On the Variance of Temporal Difference Learning and its Reduction Using Control Variates

Hsiao-Ru Pan, Bernhard Schölkopf

Keywords: temporal difference learning, Monte Carlo estimation, advantage function, control variate

Summary

We analyze the variance of temporal difference (TD) learning using the phased setting with tabular representation, and show that one of the mechanisms behind its ability to reduce variance is by effectively aggregating over a larger number of independent trajectories. Based on this insight, we demonstrate that (1) the variance of TD is asymptotically bounded from above by Monte Carlo (MC) estimators, and (2) shorter horizon updates incurs less variance for a fixed number of samples. Beyond TD, we show that Direct Advantage Estimation (DAE), a method for estimating the advantage function, can be seen as a type of regression-adjusted control variate, which achieves a tighter bound on the variance compared to TD in the large-sample limit. Finally, we numerically illustrate the behaviors of these estimators with carefully designed environments.

Contribution(s)

1. We prove that, in the phased (synchronous) and the IID (asynchronous) settings, the asymptotic variance of temporal difference (TD) learning is upper-bounded by that of Monte Carlo methods, and that one mechanism underlying TD's variance reduction is its effective aggregation over a larger pool of trajectories.
   Context: TD learning is widely assumed to reduce variance via bootstrapping (Sutton et al., 1998; Szepesvari, 2010; Dann et al., 2014). While previous works have shown that bootstrapping can be seen as a form of bias-variance tradeoff for multi-step TD (Kearns & Singh, 2000), and that batch TD is more efficient than MC (Grunewalder et al., 2007; Cheikhi & Russo, 2023), the variance comparison is less clear in the online regime. We address this gap by analyzing TD's variance reduction in the phased (synchronous) setting.
2. We show that (1) the advantage function can be used as control variates for *policy evaluation*, and (2) Direct Advantage Estimation (DAE) (Pan et al., 2022) is a form of control variate regression that simultaneously estimates both the value and advantage functions.
   Context: While Pan et al. (2022) demonstrated empirically that DAE improves deep RL performance, its theoretical justification has remained limited to convergence of the estimator. We provide a complementary perspective by showing that DAE also reduces variance by analyzing it through the lens of control variates in policy evaluation, thereby offering a theoretical explanation for its empirical performance gains.

On the Variance of Temporal Difference Learning and its Reduction Using Control Variates

Hsiao-Ru Pan[1], **Bernhard Schölkopf**[1,2,3]
{hpan,bs}@tuebingen.mpg.de

[1]**Max Planck Institute for Intelligent Systems, Tübingen**
[2]**ELLIS Institute Tübingen**
[3]**ETH Zürich**

Abstract

We analyze the variance of temporal difference (TD) learning using the phased setting with tabular representation, and show that one of the mechanisms behind its ability to reduce variance is by effectively aggregating over a larger number of independent trajectories. Based on this insight, we demonstrate that (1) the variance of TD is asymptotically bounded from above by Monte Carlo (MC) estimators, and (2) shorter horizon updates incurs less variance for a fixed number of samples. Beyond TD, we show that Direct Advantage Estimation (DAE), a method for estimating the advantage function, can be seen as a type of regression-adjusted control variate, which achieves a tighter bound on the variance compared to TD in the large-sample limit. Finally, we numerically illustrate the behaviors of these estimators with carefully designed environments.

1 Introduction

Policy evaluation is a central problem in reinforcement learning (RL) (Sutton et al., 1998). Among various methods, temporal difference (TD) learning (Sutton, 1988) stands out as a cornerstone technique for this class of problems. Traditional methods like Monte Carlo (MC) methods estimate the expected return by averaging returns of sample trajectories, typically resulting in unbiased but high variance estimates. In contrast, TD learning updates value estimates iteratively through bootstrapping (i.e., estimate based on previous estimates), thereby avoiding the need for full trajectories and tends to exhibit lower variance. Kearns & Singh (2000) showed that bootstrapping can be seen as a form of bias-variance tradeoff, where high variance estimates from sample trajectories are replaced with low variance biased bootstrapped values. While this intuition largely holds true, it is not difficult to see that the full story is more complicated. For example, if we initialize the value estimates with the *ground truth* values and update them using TD, then we essentially *inject* variance (e.g., stochastic rewards) into the estimates, and it is not immediately clear how this variance affects bootstrapping asymptotically. This shows that the way bootstrapping reduces variance may be more nuanced than simply "replace a high variance trajectory with a biased estimate".

In the present work, we analyze the variance of multi-step TD in the phased (synchronous) setting (Kearns & Singh, 1998), which abstracts away some of the complexities due to stochastic approximations. Beyond TD, we also draw connections to Direct Advantage Estimation (DAE) (Pan et al., 2022), a method that simultaneously estimates the value function and the advantage function, and show that DAE can be seen as a type of control variate regression that enjoys a tighter bound on the variance compared to TD. To summarize, we show that:

- The asymptotic variance of multi-step TD is bounded above by that of MC, and one way TD reduces variance is by effectively averaging over a larger pool of trajectories.

- The advantage function can be seen as control variates for the MC estimator, which reduces its variance, and DAE is a type of regression-adjusted control variate for TD.

Finally, we construct examples to illustrate the asymptotic behaviors of these estimators.

2 Background

We consider a discounted Markov Decision Process (MDP) (Puterman, 2014) $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$ with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, transition probability $p(s'|s,a)$, reward function $r(s,a)$, and discount factor $\gamma \in [0,1)$. For simplicity, we assume the reward function is deterministic unless otherwise stated and denote $r(s_t, a_t)$ by r_t when the context is clear. A policy $\pi(\cdot|s)$ is a function that maps states to distributions over $\mathcal{A}$, and we focus on the case where the policy π is fixed. The value functions are defined by $V^\pi(s) = \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_t | s_0{=}s]$, and $Q^\pi(s,a) = \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_t | s_0{=}s, a_0{=}a]$, respectively ($\mathbb{E}_\pi$ indicates that actions are sampled from π). The advantage function is defined by $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. For the present work, we focus on tabular representations, and use $\mathbf{V}^\pi \in \mathbb{R}^{|\mathcal{S}|}$, $\mathbf{A}^\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to denote the value and the advantage functions, respectively, and use their functional forms for indexing.

Learning value functions is at the core of various policy optimization algorithms, and one classical method is TD learning (Sutton, 1988), which updates the estimates through bootstrapping. Specifically, one-step TD estimates the value function by sampling transition tuples (s_t, a_t, s_{t+1}) and updating the values by $V(s_t) \leftarrow V(s_t) + \alpha(r_t + \gamma V(s_{t+1}) - V(s_t))$, where $\alpha \in \mathbb{R}$ is the learning rate. An extension of this is multi-step TD, which samples multiple timesteps before updating the value via $V(s_t) \leftarrow V(s_t) + \alpha(r_t + \cdots + \gamma^k V(s_{t+k}) - V(s_t))$. We denote this update by TD(k)[1].

