AAAI Press Anonymous Submission Instructions for Authors Using LaTeX

Anonymous submission

Abstract

The existing research on value-based reinforcement learning also minimizes the error. However, is error minimization really the only option for value-based reinforcement learning? We can easily observe that the policy on action choosing probabilities is often related to the relative values, and has nothing to do with their absolute values. Based on this observation, we propose the objective of variance minimization instead of error minimization, derive many new variance minimization algorithms, both including a traditional parameter ω, and conduct an analysis of the convergence rate and experiments. The experimental results show that our proposed variance minimization algorithms converge much faster.

Introduction

Reinforcement learning can be mainly divided into two categories: value-based reinforcement learning and policy gradient-based reinforcement learning. This paper focuses on temporal difference learning based on linear approximated valued functions. Its research is usually divided into two steps: the first step is to establish the convergence of the algorithm, and the second step is to accelerate the algorithm.

In terms of stability, Sutton (1988) established the convergence of on-policy TD(0), and Tsitsiklis and Van Roy (1997) established the convergence of on-policy TD(λ). However, "The deadly triad" consisting of off-policy learning, bootstrapping, and function approximation makes the stability a difficult problem (Sutton and Barto 2018). To solve this problem, convergent off-policy temporal difference learning algorithms are proposed, e.g., BR (Baird et al. 1995), GTD (Sutton, Maei, and Szepesvári 2008), GTD2 and TDC (Sutton et al. 2009), ETD (Sutton, Mahmood, and White 2016), and MRetrace (Chen et al. 2023).

In terms of acceleration, Hackman (2012) proposed Hybrid TD algorithm with on-policy matrix. Liu et al. (2015, 2016, 2018) proposed true stochastic algorithms, i.e., GTD-MP and GTD2-MP, from a convex-concave saddle-point formulation. Second-order methods are used to accelerate TD learning, e.g., Quasi Newton TD (Givchi and Palhang 2015) and accelerated TD (ATD) (Pan, White, and White 2017). Hallak et al. (2016) introduced an new parameter to reduce variance for ETD. Zhang and Whiteson (2022) proposed truncated ETD with a lower variance. Variance Reduced TD with direct variance reduction technique (Johnson and Zhang 2013) is proposed by (Korda and La 2015) and analysed by (Xu et al. 2019). How to further improve the convergence rates of reinforcement learning algorithms is currently still an open problem.

Algorithm stability is prominently reflected in the changes to the objective function, transitioning from mean squared errors (MSE) (Sutton and Barto 2018) to mean squared bellman errors (MSBE) (Baird et al. 1995), then to norm of the expected TD update (Sutton et al. 2009), and further to mean squared projected Bellman errors (MSPBE) (Sutton et al. 2009). On the other hand, algorithm acceleration is more centered around optimizing the iterative update formula of the algorithm itself without altering the objective function, thereby speeding up the convergence rate of the algorithm. The emergence of new optimization objective functions often leads to the development of novel algorithms. The introduction of new algorithms, in turn, tends to inspire researchers to explore methods for accelerating algorithms, leading to the iterative creation of increasingly superior algorithms.

The kernel loss function can be optimized using standard gradient-based methods, addressing the issue of double sampling in residual gradient algorithm (Feng, Li, and Liu 2019). It ensures convergence in both on-policy and off-policy scenarios. The logistic bellman error is convex and smooth in the action-value function parameters, with bounded gradients (Bas-Serrano et al. 2021). In contrast, the squared Bellman error is not convex in the action-value function parameters, and RL algorithms based on recursive optimization using it are known to be unstable.

It is necessary to propose a new objective function, but the mentioned objective functions above are all some form of error. Is minimizing error the only option for value-based reinforcement learning?

For policy evaluation experiments, differences in objective functions may result in inconsistent fixed points. This inconsistency makes it difficult to uniformly compare the superiority of algorithms derived from different objective functions. However, for control experiments, since the choice of actions depends on the relative values of the Q values rather than their absolute values, the presence of solution bias is acceptable.

Based on this observation, we propose alternate objective functions instead of minimizing errors. We minimize

Variance of Bellman Error (VBE) and Variance of Projected Bellman Error (VPBE) and derive Variance Minimization (VM) algorithms. These algorithms preserve the invariance of the optimal policy in the control environments, but significantly reduce the variance of gradient estimation, and thus hastening convergence.

The contributions of this paper are as follows: (1) Introduction of novel objective functions based on the invariance of the optimal policy. (2) Derived three variance minimization algorithms, including on-policy and off-policy. (3) Proof of their convergence. (4) Analysis of the convergence rate of on-policy algorithm. (5) Experiments demonstrating the faster convergence speed of the proposed algorithms.

Background

Reinforcement learning agent interacts with environment, observes state, takes sequential decision makings to influence environment, and obtains rewards. Consider an infinite-horizon discounted Markov Decision Process (MDP), defined by a tuple $\langle S, A, R, P, \gamma \rangle$, where $S = \{1, 2, \ldots, N\}$ is a finite set of states of the environment; A is a finite set of actions of the agent; $R : S \times A \times S \to \mathbb{R}$ is a bounded deterministic reward function; $P : S \times A \times S \to [0, 1]$ is the transition probability distribution; and $\gamma \in (0, 1)$ is the discount factor (Sutton and Barto 2018). Due to the requirements of online learning, value iteration based on sampling is considered in this paper. In each sampling, an experience (or transition) $\langle s, a, s', r \rangle$ is obtained.

