity.

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

[6] This follows formally as a consequence of Lemma 3 in this paper.

[7] This approach argues for using step sizes of the order of $1/\sqrt{t}$, where t is the current iteration. These are much larger than the step sizes, on the order of $1/t$, that are suggested in the classical stochastic approximation literature. This should not be confused with the approach of using even larger step sizes that do not depend on t or the total number of iterations T (e.g., see Lakshminarayanan and Szepesvári (2018) and related works of Ruppert 1988, Polyak and Juditsky 1992, and Györfi and Walk 1996).

[8] This can be seen from the fact that for any vector u with $\|u\|_2 \le 1$,

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

[9] 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 $A \le -(1-\gamma)\Sigma$, that is, 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 $A^\top A \le \Sigma$. Taking this perspective, the important part of these lemmas is that they allow us to understand TD in terms of feature covariance matrix Σ and the discount factor γ rather than the more mysterious matrix A.

[10] Recall that we assumed $\|\phi(s)\|_2 \le 1$ for all $s \in \mathcal{S}$ and $|\mathcal{R}(s,s')| \le r_{\max}$ for all $s, s' \in \mathcal{S}$.

[11] Let $\mathcal{P}_\mathcal{C}(x) = \operatorname{argmin}_{x'\in\mathcal{C}} \|x' - x\|$ denote the projection operator onto a closed, nonempty, convex set $\mathcal{C} \subset \mathbb{R}^d$. Then $\|\mathcal{P}_\mathcal{C}(x) - \mathcal{P}_\mathcal{C}(y)\| \le \|x - y\|$ for all vectors x and y.

[12] TD(0) corresponds to $\lambda = 0$.

References

Andersen L, Broadie M (2004) Primal-dual simulation algorithm for pricing multidimensional American options. *Management Sci.* 50(9):1222–1234.

Antos A, Szepesvári C, Munos R (2008) Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learn.* 71(1):89–129.

Bach F, Moulines E (2013) Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). *Adv. Neural Inform. Process. Systems* Vol. 26 (Curran Associates, Red Hook, NY), 773–781.

Baird L (1995) Residual algorithms: Reinforcement learning with function approximation. Prieditis A, Russell SJ, eds., *Proceedings of the 12th International Conference on Machine Learning*. (Morgan Kaufmann Publishers, San Francisco, CA), 30–37.

Benveniste A, Métivier M, Priouret P (2012) *Adaptive Algorithms and Stochastic Approximations*, vol. 22 (Springer Science & Business Media, Berlin).

Bertsekas DP (1995) *Dynamic Programming and Optimal Control* (Athena Scientific, Belmont, MA).

Bertsekas DP, Shreve S (1978) *Stochastic Optimal Control: The Discrete-Time Case* (Academic Press, Cambridge, MA).

Borkar VS (2009) *Stochastic Approximation: A Dynamical Systems Viewpoint*, vol. 48 (Springer, Berlin).

Borkar VS, Meyn SP (2000) The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.* 38(2):447–469.

Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. *SIAM Rev.* 60(2):223–311.

Bradtke SJ, Barto AG (1996) Linear least-squares algorithms for temporal difference learning. *Machine Learn.* 22(1–3):33–57.

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

Dalal G, Szörényi B, Thoppe G, Mannor S (2018a) Finite sample analyses for TD(0) with function approximation. McIlraith SA, Weinberger KQ, eds., *32nd AAAI Conf. Artificial Intelligence*, Vol. 34 (AAI Press, Palo Alto, California), 6144-6160.

Dalal G, Szörényi B, Thoppe G, Mannor S (2018b) Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. Bubeck S, Perchet V, Rigollet P, eds., *Proc. 31st Conf. Learn. Theory*, (PMLR, New York, NY), 1199–1233.

Dann C, Neumann G, Peters J (2014) Policy evaluation with temporal differences: A survey and comparison. *J. Machine Learning Res.* 15(24):809–883.

De Farias DP, Van Roy B (2003) The linear programming approach to approximate dynamic programming. *Oper. Res.* 51(6):850–865.