Phased TD We presently consider the *phased* setting (Kearns & Singh, 1998), which removes some of the complexities of TD leaning due to asynchronous updates and stochastic approximations. In phased TD, value estimates are updated in *phases*, where each phase consists of two steps: (1) sampling n k-step trajectories for each state: $\mathcal{D} = \{\tau_{s,i}\}_{s\in\mathcal{S}, i\in[1,\cdots,n]}$, where $\tau_{s,i} = (s, a_0^i, r_0^i, s_1^i, \cdots, s_k^i)$, and (2) updating the value of every state synchronously by

$$V^{T+1}(s) \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(r_0^i + \cdots + \gamma^{k-1} r_{k-1}^i + \gamma^k V^T\left(s_k^i\right)\right), \tag{1}$$

where T denotes the phase. It was shown that phased TD is analogous to *TD learning with a fixed learning rate* under mild assumptions. Kearns & Singh (2000) used this setting to analyze the bias-variance tradeoff of multi-step TD, and showed that the estimation error satisfies:

$$\max_s |V^{T+1}(s) - V^\pi(s)| := \Delta_{T+1} \le \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=0}^{k-1}\gamma^t r_t^i\right) - \mathbb{E}\left[\left(\sum_{t=0}^{k-1}\gamma^t r_t\right)|s_0{=}s\right]\right|}_{\text{variance}} + \gamma^k \underbrace{\Delta_T}_{\text{bias}},$$

where the variance term accounts for the stochasticity from the sample rewards, while the bias term accounts for the error from bootstrapping. The authors then derived the following probabilistic bound (assume $\Delta_0 = 1$ and $|r(\cdot)| \le 1$):

$$\Delta_T \le \frac{1-\gamma^{kT}}{1-\gamma}\sqrt{\frac{3\log(k/\delta)}{n}} + \gamma^{kT}, \tag{2}$$

with probability $1-\delta$. Intuitively, increasing k reduces the bias from bootstrapping at the cost of increased variances from the rewards. However, this bias-variance decomposition fails to consider the variance from finite samples of s_k^i, and the variance from previous estimates V^T. Furthermore, as $k \to \infty$ (MC estimation), this bound becomes vacuous even though the rewards are bounded.

[1]This is not to be confused with TD(λ).

Control Variate Control variate (Hammersley & Handscomb, 1964; Asmussen & Glynn, 2007) is a technique for reducing variances of MC methods. We briefly review the basics of control variate in the one-dimensional setting, and refer the reader to Owen (2013) for a more general treatment.

Recall that MC methods estimate $\mu_X = \mathbb{E}[X]$ by $\hat{\mu}_X = \frac{1}{n}\sum_{i=1}^n X_i$, where $X_i \overset{\text{iid}}{\sim} X$. Given another random variable Y, the control variate, that satisfies $\text{Cov}(X, Y) \neq 0$ and has known $\mathbb{E}[Y]$ (assume $\mathbb{E}[Y] = 0$, otherwise use $Z = Y - \mathbb{E}[Y]$). Let $\hat{\mu}_{X,\lambda Y} = \frac{1}{n}\sum_{i=1}^n X_i + \lambda Y_i$, where $(X_i, Y_i) \overset{\text{iid}}{\sim} (X, Y)$ and λ is a tunable constant. This new estimator remains unbiased, but has a different variance $\text{Var}(\hat{\mu}_{X,\lambda Y}) = \text{Var}(\hat{\mu}_X) + \frac{\lambda^2}{n}\text{Var}(Y) + \frac{2\lambda}{n}\text{Cov}(X, Y)$. If we choose $\lambda = \lambda^* = -\frac{\text{Cov}(X,Y)}{\text{Var}(Y)}$, then $\text{Var}(\hat{\mu}_{X,\lambda^* Y}) = \text{Var}(\hat{\mu}_X) - \frac{\text{Cov}(X,Y)^2}{n\text{Var}(Y)} \leq \text{Var}(\hat{\mu}_X)$, which is never worse than the original estimator. In practice, however, λ^* is usually unknown, and estimated by $\hat{\lambda} = -\frac{\widehat{\text{Cov}}(X,Y)}{\widehat{\text{Var}}(Y)}$. Interestingly, this estimator arises naturally from solving the following least squares:

$$(\hat{\mu}_{X,\hat{\lambda}Y}, \hat{\lambda}) = \arg\min_{\theta,\lambda} \sum_{i=1}^n (\theta - x_i - \lambda y_i)^2. \tag{3}$$

This approach is also known as regression-adjusted control variate, and can be readily generalized to more complex settings (e.g., multiple control variates). The error of this estimator is:

$$\hat{\mu}_{X,\hat{\lambda}Y} - \mu_X = \underbrace{(\hat{\lambda} - \lambda^*)\sum_{i=1}^n \frac{Y_i}{n}}_{O(1/n)} + \underbrace{\hat{\mu}_{X,\lambda^* Y} - \mu_X}_{O(1/\sqrt{n})}. \tag{4}$$

As both $\hat{\lambda} - \lambda^*$ and $\sum_{i=1}^n \frac{Y_i}{n}$ approach zero with rates $O(1/\sqrt{n})$, their product converges to 0 with rate $O(1/n)$. Consequently, the second term dominates the error asymptotically, and $\hat{\mu}_{X,\lambda^* Y}$ can be seen as a large-sample approximation of $\hat{\mu}_{X,\hat{\lambda}Y}$. Finally, we note that the estimator is, in general, biased since Y_i can be correlated with $\hat{\lambda}$ (i.e., $\mathbb{E}[Y_i\hat{\lambda}] \neq 0$); however, the estimator remains consistent as the errors approach zero for $n \to \infty$.

Direct Advantage Estimation Direct Advantage Estimation (DAE) (Pan et al., 2022) is a method for estimating the advantage function directly from sample trajectories. Similar to multi-step TD, DAE can update values by bootstrapping previous estimates. Specifically, DAE estimates the values by iteratively minimizing the following constrained least-squares:

$$\mathcal{L}(\hat{A}, \hat{V}) = \mathbb{E}_\pi\left[\left(\sum_{t=0}^{k-1}\gamma^t(r_t - \hat{A}_t) + \gamma^k V_{\text{tar}}(s_k) - \hat{V}(s_0)\right)^2\right] \text{ s.t.} \sum_a \pi(a|s)\hat{A}(s,a) = 0, \tag{5}$$

where $\hat{A}_t = \hat{A}(s_t, a_t)$, and V_{tar} is the bootstrapping target. It was shown that the minimizer of this objective can be viewed as multi-step estimates of the advantage function and the value function. In practice, the expectation is replaced with an average over sample trajectories. We note that, if we force $\hat{A} \equiv 0$ and only optimize with respect to $\hat{V}$, then DAE reduces to multi-step TD. DAE demonstrated strong empirical performance in the deep RL setting; however, it remains unclear whether estimating the value function this way is beneficial compared to classical approaches.

3 Variance of Monte Carlo (MC) and TD Learning

We begin by showing that the variance of MC can be broken down into segments. This will become useful when compared to multi-step TD, whose variance follows a similar form.

For clarity, we summarize the notations used in subsequent analyses in Table 1.

Table 1: Variables names and their definitions. i denotes the ith trajectory. k and n are constants unless otherwise stated. Note that $\mathbf{P}$ is a (row) vector of dimension $|\mathcal{S}|$.

Variable	Def.	Description
$\mathbf{G}$	$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t r_t^i$	Average of sample returns
$\mathcal{D}$	$\left\{(s, a_0^i, r_0^i, s_1^i, \cdots, s_k^i)\right\}_{s\in\mathcal{S}, i=1,\ldots,n}$	k-step partial trajectories
$\mathbf{R}$	$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{k-1}\gamma^t r_t^i$	Average of sample k-step rewards
$\mathbf{P}$	$\mathbf{P}_{s'} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(s_k^i = s')$	Empirical k-step transition distribution

Variance of MC Recall that MC estimates values via $V_{\mathrm{MC}}(s) = \mathbf{G}$. The following lemma shows that its variance is equal to the sum of the variances of partial trajectories.

Lemma 1. $\mathrm{Var}(V_{\mathrm{MC}}(s)) = \mathrm{Var}(\mathbf{G}) = \sum_{m=0}^{\infty}\gamma^{2km}\mathbb{E}\left[\mathrm{Var}(\mathbf{R} + \gamma^k\mathbf{P}\mathbf{V}^{\pi}|s_0{=}s_{km})\middle|\, s_0{=}s\right]$,

This is a multi-step ($k \geq 1$) multi-sample ($n \geq 1$) generalization of the variance recursion derived by Sobel (1982), see Appendix 8.1 for a proof. As we will see, the variance of TD shares a similar structure, except $\mathbf{V}^{\pi}$ is replaced by bootstrapping targets.