A policy is a mapping $\pi : S \times A \to [0, 1]$. The goal of the agent is to find an optimal policy π^* to maximize the expectation of a discounted cumulative rewards in a long period. State value function $V^\pi(s)$ for a stationary policy π is defined as:

$$V^\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_k | s_0 = s].$$

Linear value function for state $s \in S$ is defined as:

$$V_\theta(s) := \boldsymbol{\theta}^\top \boldsymbol{\phi}(s) = \sum_{i=1}^{m} \theta_i \phi_i(s), \tag{1}$$

where $\boldsymbol{\theta} := (\theta_1, \theta_2, \ldots, \theta_m)^\top \in \mathbb{R}^m$ is a parameter vector, $\boldsymbol{\phi} := (\phi_1, \phi_2, \ldots, \phi_m)^\top \in \mathbb{R}^m$ is a feature function defined on state space S, and m is the feature size.

Tabular temporal difference (TD) learning (Sutton and Barto 2018) has been successfully applied to small-scale problems. To deal with the well-known curse of dimensionality of large scale MDPs, value function is usually approximated by a linear model, kernel methods, decision trees, or neural networks, etc. This paper focuses on the linear model, where features are usually hand coded by domain experts.

Variance Minimization Algorithms

Motivation

As shown in Table 1, although there is a bias between the true value and the predicted value, action a_3 is still chosen under the greedy-policy. On the contrary, supervised learning is usually used to predict temperature, humidity, morbidity, etc. If the bias is too large, the consequences could be serious.

Table 1: Classification accuracies for naive Bayes and flexible Bayes on various data sets.

ACTION	Q VALUE	Q VALUE WITH BIAS
$Q(s, a_0)$	1	5
$Q(s, a_1)$	2	6
$Q(s, a_2)$	3	7
$Q(s, a_3)$	4	8
$\arg\min_a Q(s, a)$	a_3	a_3

In addition, reward shaping can significantly speed up the learning by adding a shaping reward $F(s, s')$ to the original reward r, where $F(s, s')$ is the general form of any state-based shaping reward. Static potential-based reward shaping (Static PBRS) maintains the policy invariance if the shaping reward follows from $F(s, s') = \gamma f(s') - f(s)$ (Ng, Harada, and Russell 1999).

This means that we can make changes to the TD error $\delta = r + \gamma \boldsymbol{\theta}^\top \boldsymbol{\phi}' - \boldsymbol{\theta}^\top \boldsymbol{\phi}$ while still ensuring the invariance of the optimal policy,

$$\delta - \omega = r + \gamma \boldsymbol{\theta}^\top \boldsymbol{\phi}' - \boldsymbol{\theta}^\top \boldsymbol{\phi} - \omega,$$

where ω is a constant, acting as a static PBRS. This also means that algorithms with the optimization goal of minimizing errors, after introducing reward shaping, may result in larger or smaller bias. Fortunately, as discussed above, bias is acceptable in reinforcement learning. However, the problem is that selecting an appropriate ω requires expert knowledge. This forces us to learn ω dynamically, i.e., $\omega = \omega_t$ and dynamic PBRS can also maintain the policy invariance if the shaping reward is $F(s, t, s', t') = \gamma f(s', t') - f(s, t)$, where t is the time-step the agent reaches in state s (Devlin and Kudenko 2012). However, this result requires the convergence guarantee of the dynamic potential function $f(s, t)$. If $f(s, t)$ does not converge as the time-step $t \to \infty$, the Q-values of dynamic PBRS are not guaranteed to converge.

Let $f_{\omega_t}(s) = \frac{\omega_t}{\gamma - 1}$. Thus, $F_{\omega_t}(s, s') = \gamma f_{\omega_t}(s') - f_{\omega_t}(s) = \omega_t$ is a dynamic PBRS. And if ω converges finally, the dynamic potential function $f(s, t)$ will converge. Bias is the expected difference between the predicted value and the true value. Therefore, under the premise of bootstrapping, we first think of letting $\omega \doteq \mathbb{E}[\mathbb{E}[\delta|s]] = \mathbb{E}[\delta]$.

Variance Minimization TD Learning: VMTD

For on-policy learning, a novel objective function, Variance of Bellman Error (VBE), is proposed as follows:

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}} \mathrm{VBE}(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\theta}} \mathbb{E}[(\mathbb{E}[\delta|s] - \mathbb{E}[\mathbb{E}[\delta|s]])^2] \\ &= \arg\min_{\boldsymbol{\theta}, \omega} \mathbb{E}[(\mathbb{E}[\delta|s] - \omega)^2]. \end{aligned} \tag{2}$$

Clearly, it is no longer to minimize Bellman errors.

Algorithm 1: VMTD algorithm with linear function approximation in the on-policy setting

Input: $\boldsymbol{\theta}_0$, ω_0, γ, learning rate α_t and β_t
repeat
  For any episode, initialize $\boldsymbol{\theta}_0$ arbitrarily, ω_0 to 0, $\gamma \in (0, 1]$, and α_t and β_t are constant.
  for $t = 0$ **to** $T - 1$ **do**
    Take A_t from S_t according to policy π, and arrive at S_{t+1}
    Observe sample (S_t,R_{t+1},S_{t+1}) at time step t (with their corresponding state feature vectors)
    $\delta_t = R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t' - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t$
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t(\delta_t - \omega_t)\boldsymbol{\phi}_t$
    $\omega_{t+1} \leftarrow \omega_t + \beta_t(\delta_t - \omega_t)$
    $S_t = S_{t+1}$
  end for
until terminal episode

Algorithm 2: VMTDC algorithm with linear function approximation in the off-policy setting