Desai VV, Farias VF, Moallemi CC (2012) Pathwise optimization for optimal stopping problems. *Management Sci.* 58(12):2292–2308.

Devraj AM, Meyn SP (2017) Zap Q-learning. *Adv. Neural Inform. Processing Systems* Vol. 30 (Curran Associates, Red Hook, NY) 2235–2244.

Duchi JC (2018) Introductory lectures on stochastic optimization. *Math. Data* 25:99–185.

Ghavamzadeh M, Lazaric A, Maillard O, Munos R (2010) LSTD with random projections. *Adv. Neural Inform. Processing Systems* 23: 721–729.

Goldberg DA, Chen Y (2018) Beating the curse of dimensionality in options pricing and optimal stopping. Preprint, submitted August 17, 2018, https://arxiv.org/abs/1807.02227.

Györfi L, Walk H (1996) On the averaged stochastic approximation for linear regression. *SIAM J. Control Optim.* 34(1):31–61.

Haugh MB, Kogan L (2004) Pricing American options: A duality approach. *Oper. Res.* 52(2):258–270.

Jaakkola T, Jordan MI, Singh SP (1994) Convergence of stochastic iterative dynamic programming algorithms. Cowan J, Tesauro G, Alspector J, eds., *Adv. Neural Inform. Processing Systems* 7: 703–710.

Jain P, Kar P (2017) Non-convex optimization for machine learning. *Foundations Trends Machine Learning* 10(3–4):142–336.

Konda VR (2002) Actor-critic algorithms. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Korda N, Prashanth LA (2015) On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. Bach F, Blei D, eds., *Proc. 32nd Internat. Conf. Machine Learn.*, (PMLR, New York, NY), 626–634.

Kushner H (2010) Stochastic approximation: A survey. *Wiley Interdisciplinary Rev. Comput. Statist.* 2(1):87–96.

Kushner H, Yin GG (2003) *Stochastic Approximation and Recursive Algorithms and Applications*, vol. 35 (Springer Science & Business Media, Berlin).

Lacoste-Julien S, Schmidt M, Bach F (2012) A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method. Preprint, submitted December 20, 2012, https://arxiv.org/abs/1212.2002.

Lakshminarayanan C, Szepesvári C (2017) Finite time bounds for temporal difference learning with function approximation: Problems with some "state-of-the-art" results. https://sites.ualberta.ca/~szepesva/papers/TD-issues17.pdf.

Lakshminarayanan C, Szepesvári C (2018) Linear stochastic approximation: How far does constant step-size and iterate averaging go? Storkey A, Perez-Cruz F, eds. *Proc. 21st Internat. Conf. Artificial Intelligence Statistics*, (PMLR, New York, NY), 1347–1355.

Lazaric A, Ghavamzadeh M, Munos R (2010) Finite-sample analysis of LSTD. Fürnkranz J, Joachims T, eds., *Proc. 27th Internat. Conf. Machine Learn.*, (Omnipress, Madison, WI), 615–622.

Levin DA, Peres Y (2017) *Markov Chains and Mixing Times*, vol. 107 (American Mathematical Society, Providence, RI).

Liu B, Liu J, Ghavamzadeh M, Mahadevan S, Petrik M (2015) Finite-sample analysis of proximal gradient TD algorithms. Meila M, Heskes T, eds., *Proc. 31st Conf. Uncertainty Artificial Intelligence*, (AUAI Press, Corvallis, OR), 504–513.

Macua SV, Chen J, Zazo S, Sayed AH (2014) Distributed policy evaluation under multiple behavior strategies. *IEEE Trans. Automated Control* 60(5):1260–1274.

Meyn SP, Tweedie RL (2012) *Markov Chains and Stochastic Stability* (Springer Science & Business Media, Berlin).

Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, et al (2015) Human-level control through deep reinforcement learning. *Nature* 518(7540):529–533.

Nemirovski A, Juditsky A, Lan G, Shapiro A (2009) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.* 19(4):1574–1609.

Pires BÁ, Szepesvári C (2012) Statistical linear estimation with penalized estimators: An application to reinforcement learning. Langford J, Pineau J, eds., *Proc. 29th Internat. Conf. Machine Learn.*, (Omnipress, Madison, WI), 1755–1762.