Variance of Multi-Step TD We first rewrite the phased update (Equation 1) into:

$$V^T_{\mathrm{TD}(k)}(s) = \mathbf{R} + \gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)}. \tag{6}$$

Without loss of generality, we set $\mathbf{V}^0_{\mathrm{TD}(k)} \equiv 0$. Note that $\mathbf{V}^T_{\mathrm{TD}(k)}$ is now a random vector for $T > 0$ as the update involves random variables $\mathbf{R}$ and $\mathbf{P}$, and its expectation, denoted $\bar{\mathbf{V}}^T_{\mathrm{TD}(k)}$, is equal to:

$$\bar{V}^T_{\mathrm{TD}(k)}(s) = \mathbb{E}\left[\mathbf{R}\right] + \gamma^k\mathbb{E}\left[\mathbf{P}\right]\mathbb{E}\left[\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\right] = \mathbb{E}\left[\sum_{t=0}^{k-1}\gamma^t r_t + \gamma^k\bar{V}^{T-1}_{\mathrm{TD}(k)}(s_k)\middle|\, s_0{=}s\right],$$

and we recover the k-step Bellman update. We emphasize that, while this expectation converges to $\mathbf{V}^{\pi}$ in the limit ($T \to \infty$), $\mathbf{V}^T_{\mathrm{TD}(k)}$ in general does *not*. This is due to the finite-sample variance, which causes the estimates to oscillate around $\mathbf{V}^{\pi}$. We now analyze this variance and show that, similar to Lemma 1, the upper bound of the variance of TD also satisfies a recurrence relation.

Lemma 2.

$$\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \leq \mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)}\middle|\, s_0{=}s\right) + \gamma^{2k}\mathbb{E}_{s_k}\left[\mathrm{Var}\left(V^{T-1}_{\mathrm{TD}(k)}(s_k)\right)\middle|\, s_0{=}s\right]. \tag{7}$$

Proof.

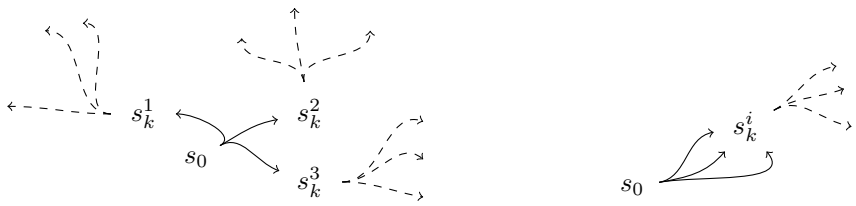
$$\begin{aligned}
&\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) = \mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\middle|\, s_0{=}s\right) && (8)\\
=&\mathrm{Var}\left(\mathbb{E}\left[\mathbf{R} + \gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\middle|\,\mathcal{D}\right]\middle|\, s_0{=}s\right) + \mathbb{E}\left[\mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\middle|\,\mathcal{D}\right)\middle|\, s_0{=}s\right] && (9)\\
=&\mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)}\middle|\, s_0{=}s\right) + \gamma^{2k}\mathbb{E}\left[\mathrm{Var}\left(\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\middle|\,\mathcal{D}\right)\middle|\, s_0{=}s\right] && (10)\\
\leq&\mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)}\middle|\, s_0{=}s\right) + \gamma^{2k}\mathbb{E}\left[\mathbf{P}\mathrm{Var}\left(\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\right)\middle|\, s_0{=}s\right] && (11)\\
=&\mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)}\middle|\, s_0{=}s\right) + \gamma^{2k}\mathbb{E}_{s_k}\left[\mathrm{Var}\left(V^{T-1}_{\mathrm{TD}(k)}(s_k)\right)\middle|\, s_0{=}s\right]. && (12)
\end{aligned}$$

□

If we expand this recursion, then

$$\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \leq \sum_{m=0}^{T-1}\gamma^{2km}\mathbb{E}\left[\mathrm{Var}\left(\mathbf{R} + \gamma^k\mathbf{P}\bar{\mathbf{V}}^{T-1-m}_{\mathrm{TD}(k)}\middle|\, s_0{=}s_{km}\right)\middle|\, s_0{=}s\right], \tag{13}$$

and we arrive at a similar expression to the MC estimator (Lemma 1). In fact, we get, in the limit:



(a) Independent bootstrapping states (b) Correlated bootstrapping states

Figure 1: Illustration of how TD can reduce variance. Solid and dashed arrows represent trajectories collected in the current phase and the previous phase, respectively. (a) s_k^i are independent, and averaging their values effectively aggregates over a larger pool of trajectories. (b) s_k^i are all the same, and averaging their values provides no variance reduction, as the effective number of trajectories remains the same.

Theorem 1. $\lim_{T\to\infty} \mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \le \mathrm{Var}(V_{\mathrm{MC}}(s))$.

See Appendix 8.2 for a proof and a discussion of how this can be extended to the asynchronous IID setting. This shows that, asymptotically, TD is no worse than MC, independent of k; however, it also suggests that TD, in the worst case, can suffer from the same variance as MC.

When does the equality hold? Let us examine the only inequality used in the derivation, namely, $\mathrm{Var}\left(\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)} \middle| \mathcal{D}\right) \le \mathbf{P}\mathrm{Var}\left(\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\right)$. Recall that the left-hand side is simply the variance of the average of the bootstrap values, namely $\mathrm{Var}\left(\frac{1}{n}\sum_i V^{T-1}_{\mathrm{TD}(k)}(s_k^i) \middle| \mathcal{D}\right)$, and this inequality becomes an equality when the value estimates have correlation 1 (e.g., s_k^i are the same for all i). This suggests that one way bootstrapping reduces variance is by effectively aggregating over diverse states with weak dependencies between their value estimates, as illustrated in Figure 1. In Section 5, we also construct toy examples to illustrate this effect.

A remark on sample efficiency It is not immediately clear whether the comparison in Theorem 1 is fair in terms of sample efficiency as the result is asymptotic. Here, we compare two settings with the same number of samples: (1) T phases with k-step updates, and (2) 1 phase with kT-step update. Both settings require $nkT|\mathcal{S}|$ number of samples, and share the same expected value given the same initialization. However, their variances satisfy the following ordering:

Corollary 1. *For $\tau \ge 0$, we have* $\mathrm{Var}(V^{\tau+T}_{\mathrm{TD}(k)}(s)|\mathbf{V}^{\tau}_{\mathrm{TD}(k)}{=}\mathbf{V}) \le \mathrm{Var}(V^{\tau+1}_{\mathrm{TD}(kT)}(s)|\mathbf{V}^{\tau}_{\mathrm{TD}(kT)}{=}\mathbf{V})$.

See Appendix 8.3 for a proof. This shows that short horizon updates incur less variance than a single long horizon update given the same number of interactions and bootstrapping targets.

4 Control Variate and the Advantage Function

We now shift our focus to the second main contribution of the present work, namely, how to use the advantage function to reduce the variance of value estimations.

We begin by considering the π-centered function class $F_\pi = \{f|\mathbb{E}_\pi[f(s,a)|s] = 0\ \forall s\}$, which has the following property (given $f \in F_\pi$): $\mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t (r_t - f(s_t,a_t))\right] = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_t\right]$. In other words, introducing f does not bias the MC estimate, and f can be seen as a control variate. A natural question is, then, what would be the optimal choice of f^* that minimizes the variance, that is, $f^* = \arg\min_{f\in F_\pi} \mathrm{Var}\left(\sum_{t=0}^\infty \gamma^t (r_t - f(s_t,a_t))\right)$. Pan et al. (2022) proved that the advantage function A^π is the unique minimizer under mild assumptions, and used it as a way to estimate A^π.