Input: $\boldsymbol{\theta}_0$, $\boldsymbol{u}_0$, ω_0, γ, learning rate α_t, ζ_t and β_t, behavior policy μ and target policy π
repeat
  For any episode, initialize $\boldsymbol{\theta}_0$ arbitrarily, $\boldsymbol{u}_0$ and ω_0 to 0, $\gamma \in (0, 1]$, and α_t, ζ_t and β_t are constant.
  for $t = 0$ **to** $T - 1$ **do**
    Take A_t from S_t according to μ, and arrive at S_{t+1}
    Observe sample (S_t,R_{t+1},S_{t+1}) at time step t (with their corresponding state feature vectors)
    $\delta_t = R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t$
    $\rho_t \leftarrow \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
    $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t\rho_t[(\delta_t - \omega_t)\boldsymbol{\phi}_t - \gamma\boldsymbol{\phi}_{t+1}(\boldsymbol{\phi}_t^\top \boldsymbol{u}_t)]$
    $\boldsymbol{u}_{t+1} \leftarrow \boldsymbol{u}_t + \zeta_t[\rho_t(\delta_t - \omega_t) - \boldsymbol{\phi}_t^\top \boldsymbol{u}_t]\boldsymbol{\phi}_t$
    $\omega_{t+1} \leftarrow \omega_t + \beta_t\rho_t(\delta_t - \omega_t)$
    $S_t = S_{t+1}$
  end for
until terminal episode

First, the parameter ω is derived directly based on stochastic gradient descent:

$$\omega_{k+1} \leftarrow \omega_k + \beta_k(\delta_k - \omega_k), \tag{3}$$

where δ_k is the TD error as follows:

$$\delta_k = r_{k+1} + \gamma \boldsymbol{\theta}_k^\top \boldsymbol{\phi}_{k+1} - \boldsymbol{\theta}_k^\top \boldsymbol{\phi}_k. \tag{4}$$

Then, based on stochastic semi-gradient descent, the update of the parameter $\boldsymbol{\theta}$ is as follows:

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha_k(\boldsymbol{\theta}_k - \omega_k)\boldsymbol{\phi}_k. \tag{5}$$

For control tasks, two extensions of VMTD are named VMSarsa and VMQ respectively, and the update formulas are shown below:

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha_k(\delta_k - \omega_k)\boldsymbol{\phi}(s_k, a_k). \tag{6}$$

and

$$\omega_{k+1} \leftarrow \omega_k + \beta_k(\delta_k - \omega_k), \tag{7}$$

where δ_k delta in VMSarsa is:

$$\delta_k = r_{k+1} + \gamma \boldsymbol{\theta}_k^\top \boldsymbol{\phi}(s_{k+1}, a_{k+1}) - \boldsymbol{\theta}_k^\top \boldsymbol{\phi}(s_k, a_k), \tag{8}$$

and δ_k delta in VMQ is:

$$\delta_k = r_{k+1} + \gamma \max_{a \in A} \boldsymbol{\theta}_k^\top \boldsymbol{\phi}(s_{k+1}, a) - \boldsymbol{\theta}_k^\top \boldsymbol{\phi}(s_k, a_k). \tag{9}$$

Variance Minimization TDC Learning: VMTDC

For off-policy learning, we employ a projection operator. The objective function is called Variance of Projected Bellman error (VPBE), and the corresponding algorithm is called VMTDC.

$$\begin{aligned} \text{VPBE}(\boldsymbol{\theta}) &= \mathbb{E}[(\delta - \mathbb{E}[\delta])\boldsymbol{\phi}]^\top \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]^{-1} \mathbb{E}[(\delta - \mathbb{E}[\delta])\boldsymbol{\phi}] \\ &= \mathbb{E}[(\delta - \omega)\boldsymbol{\phi}]^\top \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]^{-1} \mathbb{E}[(\delta - \omega)\boldsymbol{\phi}], \end{aligned} \tag{10}$$

where ω is used to estimate $\mathbb{E}[\delta]$, i.e., $\omega \doteq \mathbb{E}[\delta]$.

The derivation process of the VMTDC algorithm is the same as that of the TDC algorithm, the only difference is that the original δ is replaced by $\delta - \omega$. Therefore, we can easily get the updated formula of VMTDC, as follows:

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha_k[(\delta_k - \omega_k)\boldsymbol{\phi}(s_k) - \gamma\boldsymbol{\phi}(s_{k+1})(\boldsymbol{\phi}^\top(s_k)\boldsymbol{u}_k)], \tag{11}$$

$$\boldsymbol{u}_{k+1} \leftarrow \boldsymbol{u}_k + \zeta_k[\delta_k - \omega_k - \boldsymbol{\phi}^\top(s_k)\boldsymbol{u}_k]\boldsymbol{\phi}(s_k), \tag{12}$$

and

$$\omega_{k+1} \leftarrow \omega_k + \beta_k(\delta_k - \omega_k), \tag{13}$$

Now, we will introduce the improved version of the GQ(0) algorithm, named VMGQ(0):

$$\begin{aligned} \boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k \quad &+ \quad \alpha_k[(\delta_k - \omega_k)\boldsymbol{\phi}(s_k, a_k) \\ &- \quad \gamma\boldsymbol{\phi}(s_{k+1}, A_{k+1}^*)(\boldsymbol{\phi}^\top(s_k, a_k)\boldsymbol{u}_k)], \end{aligned} \tag{14}$$

$$\boldsymbol{u}_{k+1} \leftarrow \boldsymbol{u}_k + \zeta_k[(\delta_k - \boldsymbol{u}_k) - \boldsymbol{\phi}^\top(s_k, a_k)\boldsymbol{u}_k]\boldsymbol{\phi}(s_k, a_k), \tag{15}$$

and

$$\omega_{k+1} \leftarrow \omega_k + \beta_k(\delta_k - \omega_k), \tag{16}$$

where δ_k is (9) and $A_{k+1}^* = \arg\max_a(\boldsymbol{\theta}_k^\top \boldsymbol{\phi}(s_{k+1}, a))$.

