A finite-sample analysis of multi-step temporal difference estimates

Yaqi Duan[⋆] YAQID@MIT.EDU

Martin J. Wainwright[◇,†,⋆] WAINWRIGWORK@GMAIL.COM

Department of Electrical Engineering and Computer Sciences[⋆]*, Department of Mathematics*[†]*, MIT*

Department of Electrical Engineering and Computer Sciences[◇]*, Department of Statistics*[◇]*, UC Berkeley*

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

We consider the problem of estimating the value function of an infinite-horizon γ-discounted Markov reward process (MRP). We establish non-asymptotic guarantees for a general family of multi-step temporal difference (TD) estimates, including canonical K-step look-ahead TD for $K = 1, 2, \ldots$ and the TD(λ) family for $\lambda \in [0, 1)$ as special cases. Our bounds capture the dependence of these estimates on both the variance as defined by Bellman fluctuations, and the bias arising from possible model mis-specification. Our results reveal that the variance component shows limited sensitivity to the choice of look-ahead defining the estimator itself, while increasing the look-ahead can reduce the bias term. This highlights the benefit of using a larger look-ahead: it reduces bias but need not increase the variance.

Keywords: reinforcement learning, Markov reward process, policy evaluation, temporal difference learning

1. Introduction

Policy evaluation in reinforcement learning refers to evaluating the performance of a decision policy using existing data. The quality of a given policy can be measured by its value function V^*, corresponding to the expected sum of (discounted) rewards under a trajectory generated by running the given policy. Policy evaluation is central to many applications. For example, in the setting of clinical treatments, the value function might correspond to the expected long-term survival rate of septic patients (e.g., Komorowski et al. (2018)), whereas in inventory management, it measures the profits/losses of a company over time (e.g., Giannoccaro and Pontrandolfo (2002)).

In practice, policy evaluation is rendered challenging by the complexity of the underlying state space, which can be of finite cardinality but prohibitively large, or continuous in nature. In most cases of interest, it is essential to use some type of function approximation to compute what is known as a projected fixed point associated with the Bellman operator. In particular, in this paper, we study projected fixed point approximations within a linear function space.

Our focus is the multi-step temporal difference (TD) methods that are commonly used in practice. Recall that the value function V^* can be characterized as the unique fixed point of the Bellman operator $\mathcal{T}$, and the standard approach is to empirically approximate the projected fixed point associated with this operator. Given observations from trajectories, we can form empirical approximations to multi-step versions of the Bellman operator—of the form $\mathcal{T}^{(\boldsymbol{w})} := \sum_{k=1}^{K} w_k \mathcal{T}^{(k)}$ where the integer $K \geq 1$ is the *look-ahead parameter*, and $\boldsymbol{w} \in \mathbb{R}_+^K$ is a vector of non-negative weights summing to one, and $\mathcal{T}^{(k)}$ is the multi-step Bellman operator that looks ahead k steps. The TD(λ)-

family is a well-known instance of this type of approach. Given the wide range of possible choices of look-ahead K and weight vector $\boldsymbol{w}$, one naturally wonders how to make principled choices, and in particular ones that lead to better estimators. These types of questions, while long recognized as being important in reinforcement learning (e.g., Jaakkola et al. (1993); Baird (1995); Bertsekas and Tsitsiklis (1996); Singh and Dayan (1998); Boyan (1999); Yu and Bertsekas (2009); Mann et al. (2016); Bhandari et al. (2018)), are far from completely resolved. In particular, what would be desirable—and the goal of this paper—is theory that gives a very precise understanding of the trade-offs involved, along with some actionable guidelines for the practitioner.

In this paper, we explore these fundamental issues in the context of γ-discounted Markov reward processes. Our main contributions are to provide a non-asymptotic characterization of the statistical properties of a broad class of multi-step policy evaluation procedures, with a particular emphasis on how Bellman fluctuation (variance) and model mis-specification (bias) affect the estimation error. Our theory reveals some surprising phenomena, and also provides guidance on the choice of look-ahead in multi-step methods.

1.1. Our contributions and paper organization

Our main contribution to provide sharp upper bounds on the error in policy evaluation based on a single observed trajectory. Our result (Theorem 1) applies to broad class of projected fixed point estimators, and gives high-probability upper bounds on the associated estimation error. These upper bounds are specified in terms of a *signal-to-noise ratio*, or SNR for short, one which captures the essential difficulty of value function estimation. We identify two different types of fluctuations, denoted by σ_m and σ_a respectively, that correspond to martingale noise (variance), and error due to the Bellman residual (bias), respectively. The martingale noise exhibits behavior similar to that of independent random variables, whereas the temporal dependence in the underlying Markov chain interacts with the Bellman residual to form σ_a. Our characterization of these interactions has a number of interesting implications. As one example, consider the natural intuition about multi-step TD methods—as written about in past work on the topic (Bertsekas and Tsitsiklis, 1996; Boyan, 1999; Yu and Bertsekas, 2009)—that increasing look-ahead, which is known to reduce the (deterministic) approximation error, will increase the (stochastic) estimation error. The results in this paper reveal many scenarios in which estimation error is not increased by larger choices of look-ahead parameter; other factors dictate the limits of choosing look-ahead.