Now, if A^π is known, then we can construct a new estimator via:

$$V_{\text{MC-A}}(s) = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=0}^{\infty}\gamma^t\left(r_t^i - A^\pi(s_t^i, a_t^i)\right)\right). \tag{14}$$

One may wonder to what extent can this control variate reduce the variance. To answer this, we use the return decomposition by Pan & Schölkopf (2024):

$$\sum_{t=0}^{\infty}\gamma^t r_t = V^\pi(s_0) + \sum_{t=0}^{\infty}\gamma^t(A^\pi(s_t, a_t) + B^\pi(s_t, a_t, s_{t+1})), \tag{15}$$

where $B^\pi(s_t, a_t, r_t, s_{t+1}) = r_t + \gamma V^\pi(s_{t+1}) - \mathbb{E}[r + \gamma V^\pi(s')|s_t, a_t]$ quantifies how much of the return is caused by stochastic transitions[2]. If $B^\pi \equiv 0$ (e.g., the environment is deterministic), then:

$$\sum_{t=0}^{\infty}\gamma^t r_t = V^\pi(s_0) + \sum_{t=0}^{\infty}\gamma^t A^\pi(s_t, a_t). \tag{16}$$

Combine this with Equation 14, we have $\text{Var}(V_{\text{MC-A}}(s)) = 0$, meaning that A^π can fully explain away the variance of MC in this case. For more general cases, B^π is required to account for the variance caused by stochastic transitions, and it remains open whether B^π can be easily estimated in model-free settings. As such, we focus on the advantage function in the present work.

In practice, A^π is rarely known a priori, so we have to estimate both V^π and A^π simultaneously. Next, we show that DAE (Pan et al., 2022) can be seen as a type of regression-adjusted control variate, which achieves this.

4.1 Direct Advantage Estimation and Control Variate Regression

DAE estimates V^π and A^π by solving a constrained least-square problem (Equation 5). For the present work, we focus only on the value estimates and treat $\hat{A}$ as nuisance parameters. This allows us to remove the constraint and reformulate the objective into:

$$\mathcal{L}_T(\hat{A}, \hat{V}) = \mathbb{E}_\pi\left[\left(\sum_{t=0}^{k-1}\gamma^t\left(r_t - \breve{A}_t\right) + \gamma^k V^T_{\text{DAE}(k)}(s_k) - \hat{V}(s)\right)^2\right],$$

where $\breve{A}_t = \hat{A}(s_t, a_t) - \sum_a \pi(a|s_t)\hat{A}(s_t, a)$ (this parametrization ensures that $\sum_a \pi(a|s)\breve{A}(s, a) = 0$). Under this formulation, the minimizer of $\hat{A}$ may no longer be unique, but the minimizer of $\hat{V}$ remains unchanged. Let us now consider DAE in the phased setting, where the objective becomes:

$$\sum_{i=1}^{n}\left(\sum_{t=0}^{k-1}\gamma^t\left(r_t^i - \breve{A}_t^i\right) + \gamma^k V^T_{\text{DAE}(k)}(s_k^i) - \hat{V}(s)\right)^2.$$

Let $(\mathbf{V}^{T+1}_{\text{DAE}(k)}, \mathbf{A}^{T+1})$ be a minimizer of this objective, and $\mathbf{M} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ be the following:

$$\mathbf{M}_{(s,a)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{k-1}\gamma^t\left(\mathbb{I}(s_t^i{=}s, a_t^i{=}a) - \pi(a|s)\mathbb{I}(s_t^i{=}s)\right),$$

which compares the empirical occupancy measure to its expectation over the actions, then

$$V^{T+1}_{\text{DAE}(k)}(s) = \mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}^T_{\text{DAE}(k)} - \mathbf{M}\mathbf{A}^{T+1}. \tag{17}$$

Note that, this update rule differs from the TD update (cf. Equation 6) only by the term $\mathbf{M}\mathbf{A}^{T+1}$. Comparing this to Equation 3, we see that DAE is a case of regression-adjusted control variate,

[2]We use a slightly more general definition to incorporate stochastic rewards

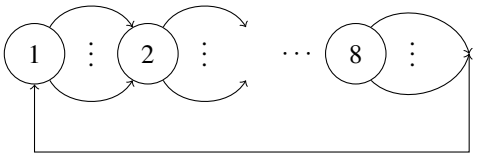


Figure 2: Chain MDP with $\mathcal{S} = \{1, 2, ..., 8\}$, $\mathcal{A} = \{1, ..., |\mathcal{A}|\}$, and $r(s, a) = \frac{(-1)^a}{4}$ (independent of s). Agents return to state 1 after state 8.

Table 2: Parameters of the experiment.

Param.	Description
$\lvert\mathcal{A}\rvert$	action space size
n	number of sample trajectories
k	backup length
p_r	probability of reward masking
p_s	probability of sticky transition

where $\mathbf{M}$ is the control variate with $\mathbb{E}[\mathbf{M}] = 0$, and $\mathbf{A}^T$ is the corresponding coefficients. Additionally, recall that this regression estimator ($\mathbf{V}^T_{\mathrm{DAE}(k)}$) behaves similarly to the one with optimal control variate coefficients ($\mathbf{V}^T_{\mathrm{DAE}^*(k)}$) under the large-sample approximation (cf. Section 2). Here, we focus on $\mathbf{V}^T_{\mathrm{DAE}^*}$ and leave it for future work to analyze the small-sample bias due to control variates.[3] The following result shows the asymptotic behavior of DAE:

Theorem 2. $\limsup_{T\to\infty} \mathrm{Var}(V^T_{\mathrm{DAE}^*(k)}(s)) \leq \mathrm{Var}(V_{\text{MC-A}}(s))$

See Appendix 8.4 for a proof. Compared to Theorem 1, this shows that DAE enjoys a tighter upper bound on the variance by using control variates (recall that $\mathrm{Var}(V_{\text{MC-A}}(s)) \leq \mathrm{Var}(V_{\mathrm{MC}}(s))$). Finally, we note that since TD is a special case of Equation 17 with $\hat{A} \equiv 0$, it follows that $\mathrm{Var}(V^{T+1}_{\mathrm{DAE}^*(k)}(s)|\mathbf{V}^T_{\mathrm{DAE}^*(k)}{=}\mathbf{V}) \leq \mathrm{Var}(V^{T+1}_{\mathrm{TD}(k)}(s)|\mathbf{V}^T_{\mathrm{TD}(k)}{=}\mathbf{V})$.

5 Empirical Illustration

In this section, we illustrate the behaviors of different estimators through experiments based on variants of the chain environment shown in Figure 2. Despite its simplicity, the environment is sufficiently expressive to illustrate various properties of the estimators analyzed in the present study, such as how the variances depend on the bootstrapping states.

All experiments are based on the phased setting, where values are updated synchronously at the end of each phase. We fix the number of phases at 2500, which we found sufficient convergence, the policy π to be uniform, the discount factor at $\gamma = 0.99$, and consider two types of stochasticity:

1. **Reward masking**: Rewards are masked out with probability p_r (i.e., $r = 0$ with probability p_r).
2. **Sticky transition**: With probability p_s, the agent stays in the current state instead of advancing to the next state (i.e., $s_{t+1} = s_t$ with probability p_s).

For simplicity, we consider cases where $|\mathcal{A}|$ is even, such that $V^\pi \equiv 0$. The variances of the MC and the MC-A estimators are equal to

$$\mathrm{Var}(V_{\mathrm{MC}}(s)) = \frac{(1-p_r)}{16(1-\gamma^2)n}, \text{ and } \mathrm{Var}(V_{\text{MC-A}}(s)) = p_r \mathrm{Var}(V_{\mathrm{MC}}(s)), \tag{18}$$

respectively, and will be used as baselines for comparing the estimators. We note that the variances do not depend on p_s due to the symmetric nature of the states.