Variance Minimization ETD Learning: VMETD

Based on the off-policy TD algorithm, a scalar, F, is introduced to obtain the ETD algorithm, which ensures convergence under off-policy conditions. This paper further introduces a scalar, ω, based on the ETD algorithm to obtain VMETD. VMETD by the following update:

$$\rho_k \leftarrow \frac{\pi(A_k|S_k)}{\mu(A_k|S_k)} \tag{17}$$

$$F_k \leftarrow \gamma\rho_{k-1}F_{k-1} + 1, \tag{18}$$

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha_k(F_k\rho_k\delta_k - \omega_k)\boldsymbol{\phi}_k, \tag{19}$$

$$\omega_{k+1} \leftarrow \omega_k + \beta_k(F_k\rho_k\delta_k - \omega_k), \tag{20}$$

where μ is behavior policy and π is target policy, F_t is a scalar variable, $F_0 = 1$, ω is used to estimate $\mathbb{E}[\delta]$, i.e.,

Algorithm 3: VMETD algorithm with linear function approximation in the off-policy setting

Input: $\boldsymbol{\theta}_0$, F_0, ω_0, γ, learning rate α_t, ζ_t and β_t, behavior policy μ and target policy π
repeat
For any episode, initialize $\boldsymbol{\theta}_0$ arbitrarily, F_0 to 1 and ω_0 to 0, $\gamma \in (0, 1]$, and α_t, ζ_t and β_t are constant.
for $t = 0$ **to** $T - 1$ **do**
Take A_t from S_t according to μ, and arrive at S_{t+1}
Observe sample (S_t,R_{t+1},S_{t+1}) at time step t (with their corresponding state feature vectors)
$\delta_t = R_{t+1} + \gamma \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t$
$\rho_t \leftarrow \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
$F_t \leftarrow \gamma \rho_t F_{t-1} + 1$
$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t + \alpha_t(F_t \rho_t \delta_t - \omega_t)\boldsymbol{\phi}_t$
$\omega_{t+1} \leftarrow \omega_t + \beta_t(F_t \rho_t \delta_t - \omega_t)$
$S_t = S_{t+1}$
end for
until terminal episode

$\omega \doteq \mathbb{E}[\delta]$, and $\mathbf{F}$ is a diagonal matrix with diagonal elements $f(s) \doteq d_\mu(s) \lim_{t\to\infty} \mathbb{E}_\mu[F_k|S_k = s]$, which we assume exists. The vector $\mathbf{f} \in \mathbb{R}^N$ with components $[\mathbf{f}]_s \doteq f(s)$ can be written as

$$\begin{aligned} \mathbf{f} &= \mathbf{d}_\mu + \gamma \mathbf{P}_\pi^\top \mathbf{d}_\mu + (\gamma \mathbf{P}_\pi^\top)^2 \mathbf{d}_\mu + \dots \\ &= (\mathbf{I} - \gamma \mathbf{P}_\pi^\top)^{-1} \mathbf{d}_\mu. \end{aligned} \tag{21}$$

Theoretical Analysis

The purpose of this section is to establish the stabilities of the VMTD algorithm and the VMTDC algorithm, and also presents a corollary on the convergence rate of VMTD.

Theorem 1. *(Convergence of VMTD). In the case of on-policy learning, consider the iterations (3) and (5) with (4) of VMTD. Let the step-size sequences α_k and β_k, $k \geq 0$ satisfy in this case $\alpha_k, \beta_k > 0$, for all k, $\sum_{k=0}^\infty \alpha_k = \sum_{k=0}^\infty \beta_k = \infty$, $\sum_{k=0}^\infty \alpha_k^2 < \infty$, $\sum_{k=0}^\infty \beta_k^2 < \infty$, and $\alpha_k = o(\beta_k)$. Assume that $(\boldsymbol{\phi}_k, r_k, \boldsymbol{\phi}'_k)$ is an i.i.d. sequence with uniformly bounded second moments, where $\boldsymbol{\phi}_k$ and $\boldsymbol{\phi}'_k$ are sampled from the same Markov chain. Let $\boldsymbol{A} = \mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi} - \gamma\boldsymbol{\phi}')$, $\boldsymbol{b} = \mathrm{Cov}(r, \boldsymbol{\phi})$. Assume that matrix $\boldsymbol{\theta}$ is non-singular. Then the parameter vector $\boldsymbol{\theta}_k$ converges with probability one to $\boldsymbol{A}^{-1}\boldsymbol{b}$.*

Proof. The proof is based on Borkar's Theorem for general stochastic approximation recursions with two time scales (Borkar 1997).

A new one-step linear TD solution is defined as:

$$0 = \mathbb{E}[(\delta - \mathbb{E}[\delta])\boldsymbol{\phi}] = -\mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}.$$

Thus, the VMTD's solution is $\boldsymbol{\theta}_{\text{VMTD}} = \mathbf{A}^{-1}\boldsymbol{b}$.