1.2. Related work

This paper builds upon our earlier work (Duan et al., 2021b), in which we studied the properties of the standard one-step ($K = 1$) least-squares temporal difference (LSTD) estimate in its kernelized form. In contrast, the major challenge addressed here is to provide a precise characterization of a much broader class of multi-step estimates.

There is large body of past work on analyzing LSTD procedures (e.g., Munos and Szepesvári (2008); Farahmand et al. (2016); Liu et al. (2015); Fan et al. (2020); Long et al. (2021)). Of most direct relevance here is a line of past work on-policy evaluation and optimization for trajectory-based models. Antos et al. (2008) studied policy iteration using single trajectory generated under a fixed policy. Under a β-mixing condition, they proved various non-asymptotic bounds on both the estimation of the value function, as well as the sub-optimality of the associated policy. Their analysis, involving VC-crossing dimension to measure the function complexity, and proved consis-

tency of policy evaluation and optimization as the trajectory length increases, but the underlying rates are slow (and hence sub-optimal). Focusing on the special case of linear function approximation, Lazaric et al. (2012) proved non-asymptotic bounds for both standard LSTD and least-squares policy iteration; their bounds involve both the feature dimension, and the smallest eigenvalue of the Gram matrix. Bhandari et al. (2018) provided non-asymptotic bounds for temporal difference learning. When applied to data from a single Markov trajectory, their bounds involve a multiplicative factor of the mixing time relative to the i.i.d. case. In application to $\mathrm{TD}(\lambda)$ algorithms, their analysis does not capture the possible benefits of increased λ in reducing statistical estimation error that we document in this work. It should be noted that bounds in the aforementioned papers (Antos et al., 2008; Lazaric et al., 2012; Bhandari et al., 2018) do not isolate the variance structure of the policy evaluation problem, which is essential to establishing the statistical optimality of the estimates. Some recent work, involving a subset of the current authors, does isolate this variance structure in the linear case. Mou et al. (2021) studied stochastic approximation procedures for solving linear fixed point equations over $\mathbb{R}^d$, given observations from a single trajectory of an underlying Markov chain. Among the consequences of their general theory are instance-dependent guarantees for the MSE of $\mathrm{TD}(\lambda)$ methods.

In this paper, we measure model mis-specification in an instance-dependent (and hence not worst-case) way, as either the L^2-distance between V^* and its projection onto $\mathcal{F}$, or the Bellman residual associated with the projected fixed point. This instance-dependence provides a more refined view than worst-case notions, such "realizability" or "completeness" (e.g., Munos and Szepesvári (2008); Farahmand et al. (2016); Chen and Jiang (2019); Duan and Wang (2020); Uehara et al. (2021); Duan et al. (2021a); Zanette (2021)), along with approximate versions thereof (Munos and Szepesvári, 2008; Chen and Jiang, 2019; Uehara et al., 2021; Duan et al., 2021a), that have been used to specify approximation error in past work on reinforcement learning. However, it should be noted that the global nature of our measure of approximation error makes it more restrictive than pointwise notions that have been used for estimating functionals of value functions (e.g., Zanette and Wainwright (2022)).

1.3. Paper organization and notation

The remainder of this paper is organized as follows. In Section 2, we begin by introducing the background of Markov reward process, value function estimation as well as multi-step Bellman equations. In Section 3, we present the statements of non-asymptotic upper bounds (Sections 3.1 and 3.2) and the interpretations of the terms that set the noise levels (Section 3.3). In Section 4, we show that various structural conditions result in different optimal choices for the look-ahead. In Section 5, we provide illustrative simulations that verify the predictions of the theory.

Notation: Throughout the paper, we use C, c, c_0 etc. to denote universal constants whose numerical values may very from line to line. Given a distribution μ, we define the $L^2(\mu)$-norm $\|f\|_\mu := \sqrt{\int f^2 \mu(dx)}$. We also make use of the supremum norm $\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$.

2. Background and problem set-up

In this section, we provide background and then set up the problem to be studied in this paper. We begin in Section 2.1 with background on Markov reward processes and value functions. Section 2.2

is devoted to the definitions of multi-step Bellman equations and operators. Section 2.3 introduces the empirical temporal difference (TD) estimates that we analyze.

2.1. Markov reward processes

For a given discount factor $\gamma \in [0, 1)$, a γ-discounted Markov reward process consists of a time-homogeneous Markov chain on a state space $\mathcal{X}$, combined with a reward function r that maps each state x to a scalar reward $r(x)$. The Markov chain is defined by a transition function $\mathcal{P}$, so that when the chain is in state x at the current time, it transitions to a random state X' drawn according to a probability distribution $\mathcal{P}(\cdot \mid x)$.