Each run (configuration) is repeated for 1000 different random seeds to ensure statistical significance. We compare the mean squared error between the true value function V^π and the estimated value ($\mathrm{MSE}(\hat{V}, V^\pi)$, MSE_π for short) , which reflects how quickly the estimator converges to V^π. We also measure the error between the estimated value and the expected Bellman update at each phase ($\mathrm{MSE}(\hat{V}, V_{\mathrm{BE}})$, $\mathrm{MSE}_{\mathrm{BE}}$ for short). For TD, $\mathrm{MSE}_{\mathrm{BE}}$ corresponds to the variance of the update, since the expected TD update follows the Bellman update (i.e., $\mathbb{E}[\mathbf{V}^T_{\mathrm{TD}}] = \mathbf{V}^T_{\mathrm{BE}}$). For DAE, it additionally captures the small-sample bias introduced by the regression estimator.

[3] The large-sample approximation is common in the literature (Lin, 2013; Owen, 2013; Davidson & MacKinnon, 1992), since the small-sample bias from imperfect control variate coefficients is often much smaller than the variance.

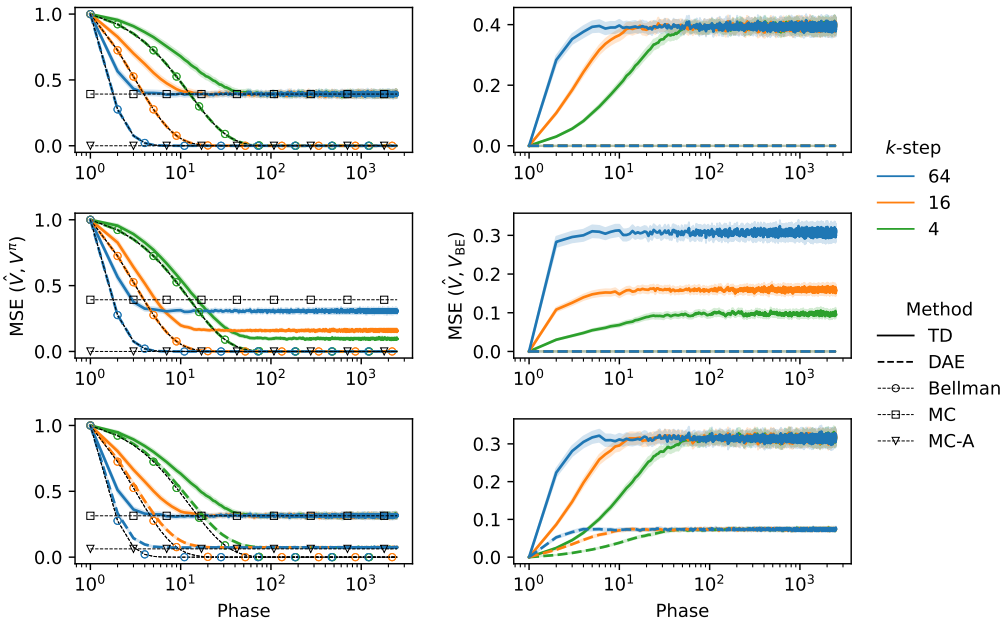


Figure 3: The **Deterministic** case (top), the **Sticky Transition** case (middle), and the **Reward Masking** case (bottom). Lines and shadings represent (mean $\pm$ 3 standard error).

The Deterministic Case ($|\mathcal{A}| = 2$, $n = 8$, $k \in \{4, 16, 64\}$, $p_s = 0$, $p_r = 0$) We first examine the effect of k in the simplest setting with a deterministic environment. Since the transitions are fully deterministic and independent of actions, the bootstrapping states are identical for a given starting state. Our analysis (Section 3) suggests that bootstrapping loses its ability to reduce variance when the value estimates of the bootstrapping states are fully correlated, and TD would behave like MC asymptotically in this case. Indeed, Figure 3 (top) shows that TD learning, independent of k, approaches the same MSE_π as MC. Note that this cannot be explained by the bound given by Equation 2, which predicts that the asymptotic error would grow as k increases. From $\mathrm{MSE}_{\mathrm{BE}}$, we also see the effect of variance injection, where the variances increase with respect to the number of updates. On the other hand, we see DAE achieving almost zero $\mathrm{MSE}_{\mathrm{BE}}$, demonstrating the effectiveness of the variance reduction.

The Sticky Transition Case ($|\mathcal{A}| = 2$, $n = 8$, $k \in \{4, 16, 64\}$, $p_s = 0.25$, $p_r = 0$) We now examine how stochastic transitions affect the estimators. Figure 3 (middle) shows that, counterintuitively, stochasticity *reduces* the variance of TD. Furthermore, the learning curves now follow the common bias-variance tradeoff intuition of multi-step learning (i.e., larger k learns faster but leads to higher variance and vice versa). We emphasize that this setting effectively differs from the deterministic case only in how the bootstrapping states are sampled. As such, the variance reduction can only be explained by TD learning's ability to effectively aggregate value estimates over more diverse states (cf. Figure 1). DAE converges with rates similar to Bellman iterations again, since $B^\pi \equiv 0$ does not depend on p_s.

So far, both cases have almost full coverage of the state-action space from the data in each phase, and have variances that can be fully explained by the advantage function (cf. Equation 16). This allowed DAE to converge almost as fast as Bellman iterations. To see when this breaks down, we next consider the stochastic reward ($B^\pi \not\equiv 0$) and the large $|\mathcal{A}|$ settings.

The Reward Masking Case ($|\mathcal{A}| = 2$, $n = 8$, $k \in \{4, 16, 64\}$, $p_s = 0$, $p_r = 0.2$) Similar to the deterministic case, Figure 3 (bottom) shows that, asymptotically, TD performs similar to MC irrespective of k, as the bootstrapping states are deterministic. On the other hand, the variance of

MC-A is no longer zero since the variance of the rewards cannot be explained by the actions. This causes the MSE_π of DAE to converge slightly above Bellman iterations asymptotically.

The Low Coverage Case ($|\mathcal{A}| \in \{4, 16, 64\}, n = 8, k = 16, p_s = 0, p_r = 0$) Finally, we consider the low coverage case, where the least squares objective become underdetermined for DAE. In the context of RL, this means that most of the actions are not sampled, and the advantage estimates become unreliable. We follow the common practice of choosing the minimum Euclidean-norm solution. Figure 4 shows how the small-sample bias affects learning in early phases (phase < 10) and its relationship to the size of the action space. Interestingly, we find that even when $|\mathcal{A}| \geq 16$, where DAE converges to suboptimal solutions compared to the full coverage cases, DAE still converges to a lower MSE_π than TD, suggesting that the advantage estimates may help reduce variance even when they are poorly estimated. [4]

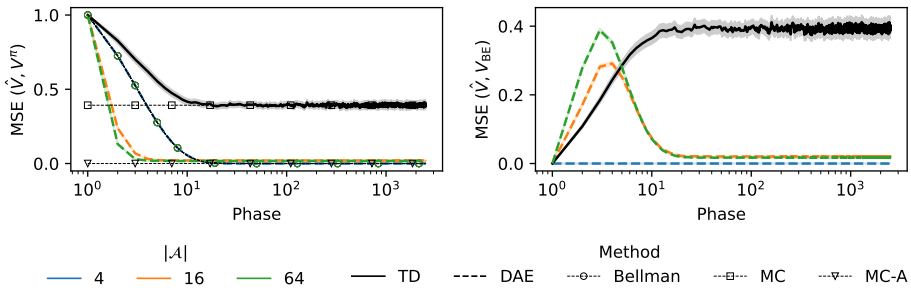


Figure 4: The low coverage case. Lines and shadings represent (mean $\pm$ 3 standard error). Note that TD (black, solid) does not depend on $|\mathcal{A}|$.

6 Related Work

The bias-variance tradeoff between MC and TD has been discussed in classical texts and surveys (Sutton et al., 1998; Szepesvari, 2010; Dann et al., 2014); however, a rigorous analysis remained desirable. Kearns & Singh (2000) analyzed the error bounds of multi-step TD learning in the phased setting (Kearns & Singh, 1998), which also serves as the basis of the present work. In the batch setting, Grunewalder et al. (2007) showed that LSTD (Bradtke & Barto, 1996) is statistically more efficient than MC when the Markovian structure of the environment can provide additional information. More recently, Cheikhi & Russo (2023) derived a more precise statistical relationship between batch TD and MC by analyzing trajectory pooling. However, the batch setting abstracts away the iterative nature of TD, and their implications in the online setting remain unclear. Similar problems have also been studied in function approximation settings (Tsitsiklis & Van Roy, 1996; Dalal et al., 2018; Bhandari et al., 2018), where convergence and finite-sample (time) error bounds were explored.