Polyak BT, Juditsky AB (1992) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.* 30(4):838–855.

Prashanth LA, Korda N, Munos R (2014) Fast LSTD using stochastic approximation: Finite time analysis and application to traffic control. Calders T, Esposito F, Hüllermeier E, Meo R, eds., *Joint European Conf. Machine Learn. Knowledge Discovery Databases*, (Springer, Berlin, Heidelberg), 66–81.

Roberts GO, Rosenthal JS (2004) General state space Markov chains and MCMC algorithms. *Probability Survey* 1:20–71.

Ruppert D (1988) *Efficient estimations from a slowly convergent Robbins-Monro process. Technical report*, Cornell University Operations Research and Industrial Engineering, (Ithaca, NY).

Russo D, Zou J (2019) How much does your data exploration overfit? Controlling bias via information usage. *IEEE Trans. Inform. Theory* 66(1):302–323.

Schapire RE, Warmuth MK (1996) On the worst-case analysis of temporal difference learning algorithms. *Machine Learn.* 22(1–3):95–121.

Seijen H, Sutton RS (2014) True online TD (λ). Xing EP, Jebara T, eds., *Proc. 31st Internat. Conf. Machine Learn.*, (PMLR, New York, NY), 692–700.

Sutton RS (1988) Learning to predict by the methods of temporal differences. *Machine Learning* 3(1):9–44.

Sutton RS, Barto AG (1998) *Introduction to Reinforcement Learning* (MIT Press, Cambridge, MA).

Sutton RS, Szepesvári C, Maei HR (2009a) A convergent O(n) temporal difference algorithm for off-policy learning with linear function approximation. Bengio Y, Schuurmans D, Lafferty J, Williams C, Culotta A, eds., *Adv. Neural Inform. Processing Systems* Vol 21. (Curran Associates, Red Hook, NY), 1609–1616.

Sutton RS, Maei HR, Precup D, Bhatnagar S, Silver D, Szepesvári C, Wiewiora E (2009b) Fast gradient descent methods for temporal difference learning with linear function approximation. Bottou L, Littman M, eds., *Proc. 26th Internat. Conf. Machine Learn.*, (Omnipress, Madison, WI), 993–1000.

Touati A, Bacon PL, Precup D, Vincent P (2018) Convergent TREE BACKUP and RETRACE with function approximation. Dy J, Krause A, eds., *Proc. 35th Internat. Conf. Machine Learn.*, (PMLR, New York, NY), 4955–4964.

Tsitsiklis JN, Van Roy B (1997) An analysis of temporal difference learning with function approximation. *IEEE Trans. Automated Control* 42(5):674–690.

Tsitsiklis JN, Van Roy B (1999) Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Trans. Automated Control* 44(10):1840–1851.

Tu S, Recht B (2018) Least-squares temporal difference learning for the linear quadratic regulator. Dy J, Krause A, eds., *Proc. 35th Internat. Conf. Machine Learn.*, (PMLR, New York, NY), 5005–5014.

Van Roy B (1998) Learning and value function approximation in complex decision processes. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

Xu A, Raginsky M (2017) Information-theoretic analysis of generalization capability of learning algorithms. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds., *Adv. Neural Inform. Processing Systems* (Curran Associates, Red Hook, NY), 2524–2533.

Yu H, Bertsekas DP (2009) Convergence results for some temporal difference methods based on least squares. *IEEE Trans. Automated Control* 54(7):1515–1531.

Jalaj Bhandari is a PhD candidate in operations research at Columbia University. His research lies at the intersection of optimization and reinforcement learning. He is broadly interested in designing efficient and robust machine learning algorithms motivated by real-world applications.

Daniel Russo is an assistant professor in the Decision, Risk, and Operations Division of Columbia Business School. His research lies at the intersection of statistical machine learning and sequential decision-making and contributes to the fields of online optimization and reinforcement learning.

Raghav Singal is a PhD candidate in operations research at Columbia University. His primary research interest is in the area of analytics. He likes to build models that help businesses understand complex systems and make better decisions.