The value function measures the expected value of a geometrically discounted sum of the rewards over a trajectory of the Markov chain. In particular, for each possible starting state $x \in \mathcal{X}$, we define $V^*(x) := \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t \, r(X_t) \;\big|\; X_0 = x\big]$, where the expectation is taken over a trajectory $(x, X_1, X_2, \ldots)$ that is governed by the probability transition operator $\mathcal{P}$. The existence and well-definedness of the value function V^* is guaranteed under mild conditions.

In this paper, we study the problem of estimating the value function V^* based on a set of observations from a single trajectory $\boldsymbol{\tau} = (x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ from the Markov chain, where x_1 is drawn from the stationary distribution. We assume that the reward function r is known, so that the rewards $r(x_i)$ are also given. Our results extend to the case of unknown reward function, but we study the known reward case for the bulk of our analysis so as to draw attention to the differences between multi-step Bellman operators (all of which share the same reward structure).

Letting μ correspond to the stationary distribution of the Markov chain, we measure the error associated with an estimate $\widehat{f}$ of V^* in terms of the squared-$L^2(\mu)$-norm

$$\|\widehat{f} - V^*\|_\mu^2 := \mathbb{E}\big[(\widehat{f}(X) - V^*(X))^2\big].$$

2.2. Multi-step Bellman operators

The estimates studied in this paper are established based on the observation that, for any positive integer $k = 1, 2, \ldots$, the value function V^* is the solution to the k^{th}-order Bellman fixed point equation $V^*(x) = r(x) + \gamma \, \mathbb{E}_{X_1|x}\big[r(X_1)\big] + \ldots + \gamma^{k-1} \, \mathbb{E}_{X_{k-1}|x}\big[r(X_{k-1})\big] + \gamma^k \, \mathbb{E}_{X_k|x}\big[V^*(X_k)\big]$. For natural reasons, we refer to the integer k as the number of *look-ahead steps*.

For future reference, we introduce a more concise formulation of this fixed point relation as $V^* = \mathcal{T}^{(k)}(V^*)$, where the *$k$-step Bellman operator* $\mathcal{T}^{(k)}$ is given by

$$\big(\mathcal{T}^{(k)}(f)\big)(x) := \mathbb{E}\Big[\sum_{\ell=0}^{k-1} \gamma^\ell r(X_\ell) + \gamma^k f(X_k) \mid X_0 = x\Big] \quad \text{for any } f \in L^2(\mu) \text{ and } x \in \mathcal{X}.$$

More generally, we can form convex combinations of operators of this type. As one possible formalization, fix a positive integer $K \geq 1$, and consider the class of all *weighted K-step Bellman operators*

$$\mathcal{T}^{(\boldsymbol{w})} := \textstyle\sum_{k=1}^{K} w_k \mathcal{T}^{(k)}, \tag{1}$$

where the non-negative weight vector $\boldsymbol{w} = \begin{pmatrix} w_1 & \ldots & w_K \end{pmatrix}$ ranges over the probability simplex in $\mathbb{R}^K$. Given these constraints, it can be verified that any such weighted operator $\mathcal{T}^{(\boldsymbol{w})}$ also has the original value function V^* as its unique fixed point. Notice that if we observe a single trajectory of length n, we can (in principle) try to approximate a K-step weighted Bellman operator for any $K \in \{1, 2, \ldots, n-1\}$.

Given any weight vector $\boldsymbol{w} \in \mathbb{R}^K$, we define the *effective discount factor* of the $\boldsymbol{w}$-weighted TD estimate as

$$\overline{\gamma} := \sum_{k=1}^{K} w_k \gamma^k. \tag{2}$$

Note that $\overline{\gamma} \leq \gamma$ for any choice of the weight vector in the probability simplex.

Let us consider a few examples to illustrate. As a first example, in the standard K-step temporal difference method, the weight vector is given by $w_K = 1$, and $w_\ell = 0$ for $\ell \neq K$. This choice leads to the effective discount factor $\overline{\gamma} = \gamma^K$. Given an integer $K \geq 1$, a second example is the K-truncated TD(λ) method, in which the weight vector takes the form $\boldsymbol{w} = \frac{1-\lambda}{1-\lambda^K}\begin{bmatrix} 1 & \lambda & \ldots & \lambda^{K-1} \end{bmatrix}$ for some $\lambda \in [0, 1)$. This choice leads to an effective discount factor $\overline{\gamma} = \frac{\gamma(1-\lambda)}{1-\lambda\gamma}\frac{1-\lambda^K\gamma^K}{1-\lambda^K}$. If we take the limit as $K \to \infty$, then we see that $\overline{\gamma} \to \frac{\gamma(1-\lambda)}{1-\lambda\gamma}$ and $(1-\overline{\gamma})^{-1} \to \frac{1-\lambda\gamma}{1-\gamma}$.