The advantage function (Baird, 1995) is commonly used as control variates for policy gradient methods (Sutton et al., 1999; Greensmith et al., 2004). The present work demonstrates that the advantage function can also be used as control variates for policy evaluation, and shows that DAE (Pan et al., 2022) can be seen as regression-adjusted control variates.

7 Discussion

We analyzed the behaviors of MC, TD, and DAE, and revealed one mechanism behind the variance reduction property of bootstrapping, namely, the ability to aggregate over a larger number of independent trajectories. Furthermore, we established a connection between DAE and control variate regression, demonstrating how it can further reduce the variance of TD learning. At its core, DAE

[4] We find DAE converging to the true value function again using iterative solvers (see Appendix 9.1).

exploits our knowledge of the policy to reduce variances and an interesting future direction would be to explore other types of control variates for policy evaluation.

Finally, we note some limitations: (1) In the present work, we considered only the phased and the IID settings, and it remains unclear to what extent the results generalize to more realistic sampling schemes (e.g., Markovian sampling). (2) We focused mainly on the variance reduction property of DAE, but it should be noted that DAE also incurs additional space (store $\mathbf{M}$) and time (solve least-squares) complexities. (3) The theoretical results for DAE only hold up to the large-sample approximation case where the control variate coefficients are optimal, and while empirical results seem to suggest that the variance reduction is beneficial even in the small-sample regime, a more rigorous analysis remains desirable.

Acknowledgments

Hsiao-Ru Pan thanks Claire Vernade and Onno Eberhard for the fruitful discussions. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hsiao-Ru Pan.

References

E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999. ISBN 0-89871-447-8.

Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

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

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

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Steven J Bradtke and Andrew G Barto. Linear least-squares algorithms for temporal difference learning. *Machine learning*, 22(1):33–57, 1996.

Robert M Burton and Uwe Rösler. An l2 convergence theorem for random affine mappings. *Journal of applied probability*, 32(1):183–192, 1995.

David Cheikhi and Daniel Russo. On the statistical benefits of temporal difference learning. In *International Conference on Machine Learning*, pp. 4269–4293. PMLR, 2023.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Christoph Dann, Gerhard Neumann, and Jan Peters. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research*, 15(1):809–883, 2014.

Russell Davidson and James G MacKinnon. Regression-based methods for using control variates in monte carlo experiments. *Journal of Econometrics*, 54(1-3):203–222, 1992.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9), 2004.

Steffen Grunewalder, Sepp Hochreiter, and Klaus Obermayer. Optimality of lstd and its relation to mc. In *2007 International Joint Conference on Neural Networks*, pp. 338–343. IEEE, 2007.

J. M. Hammersley and D. C. Handscomb. *Monte Carlo Methods*. Springer Netherlands, 1964. ISBN 9789400958197. DOI: 10.1007/978-94-009-5819-7. URL http://dx.doi.org/10.1007/978-94-009-5819-7.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. DOI: 10.1038/s41586-020-2649-2. URL https://doi.org/10.1038/s41586-020-2649-2.

Michael Kearns and Satinder Singh. Finite-sample convergence rates for q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

Michael J Kearns and Satinder Singh. Bias-variance error bounds for temporal difference updates. In *COLT*, pp. 142–147, 2000.

Winston Lin. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, pp. 295–318, 2013.

Art B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

Christopher C Paige and Michael A Saunders. Lsqr: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

Hsiao-Ru Pan and Bernhard Schölkopf. Skill or luck? return decomposition via advantage functions. *arXiv preprint arXiv:2402.12874*, 2024.

Hsiao-Ru Pan, Nico Gürtler, Alexander Neitz, and Bernhard Schölkopf. Direct advantage estimation. *Advances in Neural Information Processing Systems*, 35:11869–11880, 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Matthew J Sobel. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

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

Richard S Sutton, Andrew G Barto, et al. Introduction to reinforcement learning. 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Csaba Szepesvari. *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2010.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. DOI: 10.1038/s41592-019-0686-2.

Supplementary Materials

The following content was not necessarily subject to peer review.

8 Proofs

8.1 Proof of Lemma 1

Lemma 1

$$\mathrm{Var}(V_{\mathrm{MC}}(s)) = \mathrm{Var}(\mathbf{G}) = \sum_{m=0}^{\infty} \gamma^{2km} \mathbb{E}\left[\mathrm{Var}(\mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}^{\pi} | s_0{=}s_{km}) \middle| s_0{=}s\right].$$

Proof. Note that, when conditioned on $\mathcal{D}$, both $\mathbf{R}$ and $\mathbf{P}$ become constants. Consequently, we have:

$$\begin{aligned}
&\mathrm{Var}(V_{\mathrm{MC}}(s)) = \mathrm{Var}\left(\mathbf{G} \middle| s_0{=}s\right) \\
&= \mathrm{Var}\left(\mathbb{E}\left[\mathbf{G} \middle| \mathcal{D}, s_0{=}s\right] \middle| s_0{=}s\right) + \mathbb{E}\left[\mathrm{Var}\left(\mathbf{G} \middle| \mathcal{D}, s_0{=}s\right) \middle| s_0{=}s\right] \\
&= \mathrm{Var}\left(\mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}^{\pi} \middle| s_0{=}s\right) + \mathbb{E}\left[\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{t=k}^{\infty}\gamma^t r_t^i \middle| \mathcal{D}\right) \middle| s_0{=}s\right] \\
&= \mathrm{Var}\left(\mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}^{\pi} \middle| s_0{=}s\right) + \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\mathrm{Var}\left(\gamma^k V_{\mathrm{MC}}(s_k^i)\right) \middle| s_0{=}s\right] \\
&= \mathrm{Var}\left(\mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}^{\pi} \middle| s_0{=}s\right) + \gamma^{2k}\mathbb{E}\left[\mathrm{Var}\left(V_{\mathrm{MC}}(s_k)\right) | s_0{=}s\right].
\end{aligned}$$

Since $\mathrm{Var}(V_{\mathrm{MC}}(\cdot))$ is bounded, Lemma 1 follows from expanding this recursion. □

8.2 Proof of Theorem 1

Theorem 1

$$\lim_{T\to\infty} \mathrm{Var}(V_{\mathrm{TD}(k)}^T(s)) \leq \mathrm{Var}(V_{\mathrm{MC}}(s)).$$

Proof. First, we show that $\lim_{T\to\infty} \mathrm{Var}(V_{\mathrm{TD}(k)}^T(s))$ converges. Note that

$$V_{\mathrm{TD}(k)}^T(s) = \mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}_{\mathrm{TD}(k)}^{T-1},$$

is a special case of random affine iterated system of the form:

$$X_t = M_t X_{t-1} + N_t, \tag{19}$$

where (M_t, N_t) are IID random variables. Furthermore, since $\gamma^k\mathbf{P}$ is a contraction respect to $||\cdot||_\infty$ and $\mathbf{R}$ is bounded, we know that the random vector $\mathbf{V}_{\mathrm{TD}(k)}^T$ converges in distribution with respect to the Wasserstein distance W_∞ (Burton & Rösler, 1995). Consequently, all finite moments of $\mathbf{V}_{\mathrm{TD}(k)}^T$ also converges as $T \to \infty$.