First, note that recursion (5) can be rewritten as

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \beta_k \boldsymbol{\xi}(k),$$

where

$$\boldsymbol{\xi}(k) = \frac{\alpha_k}{\beta_k}(\delta_k - \omega_k)\boldsymbol{\phi}_k$$

Due to the settings of step-size schedule $\alpha_k = o(\beta_k)$, $\boldsymbol{\xi}(k) \to 0$ almost surely as $k \to \infty$. That is the increments in iteration (3) are uniformly larger than those in (5), thus (3) is the faster recursion. Along the faster time scale, iterations of (3) and (5) are associated to ODEs system as follows:

$$\dot{\boldsymbol{\theta}}(t) = 0, \tag{22}$$

$$\dot{\omega}(t) = \mathbb{E}[\delta_t|\boldsymbol{\theta}(t)] - \omega(t). \tag{23}$$

Based on the ODE (22), $\boldsymbol{\theta}(t) \equiv \boldsymbol{\theta}$ when viewed from the faster timescale. By the Hirsch lemma (Hirsch 1989), it follows that $||\boldsymbol{\theta}_k - \boldsymbol{\theta}|| \to 0$ a.s. as $k \to \infty$ for some $\boldsymbol{\theta}$ that depends on the initial condition $\boldsymbol{\theta}_0$ of recursion (5). Thus, the ODE pair (22)-(23) can be written as

$$\dot{\omega}(t) = \mathbb{E}[\delta_t|\boldsymbol{\theta}] - \omega(t). \tag{24}$$

Consider the function $h(\omega) = \mathbb{E}[\delta|\boldsymbol{\theta}] - \omega$, i.e., the driving vector field of the ODE (24). It is easy to find that the function h is Lipschitz with coefficient -1. Let $h_\infty(\cdot)$ be the function defined by $h_\infty(\omega) = \lim_{x\to\infty} \frac{h(x\omega)}{x}$. Then $h_\infty(\omega) = -\omega$, is well-defined. For (24), $\omega^* = \mathbb{E}[\delta|\boldsymbol{\theta}]$ is the unique globally asymptotically stable equilibrium. For the ODE

$$\dot{\omega}(t) = h_\infty(\omega(t)) = -\omega(t), \tag{25}$$

apply $\vec{V}(\omega) = (-\omega)^\top(-\omega)/2$ as its associated strict Liapunov function. Then, the origin of (25) is a globally asymptotically stable equilibrium.

Consider now the recursion (3). Let $M_{k+1} = (\delta_k - \omega_k) - \mathbb{E}[(\delta_k - \omega_k)|\mathcal{F}(k)]$, where $\mathcal{F}(k) = \sigma(\omega_l, \boldsymbol{\theta}_l, l \leq k; \boldsymbol{\phi}_s, \boldsymbol{\phi}'_s, r_s, s < k)$, $k \geq 1$ are the sigma fields generated by $\omega_0, \boldsymbol{\theta}_0, \omega_{l+1}, \boldsymbol{\theta}_{l+1}, \boldsymbol{\phi}_l, \boldsymbol{\phi}'_l$, $0 \leq l < k$. It is easy to verify that $M_{k+1}, k \geq 0$ are integrable random variables that satisfy $\mathbb{E}[M_{k+1}|\mathcal{F}(k)] = 0$, $\forall k \geq 0$. Because $\boldsymbol{\phi}_k$, r_k, and $\boldsymbol{\phi}'_k$ have uniformly bounded second moments, it can be seen that for some constant $c_1 > 0$, $\forall k \geq 0$,

$$\mathbb{E}[||M_{k+1}||^2|\mathcal{F}(k)] \leq c_1(1 + ||\omega_k||^2 + ||\boldsymbol{\theta}_k||^2).$$

Now Assumptions (A1) and (A2) of (Borkar and Meyn 2000) are verified. Furthermore, Assumptions (TS) of (Borkar and Meyn 2000) is satisfied by our conditions on the step-size sequences α_k, β_k. Thus, by Theorem 2.2 of (Borkar and Meyn 2000) we obtain that $||\omega_k - \omega^*|| \to 0$ almost surely as $k \to \infty$.

Consider now the slower time scale recursion (5). Based on the above analysis, (5) can be rewritten as

$$\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \alpha_k(\delta_k - \mathbb{E}[\delta_k|\boldsymbol{\theta}_k])\boldsymbol{\phi}_k.$$

Let $\mathcal{G}(k) = \sigma(\boldsymbol{\theta}_l, l \leq k; \boldsymbol{\phi}_s, \boldsymbol{\phi}'_s, r_s, s < k)$, $k \geq 1$ be the sigma fields generated by $\boldsymbol{\theta}_0, \boldsymbol{\theta}_{l+1}, \boldsymbol{\phi}_l, \boldsymbol{\phi}'_l$, $0 \leq l < k$. Let $Z_{k+1} = Y_k - \mathbb{E}[Y_k|\mathcal{G}(k)]$, where

$$Y_k = (\delta_k - \mathbb{E}[\delta_k|\boldsymbol{\theta}_k])\boldsymbol{\phi}_k.$$

Consequently,

$$\begin{aligned} \mathbb{E}[Y_k|\mathcal{G}(k)] &= \mathbb{E}[(\delta_k - \mathbb{E}[\delta_k|\boldsymbol{\theta}_k])\boldsymbol{\phi}_k|\mathcal{G}(k)] \\ &= \mathbb{E}[\delta_k\boldsymbol{\phi}_k|\boldsymbol{\theta}_k] - \mathbb{E}[\mathbb{E}[\delta_k|\boldsymbol{\theta}_k]\boldsymbol{\phi}_k] \\ &= \mathbb{E}[\delta_k\boldsymbol{\phi}_k|\boldsymbol{\theta}_k] - \mathbb{E}[\delta_k|\boldsymbol{\theta}_k]\mathbb{E}[\boldsymbol{\phi}_k] \\ &= \mathrm{Cov}(\delta_k|\boldsymbol{\theta}_k, \boldsymbol{\phi}_k), \end{aligned}$$

where $\mathrm{Cov}(\cdot,\cdot)$ is a covariance operator.