2.3. Multi-step temporal difference estimates

In this paper, we study multi-step temporal difference (TD) estimates that are based on linear function approximation. Any such function space is defined by a feature mapping $\Phi : \mathcal{X} \to \mathbb{R}^d$, where d denotes the (finite) dimension. The linear space $\mathbb{H}$ consists of linear combinations of the features, i.e. $\mathbb{H} := \left\{ f(\cdot) = \Phi(\cdot)^\top \boldsymbol{f} \;\middle|\; \boldsymbol{f} \in \mathbb{R}^d \right\}$. Define an associated function norm $\|f\|_{\mathbb{H}} := \|\boldsymbol{f}\|_2$. Throughout this paper, we assume that the linear function space $\mathbb{H}$ contains all constant functions.

The multi-step TD estimator $\widehat{f}$ is defined as the fixed point of an empirical Bellman operator $\widehat{\mathcal{T}}^{(\boldsymbol{w})}$. For each $k \in \{1, \ldots, K\}$, and time step $t \in \{1, 2, \ldots, n-k\}$, define the k-step future return as $\widehat{G}_{t+1}^{t+k}(f) := \sum_{\ell=1}^{k-1} \gamma^\ell r(x_{t+\ell}) + \gamma^k f(x_{t+k})$. In terms of these future returns, the *empirical Bellman operator* is given by

$$f \mapsto \widehat{\mathcal{T}}^{(\boldsymbol{w})}(f) := r + \arg\min_{h \in \mathbb{H}} \left\{ \frac{1}{n-K} \sum_{t=1}^{n-K} \Big(h(x_t) - \sum_{k=1}^{K} w_k \, \widehat{G}_{t+1}^{t+k}(f) \Big)^2 + \lambda_n \|h\|_{\mathbb{H}}^2 \right\},$$

where $\lambda_n > 0$ is a user-defined regularization parameter. The estimate $\widehat{f}$ is then the solution to the fixed point equation $\widehat{f} = \widehat{\mathcal{T}}^{(\boldsymbol{w})}(\widehat{f})$. We use f^* to denote the population-level estimate, i.e. the projected fixed point of Bellman operator $\mathcal{T}^{(\boldsymbol{w})}$. When there exists any model mis-specification, i.e. $V^* \notin \mathbb{H}$, we may have $f^* \neq V^*$.

The estimator $\widehat{f}$ can also be written as the solution of a linear operator equation defined in terms of covariance and cross-covariance operators associated with the RKHS. It is also closely connected with the standard description of temporal difference learning as a form of stochastic approximation.

3. Non-asymptotic upper bounds on multi-step LSTD

In this section, we develop some non-asymptotic theory for the estimation error associated with the function $\widehat{f}$ computed using multi-step LSTD method. From the introduction, its overall error as an estimate of the true value function V^* is upper bounded as

$$\|\widehat{f} - V^*\|_\mu \leq \underbrace{\|\widehat{f} - f^*\|_\mu}_{\text{Estimation error}} + \underbrace{\|f^* - V^*\|_\mu}_{\text{Approximation error}}. \tag{3}$$

The approximation error $\|f^* - V^*\|_\mu$ is deterministic in nature, and controlled by the richness of the underlying RKHS, as well as the choice of weight vector $\boldsymbol{w}$ in a multi-step TD method. The goal of this section is to characterize the statistical estimation error $\|\widehat{f} - f^*\|_\mu$ associated with estimating

the projected fixed point f^*. In the sequel, Section 3.1 provides the statement of the upper bound. Section 3.2 presents bounds on the noise level that appears in the theorem. In Section 3.3, we provide intuition for the noise terms.

3.1. Statement of upper bound

Our analysis relies on the following mixing condition, which involves a scalar $\tau_* \geq 1$, known as the *mixing time*, and a nonnegative constant $C_{\mathrm{v}} < \infty$.

(MIX(τ_*)) The Markov chain is uniformly ergodic, meaning that

$$\left\|\mathcal{P}^t(\cdot \mid x) - \mu(\cdot)\right\|_{\mathrm{TV}} \leq C_{\mathrm{v}}\,(1-\tau_*^{-1})^t \qquad \text{for any state } x \in \mathcal{X} \text{ and step } t \in \mathbb{N}\,. \tag{4}$$

In addition to this mixing condition, our analysis imposes some boundedness conditions on the feature mapping Φ, as well as the covariance matrix $\Sigma_{\mathrm{cov}} = \int_{\mathcal{X}} \Phi(x)\,\Phi(x)^\top \mu(dx) \in \mathbb{R}^{d\times d}$ that it induces. This covariance matrix has a collection of eigenvalues $\{\mu_j\}_{j=1}^d \subset \mathbb{R}$ along with associated eigenvectors $\{\boldsymbol{v}_j\}_{j=1}^d \subset \mathbb{R}^d$. We impose the following regularity condition:

(BD(b,κ)) The feature mapping $\Phi : \mathcal{X} \to \mathbb{R}^d$ and eigenfunctions are uniformly bounded—viz.

$$\sup_{x\in\mathcal{X}} \|\Phi(x)\|_2 \leq b \qquad \text{and} \qquad \sup_{x\in\mathcal{X}} \sup_{1\leq j\leq d} \left|\langle \boldsymbol{v}_j, \Phi(x)\rangle\right| / \sqrt{\mu_j} \leq \kappa\,. \tag{5}$$

We now turn to the other ingredients that underlie our main result:

Effective timescale: Recalling the definition (2) of the effective discount factor $\overline{\gamma} \equiv \overline{\gamma}(\boldsymbol{w})$, we use $\overline{H} := (1-\overline{\gamma})^{-1}$ to denote the effective timescale associated with a $\boldsymbol{w}$-weighted TD method.

Bellman fluctuations: We measure the variability of the K-step Bellman operator via

$$\sigma_m(f^*) := \sum_{\ell=1}^{K} \gamma^\ell \sqrt{\mathbb{E}\Big[\mathrm{Var}\big[\big(\sum_{k=\ell}^{K} w_k\, \mathcal{T}^{(k-\ell)}(f^*)\big)(X') \;\big|\; X\big]\Big]}\,, \tag{6a}$$

where X is drawn from the stationary distribution μ and (X, X') are successive samples from the Markov chain $\mathcal{P}$.

Bellman residual and mixing: When the value function V^* does not belong to the space $\mathbb{H}$, the projected fixed point f^* differs from V^*, and hence the Bellman residual $\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*$ is non-zero. In this case, our bounds involve an additional noise term, given by

$$\sigma_a(f^*) := 2\sqrt{\tau_*}\left\|\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*\right\|_\mu \left\{1 + \tfrac{1}{4}\log \tfrac{\|\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*\|_\infty}{\|\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*\|_\mu}\right\} \tag{6b}$$

where τ_* is the mixing time.

We are now ready to present our main result. Consider a user-defined radius R such that

$$R \geq \max\left\{\|f^* - r\|_{\mathbb{H}},\ \tfrac{\|r\|_\infty}{b}\right\}, \tag{7}$$

along with the *effective noise level*

$$\zeta_0 := \overline{H}\left\{\sigma_m(f^*) + \sigma_a(f^*)\right\}. \tag{8}$$

Theorem 1 (Non-asymptotic upper bound) *Under the mixing condition* **(MIX(τ_*))** *and the feature boundedness condition* **(BD(b, κ))**, *consider the multi-step LSTD method. Suppose that the sample size n is large enough to ensure that $n \geq c\, \kappa^4 d^2 (\tau_* + K)(1-\overline{\gamma})^{-2}$. Then for any regularization parameter $\lambda_n \geq c_0\,(1-\overline{\gamma})\,\frac{\kappa^2\,\zeta_0^2}{R^2}\,\frac{d}{n}\log n$, the projected fixed point $\widehat{f} \equiv \widehat{f}(\lambda_n)$ satisfies the bound*

$$\|\widehat{f} - f^*\|_\mu^2 \leq c_1\, R^2 \left\{ \frac{\kappa^2\,\zeta_0^2}{R^2}\,\frac{d}{n}\log^2 n + \frac{\lambda_n}{1-\overline{\gamma}} \right\} \tag{9}$$

with probability at least $1 - c_2\,\exp\big(-c^\dagger\,\frac{\kappa^2\,\zeta_0^2}{b^2\,R^2}\,d\big)$, *where* $c^\dagger := c_3\,\frac{(1-\overline{\gamma})^2(1-\gamma)^2}{\tau_*+K}$.

With the minimal choice of regularization parameter $\lambda_n = c_0\,(1-\overline{\gamma})\,\frac{\kappa^2\,\zeta_0^2}{R^2}\,\frac{d}{n}\log n$, Theorem 1 guarantees that $\|\widehat{f} - f^*\|_\mu^2 \lesssim \kappa^2\,\zeta_0^2\,(d/n)\log^2 n$ with high probability.