By Lemma 2, we have:

$$\mathrm{Var}(V_{\mathrm{TD}(k)}^T(s)) \leq \sum_{m=0}^{T-1} \gamma^{2km} \mathbb{E}\left[\mathrm{Var}\left(\mathbf{R} + \gamma^k \mathbf{P}\bar{\mathbf{V}}_{\mathrm{TD}(k)}^{T-1-m} \middle| s_0{=}s_{km}\right) \middle| s_0{=}s\right].$$

It is enough to show that the summation on the right hand side converges to $\mathrm{Var}(V_{\mathrm{MC}}(s))$ as $T \to \infty$. Let $x_{T-m,m} = \mathbb{E}\left[\mathrm{Var}\left(\mathbf{R} + \gamma^k \mathbf{P}\bar{\mathbf{V}}_{\mathrm{TD}(k)}^{T-1-m} \middle| s_0{=}s_{km}\right) \middle| s_0{=}s\right]$, we are interested in the following limit

$$\lim_{T\to\infty} \sum_{m=0}^{T-1} \gamma^{2km} x_{T-m,m}.$$

Since $x_{\infty,m} \coloneqq \lim_{T\to\infty} x_{T-m,m} = \mathbb{E}\left[\mathrm{Var}\left(\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^\pi \middle| s_0{=}s_{km}\right)\middle| s_0{=}s\right]$, there exists $N \in \mathbb{N}$ such that if $T-m > N$ then $|x_{T-m,m} - x_{\infty,m}| < \epsilon$. In addition, since $\mathbf{R}$, $\mathbf{P}$ and $\mathbf{V}^\pi$ are all bounded, there exists $M \in \mathbb{R}$ such that $|x_{\infty,m}| < M$ and $|x_{T-m,m}| < M$ for all $m, T-m \in \mathbb{Z}_+$. Consequently,

$$
\begin{aligned}
&\left|\sum_{m=0}^{T-1}\gamma^{2km}x_{T-m,m} - \sum_{m=0}^{\infty}\gamma^{2km}x_{\infty,m}\right| \\
\leq&\left|\sum_{m=0}^{T-1}\gamma^{2km}(x_{T-m,m}-x_{\infty,m})\right| + \left|\sum_{m=T}^{\infty}\gamma^{2km}x_{\infty,m}\right| \\
\leq&\left|\sum_{m=0}^{T-n-1}\gamma^{2km}(x_{T-m,m}-x_{\infty,m})\right| + \left|\sum_{m=T-n}^{T-1}\gamma^{2km}(x_{T-m,m}-x_{\infty,m})\right| + \frac{M\gamma^{2kT}}{1-\gamma^{2k}} \\
\leq&\frac{\epsilon}{1-\gamma^{2k}} + \frac{2M\gamma^{2k(T-n)}}{1-\gamma^{2k}} + \frac{M\gamma^{2kT}}{1-\gamma^{2k}}
\end{aligned}
$$

is arbitrarily small as $T \to \infty$, and

$$
\lim_{T\to\infty}\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \leq \lim_{T\to\infty}\sum_{m=0}^{T-1}\gamma^{2km}x_{T-m,m} = \sum_{m=0}^{\infty}\gamma^{2km}x_{\infty,m} = \mathrm{Var}(V_{\mathrm{MC}}(s)).
$$

□

Extension to IID Sampling In the IID setting, we sample states independently from a fixed distribution μ (e.g., the occupancy measure) and update their values accordingly. Assuming that μ has full support over the state space $\mathcal{S}$, the above analysis can be extended to the asynchronous setting with minor modifications. Specifically, at each iteration T, we sample a starting state $s \sim \mu$ and update its value using the k-step TD learning rule. Accordingly, $\mathcal{D}$ now denotes a collection of k-step trajectories originating from s, rather than samples covering the entire state space $\mathcal{S}$. We are now ready to proof the following theorem:

Theorem 3. *Given distribution μ over $\mathcal{S}$. We define the IID phased TD iteration as a two-step process: (1) sample a state $s \sim \mu$, and (2) update its value estimate by*

$$
V^T_{\mathrm{TD}(k)}(\tilde{s}) = \begin{cases} \mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)} & \tilde{s}=s \\ V^{T-1}_{\mathrm{TD}}(\tilde{s}) & \textit{otherwise} \end{cases}.
$$

If μ covers $\mathcal{S}$, then $\lim_{T\to\infty}\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \leq \mathrm{Var}(V_{\mathrm{MC}}(s))$.

Proof. Under this sampling scheme, the expected value vector evolves according to a slightly different equation $\bar{V}^T_{\mathrm{TD}(k)}(s) = \mu(s)(\mathbb{E}[\mathbf{R}] + \gamma^k\mathbb{E}[\mathbf{P}]\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)}) + (1-\mu(s))\bar{V}^{T-1}_{\mathrm{TD}(k)}(s)$. If μ covers $\mathcal{S}$ (i.e., $\mu(s) > 0$ for all $s \in \mathcal{S}$), then $\bar{V}^T_{\mathrm{TD}(k)}(s)$ converges to $\mathbf{V}^\pi$. Note that Lemma 2 does not depend on the sampling distribution μ, and continues to hold for the sampled state. However, we now need to account for the variance from whether a state gets updated or not. More precisely, we have

$$
\begin{aligned}
\mathrm{Var}(V^{T+1}_{\mathrm{TD}(k)}(s)) =& \mu(s)\mathrm{Var}(\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^T_{\mathrm{TD}(k)}) + (1-\mu(s))\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \\
&+ \mu(s)(1-\mu(s))\left(\mathbb{E}[\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^T_{\mathrm{TD}(k)}] - \mathbb{E}[V^T_{\mathrm{TD}(k)}(s)]\right)^2.
\end{aligned}
$$

We now denote

$$
\Delta^T = \left(\mathbb{E}[\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^T_{\mathrm{TD}(k)}] - \mathbb{E}[V^T_{\mathrm{TD}(k)}(s)]\right)^2.
$$

Since the expectation of $\mathbf{V}^T_{\mathrm{TD}(k)}$ converges to $\mathbf{V}^\pi$, we have $\Delta^T \to 0$. Expanding this recursion gives

$$\mathrm{Var}(V^{T+1}_{\mathrm{TD}(k)}(s)) = \mu(s)\left(\sum_{t=0}^{T}(1-\mu(s))^t \mathrm{Var}(\mathbf{R}+\gamma^k \mathbf{P}\mathbf{V}^{T-t}_{\mathrm{TD}(k)}) + \sum_{t=0}^{T}(1-\mu(s))^{t+1}\Delta^{T-t}\right).$$

In the limit $T \to \infty$, the first sum converges to $\lim_{T\to\infty}\mathrm{Var}(\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^T_{\mathrm{TD}(k)})$, and the second sum converges to $\lim_{T\to\infty}\Delta^{T-t}=0$. The results then follow from Lemma 2. □

Remark: The key requirement for this result is that trajectory sampling be independent of the value estimates. This requirement may be violated in the Markovian sampling setting, where the value estimates can be correlated with the next-state distribution (i.e., $\mathbf{P}$ and $\mathbf{V}^{T-1}$ are dependent). Such dependence invalidates the conditional variance bound $\mathrm{Var}\left(\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\middle|\mathcal{D}\right) \le \mathbf{P}\mathrm{Var}\left(\mathbf{V}^{T-1}_{\mathrm{TD}(k)}\right)$, which is crucial for the analysis.