Thus,

$$Z_{k+1} \quad = \quad (\delta_k - \mathbb{E}[\delta_k|\boldsymbol{\theta}_k])\boldsymbol{\phi}_k - \mathrm{Cov}(\delta_k|\boldsymbol{\theta}_k, \boldsymbol{\phi}_k).$$

It is easy to verify that $Z_{k+1}, k \geq 0$ are integrable random variables that satisfy $\mathbb{E}[Z_{k+1}|\mathcal{G}(k)] = 0, \forall k \geq 0$. Also, because $\boldsymbol{\phi}_k$, r_k, and $\boldsymbol{\phi}'_k$ have uniformly bounded second moments, it can be seen that for some constant $c_2 > 0, \forall k \geq 0$,

$$\mathbb{E}[||Z_{k+1}||^2|\mathcal{G}(k)] \leq c_2(1 + ||\boldsymbol{\theta}_k||^2).$$

Consider now the following ODE associated with (5):

$$\begin{aligned}
\dot{\boldsymbol{\theta}}(t) &= \mathrm{Cov}(\delta|\boldsymbol{\theta}(t), \boldsymbol{\phi}) \\
&= \mathrm{Cov}(r + (\gamma\boldsymbol{\phi}' - \boldsymbol{\phi})^\top\boldsymbol{\theta}(t), \boldsymbol{\phi}) \\
&= \mathrm{Cov}(r, \boldsymbol{\phi}) - \mathrm{Cov}(\boldsymbol{\theta}(t)^\top(\boldsymbol{\phi} - \gamma\boldsymbol{\phi}'), \boldsymbol{\phi}) \\
&= \mathrm{Cov}(r, \boldsymbol{\phi}) - \boldsymbol{\theta}(t)^\top\mathrm{Cov}(\boldsymbol{\phi} - \gamma\boldsymbol{\phi}', \boldsymbol{\phi}) \\
&= \mathrm{Cov}(r, \boldsymbol{\phi}) - \mathrm{Cov}(\boldsymbol{\phi} - \gamma\boldsymbol{\phi}', \boldsymbol{\phi})^\top\boldsymbol{\theta}(t) \\
&= \mathrm{Cov}(r, \boldsymbol{\phi}) - \mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi} - \gamma\boldsymbol{\phi}')\boldsymbol{\theta}(t) \\
&= -\mathbf{A}\boldsymbol{\theta}(t) + \boldsymbol{b}.
\end{aligned} \tag{26}$$

Let $\vec{h}(\boldsymbol{\theta}(t))$ be the driving vector field of the ODE (26).

$$\vec{h}(\boldsymbol{\theta}(t)) = -\mathbf{A}\boldsymbol{\theta}(t) + \boldsymbol{b}.$$

Consider the cross-covariance matrix,

$$\begin{aligned}
\mathbf{A} &= \mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi} - \gamma\boldsymbol{\phi}') \\
&= \frac{\mathrm{Cov}(\boldsymbol{\phi},\boldsymbol{\phi})+\mathrm{Cov}(\boldsymbol{\phi}-\gamma\boldsymbol{\phi}',\boldsymbol{\phi}-\gamma\boldsymbol{\phi}')-\mathrm{Cov}(\gamma\boldsymbol{\phi}',\gamma\boldsymbol{\phi}')}{2} \\
&= \frac{\mathrm{Cov}(\boldsymbol{\phi},\boldsymbol{\phi})+\mathrm{Cov}(\boldsymbol{\phi}-\gamma\boldsymbol{\phi}',\boldsymbol{\phi}-\gamma\boldsymbol{\phi}')-\gamma^2\mathrm{Cov}(\boldsymbol{\phi}',\boldsymbol{\phi}')}{2} \\
&= \frac{(1-\gamma^2)\mathrm{Cov}(\boldsymbol{\phi},\boldsymbol{\phi})+\mathrm{Cov}(\boldsymbol{\phi}-\gamma\boldsymbol{\phi}',\boldsymbol{\phi}-\gamma\boldsymbol{\phi}')}{2},
\end{aligned} \tag{27}$$

where we eventually used $\mathrm{Cov}(\boldsymbol{\phi}', \boldsymbol{\phi}') = \mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi})$ [1]. Note that the covariance matrix $\mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi})$ and $\mathrm{Cov}(\boldsymbol{\phi} - \gamma\boldsymbol{\phi}', \boldsymbol{\phi} - \gamma\boldsymbol{\phi}')$ are semi-positive definite. Then, the matrix $\mathbf{A}$ is semi-positive definite because $\mathbf{A}$ is linearly combined by two positive-weighted semi-positive definite matrice (27). Furthermore, $\mathbf{A}$ is nonsingular due to the assumption. Hence, the cross-covariance matrix $\mathbf{A}$ is positive definite.