3.2. Bounding the noise level ζ_0

The bound (9) from Theorem 1 holds, in weakened form, for any upper bound on the noise level ζ_0. Accordingly, in order to develop intuition for the behavior of our bounds, it is useful to derive such an upper bound that decouples into a variance term along with a form of approximation error. In particular, let us define the *expected Bellman variance*

$$\sigma^2(V^*) := \mathbb{E}_{X\sim\mu}\big[\mathrm{Var}[V^*(X') \;\big|\; X]\big], \tag{10}$$

associated with the true value function. Recall that $\tau_* \geq 1$ is the mixing time, $H = (1-\gamma)^{-1}$ stands for the effective horizon, and define the error $V_\perp^* := V^* - \Pi_{\mathbb{H}}(V^*)$ in the projection[1] of V^* onto the function class. With this notation, it can be shown that the effective noise ζ_0 defined in equation (8) is upper bounded as

$$\zeta_0 \leq \widetilde{\zeta}_0 := c' \Big\{ \underbrace{H\,\sigma(V^*)}_{\text{uncertainty}} + \underbrace{\overline{H}\sqrt{\max\{H,\,\tau_*\}}\,\|V_\perp^*\|_\mu}_{\text{model error}} \Big\} \tag{11}$$

where the pre-factor $c' \equiv c'(f^*)$ depends only on the logarithmic quantity $\log \frac{\|\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*\|_\infty}{\|\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*\|_\mu}$.

We notice that the term $H\,\sigma(V^*)$ remains invariant to the choice of the weight vector $\boldsymbol{w}$. Consequently, in the regime of negligible mis-specification, no matter what type of TD method is chosen—with possibilities including K-step TD method for $K \in \mathbb{Z}_+$, or TD(λ) for any $\lambda \in [0,1)$—the estimation error $\|\widehat{f} - f^*\|_\mu^2$ should scale in a similar manner. Thus, the flexibility in the choice of TD method does not have any benefits for reducing estimation error. To be clear, it can still reduce the approximation error in the decomposition (3), since the effective discount factor can be reduced.

It should be emphasized that in other regimes, careful choices of the weight vector $\boldsymbol{w}$ can reduce the estimation error. More precisely, this choice can reduce the effective horizon $\overline{H}$, which in turn can reduce the model error portion of the effective noise bound $\widetilde{\zeta}_0$, as well as the approximation error $\|f^* - V^*\|_\mu^2$. Reductions in $\overline{H}$ can be achieved by choosing a larger look-ahead parameter K in a multi-step TD method, or a larger value of $\lambda \in [0,1)$ in the TD(λ) family of methods.

1. To be clear, the projection $\Pi_{\mathbb{H}}(V^*)$ is, in general, *not* the same as the projected fixed point f^*.

3.3. Intuition for $\sigma_m(f^*)$ and $\sigma_a(f^*)$

Let us provide some intuition for how the standard deviation $\sigma_m(f^*)$ and model mis-specification error $\sigma_a(f^*)$, from equations (6a) and equation (6b) respectively, enter the upper bound.

Introducing the shorthand $\widetilde{n} := n - K$. The proof of Theorem 1 involves bounding an empirical process $\sup_{g \in \mathcal{G}} \left\{ \frac{1}{\widetilde{n}} \sum_{t=1}^{\widetilde{n}} g(x_t)\, \nu_t \right\}$, where $\mathcal{G}$ is a carefully chosen function class and the random variable ν_t is given by

$$\nu_t = \underbrace{\left\{ \sum_{k=1}^{K} w_k \, \widehat{G}_{t+1}^{t+k}(f^*) - \mathbb{E}\Big[\sum_{k=1}^{K} w_k \, \widehat{G}_{t+1}^{t+k}(f^*) \;\Big|\; x_t \Big] \right\}}_{m_t \equiv m(x_{t+1}^{t+K})} + \underbrace{\big(\mathcal{T}^{(\boldsymbol{w})}(f^*) - f^*\big)(x_t)}_{a_t \equiv a(x_t)} . \tag{12}$$

In equation (12), the random variable ν_t is decomposed into the sum of two parts, m_t and a_t. In our analysis, we show that fluctuations associated with first term m_t lead to standard deviation $\sigma_m(f^*)$ in definition (8) of noise level ζ_0, while the term a_t gives rise to the model mis-specification error $\sigma_a(f^*)$.

4. Choices of look-ahead under various uniform structural conditions

In the analysis of RL algorithms, it is standard to impose various types of uniform structural conditions on the MRP. In this section, we explore the consequences of our instance-dependent results for two such structural constraints: (i) a uniform bound on the reward function r; and (ii) a $L^2(\mu)$-norm upper bound on the value function V^*. Our theory shows that different choices of TD parameters should be made in these two settings.

4.1. Uniformly bounded reward

Suppose that the reward function is uniformly bounded—viz. $\|r\|_\infty \leq \varrho_r$ for some finite constant ϱ_r—and that the weight vector $\boldsymbol{w}$ is chosen to ensure that

$$\overline{H} \equiv \overline{H}(\boldsymbol{w}) \lesssim \left\{ 1 + \frac{H}{\tau_*} \right\}. \tag{13a}$$

The bound (13a) can be guaranteed by setting

$$K \gtrsim \min\{H, \tau_*\} \qquad \text{for } K\text{-step TD, or} \qquad (1-\lambda)^{-1} \gtrsim \min\{H, \tau_*\} \quad \text{for TD}(\lambda). \tag{13b}$$

With these choices, the noise level ζ_0 is bounded as

$$\zeta_0 \lesssim H \sqrt{\max\{H, \tau_*\}}\, \varrho_r \,, \tag{13c}$$

which, in turn, implies that

$$\|\widehat{f} - f^*\|_\mu^2 \lesssim \underbrace{\varrho_r^2 \, H^2 \, \max\{H, \tau_*\} \, \frac{d}{n}}_{\epsilon^2} \log^2 n \tag{14}$$

holds with probability at least $1 - c_2 \exp\big(- c^\dagger \frac{n \epsilon^2}{b^2 H^2 \varrho_r^2} \big)$.