8.3 Proof of Corollary 1

Corollary 1 If $\mathbf{V}^\tau_{\mathrm{TD}(k)}=\mathbf{V}^\tau_{\mathrm{TD}(kT)}=\mathbf{V}$, then

$$\mathrm{Var}(V^{\tau+T}_{\mathrm{TD}(k)}(s)|\mathbf{V}^\tau_{\mathrm{TD}(k)}=\mathbf{V}) \le \mathrm{Var}(V^{\tau+1}_{\mathrm{TD}(kT)}(s)|\mathbf{V}^\tau_{\mathrm{TD}(kT)}=\mathbf{V}).$$

Proof. Without loss of generality, we let $\tau=0$, since $\mathbf{V}^0$ is arbitrary. We proof by induction on T. For $T=1$, equality holds. For $T>1$, assume the inequality holds for $T-1$. By Lemma 2, we have:

$$\mathrm{Var}(V^T_{\mathrm{TD}(k)}(s)) \le \mathrm{Var}\left(\mathbf{R}+\gamma^k\mathbf{P}\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)}\middle|s_0{=}s\right) + \gamma^{2k}\mathbb{E}_{s_k}\left[\mathrm{Var}\left(V^{T-1}_{\mathrm{TD}(k)}(s_k)\right)\middle|s_0{=}s\right].$$

Similar to Lemma 1, we can break down the variance of long trajectories into segments, which leads to:

$$\begin{aligned}\mathrm{Var}(V^1_{\mathrm{TD}(kT)}(s)) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=0}^{kT-1}\gamma^t r^i_t + \gamma^{kT}V^0_{\mathrm{TD}(kT)}(s^i_{kT})\right)\middle|s_0{=}s\right)\\ &= \mathrm{Var}\left(\mathbf{R}+\gamma^k\mathbf{P}\bar{\mathbf{V}}^1_{\mathrm{TD}(k(T-1))}\middle|s_0{=}s\right) + \gamma^{2k}\mathbb{E}_{s_k}\left[\mathrm{Var}(V^1_{\mathrm{TD}(k(T-1))}(s_k))|s_0{=}s\right]\end{aligned}$$

Recall that the expected value estimates follow the Bellman update, so $\bar{\mathbf{V}}^{T-1}_{\mathrm{TD}(k)} = \bar{\mathbf{V}}^1_{\mathrm{TD}(k(T-1))}$. Since the inequality holds for $T-1$, we have:

$$\mathbb{E}_{s_k}\left[\mathrm{Var}\left(V^{T-1}_{\mathrm{TD}(k)}(s_k)\right)\middle|s_0{=}s\right] \le \mathbb{E}_{s_k}\left[\mathrm{Var}\left(V^1_{\mathrm{TD}(k(T-1))}(s_k)\right)\middle|s_0{=}s\right],$$

which concludes the proof. □

8.4 Proof of Theorem 2

Theorem 2

$$\limsup_{T\to\infty}\mathrm{Var}(V^T_{\mathrm{DAE}^*(k)}(s)) \le \mathrm{Var}(V_{\text{MC-A}}(s)).$$

Proof. Note that $V^T_{\mathrm{DAE}^*(k)}(s)$ is DAE with the optimal control variate coefficients A^{*T}. Consequently, we must have:

$$\mathrm{Var}(V^T_{\mathrm{DAE}^*(k)}(s)) = \mathrm{Var}\left(\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{DAE}(k)} - \mathbf{M}\mathbf{A}^{*T}\middle|s_0{=}s\right) \tag{20}$$

$$\le \mathrm{Var}\left(\mathbf{R}+\gamma^k\mathbf{P}\mathbf{V}^{T-1}_{\mathrm{DAE}(k)} - \mathbf{M}\mathbf{A}^{\pi}\middle|s_0{=}s\right), \tag{21}$$

where $\mathbf{A}^\pi$ is the true advantage function. Since $\mathbf{A}^\pi$ is a constant vector, and $\mathbf{M}$ is uniformly bounded, we can rewrite the right hand side of this inequality by

$$\text{Var}\left(\mathbf{R}' + \gamma^k \mathbf{P}\mathbf{V}^{T-1}_{\text{DAE}(k)} \middle| s_0 = s\right), \tag{22}$$

where $\mathbf{R}' = \mathbf{R} - \mathbf{M}\mathbf{A}^\pi$. By Corollay 1, we know that this variance is bounded above by the variance of the MC estimator with this new reward function, which is precisely the variance of $V_{\text{MC-A}}$. □

Finally, we make a remark about this corollary. Since the advantage estimate $\mathbf{A}^T$ now also depends on the bootstrapping value function, it is not clear whether the update remains a contraction, or whether higher moments (e.g., variance) also converge. As such, we only prove the supremum limit is upper bounded by $V_{\text{MC-A}}$.

9 Experimental Details

Algorithm 1 shows the pseudocode. All experiments are based on Python with least-square solvers implemented by `NumPy` (Harris et al., 2020), `SciPy` (Virtanen et al., 2020) or `JAX` (Bradbury et al., 2018). A single run (1 seed, 2500 phases) takes less than a minute on commercial CPUs, except for the large $|\mathcal{A}|$ experiment, where we leveraged GPUs (Nvidia A100) to parallelize the least-square solver. We use LSQR (Paige & Saunders, 1982) as the default least-square solver as we found it to be slightly faster. The only exception is the large $|\mathcal{A}|$ experiment, where we used the SVD-based minimum norm solver (Anderson et al., 1999) to ensure reproducibility.

Algorithm 1 Phased TD/DAE

Require: n, k, `alg`$\in$\{TD, DAE\}, `LSTSQ_SOLVER`

Initialize $\mathbf{V} \equiv 0$
for $T = 1, 2, \ldots$ **do**
  $\mathcal{D} = \{\}$
  for $s \in \mathcal{S}$ **do**
    for $i = 1, \ldots, n$ **do**
      Sample k-step trajectory τ from environment
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{\tau\}$
    end for
  end for
  if `alg` == TD **then**
    Compute $\mathbf{R}$, $\mathbf{P}$ from $\mathcal{D}$
    $\mathbf{V} \leftarrow \mathbf{R} + \gamma^k \mathbf{P}\mathbf{V}$
  else
    Compute $\mathbf{M}$, $\mathbf{R}$, $\mathbf{P}$ from $\mathcal{D}$
    $\mathbf{V}, \mathbf{A} \leftarrow$ `LSTSQ_SOLVER`$(||\mathbf{R} + \gamma^k \mathbf{P}\mathbf{V} - \hat{\mathbf{V}} - \mathbf{M}\hat{\mathbf{A}}||^2)$
  end if
end for

9.1 Additional Experiments

The Low Coverage Case With an Iterative Solver ($|\mathcal{A}| \in \{4, 16, 64\}$, $n = 8$, $k = 16$, $p_s = 0$, $p_r = 0$) In Section 5, we showed that increasing the size of the action space results in DAE converging to suboptimal solutions when regularized with minimum norm solutions. In Figure 5, we rerun the same experiment but with an iterative least-sqaures solver (LSQR (Paige & Saunders, 1982) in this case), where the optimum in the previous phase is used as the initialization for the current phase. We find that DAE converges again to the true value function, although at a slower rate as $|\mathcal{A}|$ increases. This might partially explain the success of DAE in the deep RL setting (Pan et al., 2022), where gradient-based optimization is used.

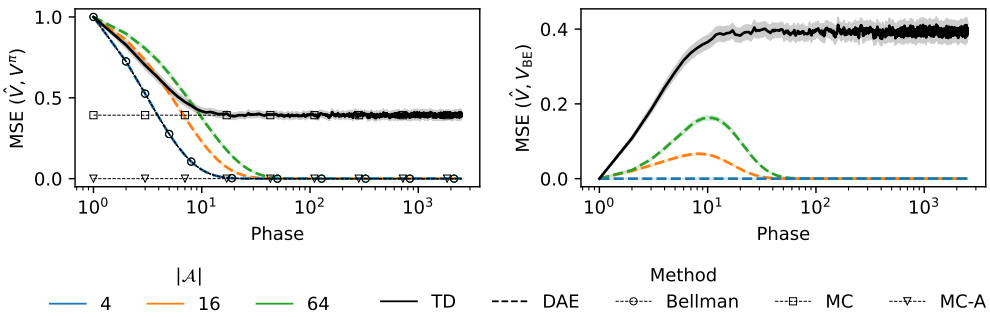


Figure 5: The low coverage case with an iterative solver. Lines and shadings represent (mean $\pm$ 3 standard error). Note that TD does not depend on $|\mathcal{A}|$.