Therefore, $\boldsymbol{\theta}^* = \mathbf{A}^{-1}\boldsymbol{b}$ can be seen to be the unique globally asymptotically stable equilibrium for ODE (26). Let $\vec{h}_\infty(\boldsymbol{\theta}) = \lim_{r\to\infty}\frac{\vec{h}(r\boldsymbol{\theta})}{r}$. Then $\vec{h}_\infty(\boldsymbol{\theta}) = -\mathbf{A}\boldsymbol{\theta}$ is well-defined. Consider now the ODE

$$\dot{\boldsymbol{\theta}}(t) = -\mathbf{A}\boldsymbol{\theta}(t). \tag{28}$$

The ODE (28) has the origin as its unique globally asymptotically stable equilibrium. Thus, the assumption (A1) and (A2) are verified. □

Theorem 2. *(Convergence of VMTDC). In the case of off-policy learning, consider the iterations (13), (12) and (11) of VMTDC. Let the step-size sequences α_k, ζ_k and β_k, $k \geq 0$ satisfy in this case $\alpha_k, \zeta_k, \beta_k > 0$, for all k, $\sum_{k=0}^\infty \alpha_k = \sum_{k=0}^\infty \beta_k = \sum_{k=0}^\infty \zeta_k = \infty$, $\sum_{k=0}^\infty \alpha_k^2 < \infty$, $\sum_{k=0}^\infty \zeta_k^2 < \infty$, $\sum_{k=0}^\infty \beta_k^2 < \infty$, and $\alpha_k = o(\zeta_k)$, $\zeta_k = o(\beta_k)$. Assume that $(\boldsymbol{\phi}_k, r_k, \boldsymbol{\phi}'_k)$ is an i.i.d. sequence with uniformly bounded second moments. Let $\boldsymbol{A} = \mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi} - \gamma\boldsymbol{\phi}')$, $\boldsymbol{b} = \mathrm{Cov}(r, \boldsymbol{\phi})$, and $\boldsymbol{A} = \mathbb{E}[\boldsymbol{\phi}\boldsymbol{\phi}^\top]$. Assume that $\boldsymbol{A}$ and $\boldsymbol{C}$ are non-singular matrices. Then the parameter vector $\boldsymbol{\theta}_k$ converges with probability one to $\boldsymbol{A}^{-1}\boldsymbol{b}$.*

[1] The covariance matrix $\mathrm{Cov}(\boldsymbol{\phi}', \boldsymbol{\phi}')$ is equal to the covariance matrix $\mathrm{Cov}(\boldsymbol{\phi}, \boldsymbol{\phi})$ if the initial state is re-reachable or initialized randomly in a Markov chain for on-policy update.

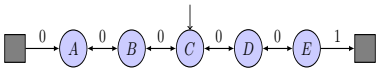

Figure 1: Random walk.

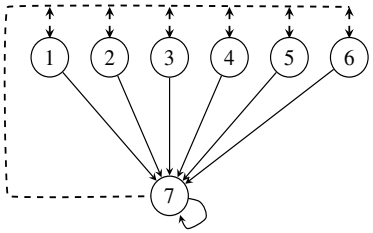

Figure 2: 7-state version of Baird's off-policy counterexample.

Theorem 3. *(Convergence of VMETD). In the case of off-policy learning, consider the iterations (18), (20) and (19) with (4) of VMETD. Let the step-size sequences α_k and β_k, $k \geq 0$ satisfy in this case $\alpha_k, \beta_k > 0$, for all k, $\sum_{k=0}^\infty \alpha_k = \sum_{k=0}^\infty \beta_k = \infty$, $\sum_{k=0}^\infty \alpha_k^2 < \infty$, $\sum_{k=0}^\infty \beta_k^2 < \infty$, and $\alpha_k = o(\beta_k)$. Assume that $(\boldsymbol{\phi}_k, r_k, \boldsymbol{\phi}'_k)$ is an i.i.d. sequence with uniformly bounded second moments, where $\boldsymbol{\phi}_k$ and $\boldsymbol{\phi}'_k$ are sampled from the same Markov chain. Let $\boldsymbol{A_{VMETD}} = \boldsymbol{\Phi}^\top(\boldsymbol{F}(\boldsymbol{I} - \gamma\boldsymbol{P}_\pi) - \boldsymbol{d}_\mu\boldsymbol{d}_\mu^\top)\boldsymbol{\Phi}$, $\boldsymbol{b_{VMETD}} = \boldsymbol{\Phi}^\top(\boldsymbol{F} - \boldsymbol{d}_\mu\boldsymbol{d}_\mu^\top)\boldsymbol{r}_\pi$. Assume that matrix A is non-singular. Then the parameter vector $\boldsymbol{\theta}_k$ converges with probability one to $\boldsymbol{A}_{\boldsymbol{VMETD}}^{-1}\boldsymbol{b}_{\boldsymbol{VMETD}}$.*

Experimental Studies

This section assesses algorithm performance through experiments, which are divided into policy evaluation experiments and control experiments.

Testing Tasks

Random-walk: as shown in Figure 1, all episodes start in the center state, C, and proceed to left or right by one state on each step, equiprobably. Episodes terminate either on the extreme left or the extreme right, and get a reward of $+1$ if terminate on the right, or 0 in the other case. In this task, the true value for each state is the probability of starting from that state and terminating on the right (Sutton and Barto 2018). Thus, the true values of states from A to E are $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}$, respectively. The discount factor $\gamma = 1.0$. There are three standard kinds of features for random-walk problems: tabular feature, inverted feature and dependent feature (Sutton et al. 2009). The feature matrices corresponding to three random walks are shown in Appendix. Conduct experiments using an on-policy approach in the Random-walk environment.