4.2. Value function with bounded $L^2(\mu)$-norm

Now suppose that $\|V^*\|_\mu \leq \varrho_V$ for some finite ϱ_V, and the weight vector $\boldsymbol{w}$ is chosen to ensure that

$$\overline{H} \equiv \overline{H}(\boldsymbol{w}) \lesssim \min\Big\{\sqrt{H},\, 1 + \frac{H}{\sqrt{\tau_*}}\Big\}. \tag{15a}$$

The bound (15a) can be satisfied by letting

$$K \gtrsim \min\big\{H, \sqrt{H+\tau_*}\big\} \quad \text{in } K\text{-step TD, or} \quad (1-\lambda)^{-1} \gtrsim \min\big\{H, \sqrt{H+\tau_*}\big\} \quad \text{in TD}(\lambda)\,. \tag{15b}$$

We therefore prove that the effective noise level is bounded as

$$\zeta_0 \lesssim \max\{H, \sqrt{\tau_*}\}\, \varrho_V. \tag{15c}$$

As before, for the LSTD estimate $\widehat{f}$, we have

$$\|\widehat{f} - f^*\|_\mu^2 \;\lesssim\; \underbrace{\varrho_V^2\; \max\{H^2,\, \tau_*\}\; \frac{d}{n}}_{\epsilon^2} \log^2 n \tag{16}$$

with probability at least $1 - c_2\, \exp\big(-c^\dagger \frac{n\,\epsilon^2}{b^2\,\varrho_V^2}\big)$. By comparison with the bound (14) for the bounded reward case, we see that estimation error is increased; this change is to be expected, since we have imposed only the milder condition of a L^2-bounded value function.

5. Some illustrative simulations

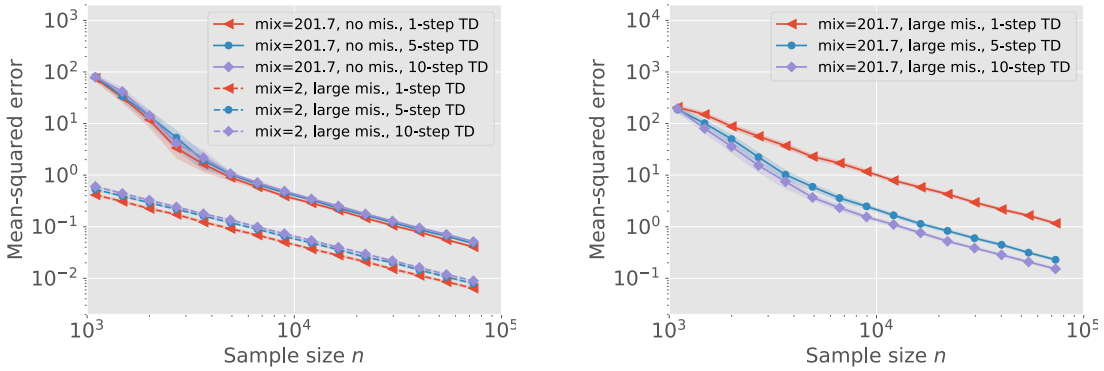
It is helpful to examine some simulations, so as to reveal the phenomena predicted by our theory. Here we show some plots of the mean-squared estimation error $\mathbb{E}\|\widehat{f} - f^*\|_\mu^2$ for estimates $\widehat{f}$ obtained using various types of multi-step LSTD estimates of the value function.

Figure 1 provides comparisons of TD estimates with look-ahead lengths $K \in \{1, 5, 10\}$, as applied to a discounted MRP with $\gamma = 0.9$. We conducted three groups of experiments in total, corresponding to the following types of MRP instances: (i) slowly mixing but well-specified (no mis-specification, i.e. $V^* \in \mathbb{H}$); (ii) large mis-specification but rapidly mixing; or (iii) large mis-specification and slowly mixing. As indicated in the figure, panel (a) involves the first two cases (i) and (ii), whereas panel (b) provides results for case (iii).

From panel (a), we see that, for both cases (i) and (ii), the choice of look-ahead K has little effect; all three methods ($K \in \{1, 5, 10\}$) behave very similarly. This behavior should be contrasted with case (iii): as shown in panel (b), in this setting, increasing the look-ahead K leads to substantial reductions in the MSE. Thus, while some settings are unaffected by look-ahead choice, changing K does have a very significant effect for a model that is both mis-specified and slowly mixing.

6. Discussion

In this paper, we analyzed non-asymptotic statistical properties of multi-step temporal difference (TD) methods. In particular, we investigated how variance (Bellman fluctuation) and bias (model



(a) No mis-specification *or* fast mixing

(b) Large mis-specification *and* slow mixing.

Figure 1. Log-log plots of the mean-square error (MSE) versus the sample size n for different multi-step temporal difference (TD) estimates when using data from a single path. For each point on each curve on each plot, the MSE was approximated by taking a Monte Carlo average over 5000 trials; 3 times sample errors are shown by the shaded area. (a) No mis-specification *or* fast mixing: When there are enough samples, the MSEs of TD estimates with different look-ahead have similar scales. (b) Large mis-specification *and* slowly mixing: The MSE is smaller for larger step K in the TD estimates.

mis-specification) influence the statistical estimation error. Our theory shows when and to what extent multi-step TD methods improve the quality of estimates.

Our work leaves open a number of intriguing questions; let us mention a few of them here to conclude. First, it would be interesting to develop a principled method for parameter selection in $\boldsymbol{w}$-weighted TD that can be implemented without population-level knowledge. Currently, our theory involves some quantities that are non-trivial to estimate using data, for instance, the norm of Bellman residual and the mixing time. Second, the scope of the paper is restricted to the on-policy setting in reinforcement learning. The generalization of the theory to off-policy evaluation remains challenging. It is interesting to determine whether, and if so under what conditions, off-policy procedures can be devised to benefit from multi-step predictive models. Another interesting direction is how to use possible freedom in data collection so as to develop adaptive procedures that minimize the estimation error.

Acknowledgments

This work was partially supported by Office of Naval Research Grant ONR-N00014-21-1-2842, NSF-CCF grant 1955450, and NSF-DMS grant 2015454 to MJW.

References

Radoslaw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electronic Journal of Probability*, 13:1000–1034, 2008.

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

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume II. Athena Scientific, 3rd edition, 2011.

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

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

Justin A Boyan. Least-squares temporal difference learning. In *ICML*, pages 49–56, 1999.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Olga Chilina. f-uniform ergodicity of Markov chains. *Supervised Project, Unversity of Toronto*, 2006.

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

Yaqi Duan, Chi Jin, and Zhiyuan Li. Risk bounds and Rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pages 2892–2902. PMLR, 2021a.

Yaqi Duan, Mengdi Wang, and Martin J Wainwright. Optimal policy evaluation using kernel-based temporal difference methods. *arXiv preprint arXiv:2109.12002*, 2021b.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR, 2020.

Amir-Massoud Farahmand, Mohammad Ghavamzadeh, Csaba Szepesvári, and Shie Mannor. Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research*, 17(1):4809–4874, 2016.

Ilaria Giannoccaro and Pierpaolo Pontrandolfo. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2):153–161, 2002.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient TD algorithms. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 504–513, 2015.

Jihao Long, Jiequn Han, and Weinan E. An L^2 analysis of reinforcement learning in high dimensions with kernel and neural network approximation. *arXiv preprint arXiv:2104.07794*, 2021.

Timothy A Mann, Hugo Penedones, Shie Mannor, and Todd Hester. Adaptive lambda least-squares temporal difference learning. *arXiv preprint arXiv:1612.09465*, 2016.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Stephen J Montgomery-Smith. Comparison of sums of independent identically distributed random variables. *Probability and Mathematical Statistics*, 14 no.2:281–285, 1993.

Wenlong Mou, Ashwin Pananjady, Martin J Wainwright, and Peter L Bartlett. Optimal and instance-dependent guarantees for Markovian linear stochastic approximation. *arXiv preprint arXiv:2112.12770*, 2021.

Wenlong Mou, Ashwin Pananjady, and Martin J. Wainwright. Optimal oracle inequalities for solving projected fixed-point equations. *Mathematics of Operations Research*, 2022. To appear; Posted originally as *arXiv preprint arXiv:2021.05299*, 2021.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

Satinder Singh and Peter Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32(1):5–40, 1998.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. 42(5), 1997.

Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

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

Huizhen Yu and Dimitri P Bertsekas. Error bounds for approximations from projected linear equations. *Mathematics of Operations Research*, 35(2):306–329, 2010.

A. Zanette and M. J. Wainwright. Bellman residual orthogonalization for offline reinforcement learning. In *Neural Information Processing Systems*, December 2022. Long version posted as arxiv:2203.12786.

Andrea Zanette. Exponential lower bounds for batch reinforcement learning: Batch RL can be exponentially harder than online RL. In *International Conference on Machine Learning*, pages 12287–12297. PMLR, 2021.