Baird's off-policy counterexample: This task is well known as a counterexample, in which TD diverges (Baird

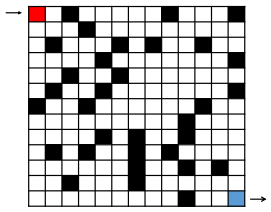

Figure 3: Maze.

et al. 1995; Sutton et al. 2009). As shown in Figure 2, reward for each transition is zero. Thus the true values are zeros for all states and for any given policy. The behaviour policy chooses actions represented by solid lines with a probability of $\frac{1}{7}$ and actions represented by dotted lines with a probability of $\frac{6}{7}$. The target policy is expected to choose the solid line with more probability than $\frac{1}{7}$, and it chooses the solid line with probability of 1 in this paper. The discount factor $\gamma = 0.99$, and the feature matrix is defined in Appendix.

Maze: The learning agent should find a shortest path from the upper left corner to the lower right corner. In each state, there are four alternative actions: up, $down$, $left$, and $right$, which takes the agent deterministically to the corresponding neighbour state, except when a movement is blocked by an obstacle or the edge of the maze. Rewards are -1 in all transitions until the agent reaches the goal state. The discount factor $\gamma = 0.99$, and states s are represented by tabular features.The maximum number of moves in the game is set to 1000.

The other three control environments: Cliff Walking, Mountain Car, and Acrobot are selected from the gym official website and correspond to the following versions: "CliffWalking-v0", "MountainCar-v0" and "Acrobot-v1". For specific details, please refer to the gym official website. The maximum number of steps for the Mountain Car environment is set to 1000, while the default settings are used for the other two environments. In Mountain car and Acrobot, features are generated by tile coding.

Please, refer to the Appendix for the selection of learning rates for all experiments.

Experimental Results and Analysis

The experiment needs further elaboration.

Related Work

Difference between VMQ and R-learning

Tabular VMQ's update formula bears some resemblance to R-learning's update formula. As shown in Table **??**, the update formulas of the two algorithms have the following differences:
(1) The goal of the R-learning algorithm (Schwartz 1993) is to maximize the average reward, rather than the cumulative reward, by learning an estimate of the average reward. This estimate m is then used to update the Q-values. On the contrary, the ω in the tabular VMQ update formula eventually converges to $\mathbb{E}[\delta]$.
(2) When $\gamma = 1$ in the tabular VMQ update formula, the R-learning update formula is formally the same as the tabular VMQ update formula. Therefore, R-learning algorithm can be considered as a special case of VMQ algorithm in form.

Variance Reduction for TD Learning

The TD with centering algorithm (CTD) (Korda and La 2015) was proposed, which directly applies variance reduction techniques to the TD algorithm. The CTD algorithm updates its parameters using the average gradient of a batch of Markovian samples and a projection operator. Unfortunately, the authors' analysis of the CTD algorithm contains technical errors. The VRTD algorithm (Xu et al. 2020) is also a variance-reduced algorithm that updates its parameters using the average gradient of a batch of i.i.d. samples. The authors of VRTD provide a technically sound analysis to demonstrate the advantages of variance reduction.

Variance Reduction for Policy Gradient Algorithms

Policy gradient algorithms are a class of reinforcement learning algorithms that directly optimize cumulative rewards. REINFORCE is a Monte Carlo algorithm that estimates gradients through sampling, but may have a high variance. Baselines are introduced to reduce variance and to accelerate learning (Sutton and Barto 2018). In Actor-Critic, value function as a baseline and bootstrapping are used to reduce variance, also accelerating convergence (Sutton and Barto 2018). TRPO (Schulman et al. 2015) and PPO (Schulman et al. 2017) use generalized advantage estimation, which combines multi-step bootstrapping and Monte Carlo estimation to reduce variance, making gradient estimation more stable and accelerating convergence.

In Variance Minimization, the incorporation of $\omega \doteq \mathbb{E}[\delta]$ bears a striking resemblance to the use of a baseline in policy gradient methods. The introduction of a baseline in policy gradient techniques does not alter the expected value of the update; rather, it significantly impacts the variance of gradient estimation. The addition of $\omega \doteq \mathbb{E}[\delta]$ in Variance Minimization preserves the invariance of the optimal policy while stabilizing gradient estimation, reducing the variance of gradient estimation, and hastening convergence.

Conclusion and Future Work

Value-based reinforcement learning typically aims to minimize error as an optimization objective. As an alternation, this study proposes new objective functions: VBE and VPBE, and derives many variance minimization algorithms, including VMTD, VMTDC and VMETD. All algorithms demonstrated superior performance in policy evaluation and control experiments. Future work may include, but are not limited to, (1) analysis of the convergence rate of VMTDC and VMETD. (2) extensions of VBE and VPBE to multi-step returns. (3) extensions to nonlinear approximations, such as neural networks.

References

Baird, L.; et al. 1995. Residual algorithms: Reinforcement learning with function approximation. In *Proc. 12th Int. Conf. Mach. Learn.*, 30–37.

Bas-Serrano, J.; Curi, S.; Krause, A.; and Neu, G. 2021. Logistic Q-Learning. In *International Conference on Artificial Intelligence and Statistics*, 3610–3618.

Borkar, V. S. 1997. Stochastic approximation with two time scales. *Syst. & Control Letters*, 29(5): 291–294.

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

Chen, X.; Ma, X.; Li, Y.; Yang, G.; Yang, S.; and Gao, Y. 2023. Modified Retrace for Off-Policy Temporal Difference Learning. In *Uncertainty in Artificial Intelligence*, 303–312. PMLR.

Devlin, S.; and Kudenko, D. 2012. Dynamic potential-based reward shaping. In *Proc. 11th Int. Conf. Autonomous Agents and Multiagent Systems*, 433–440.

Feng, Y.; Li, L.; and Liu, Q. 2019. A kernel loss for solving the Bellman equation. In *Advances in Neural Information Processing Systems*, 15430–15441.

Givchi, A.; and Palhang, M. 2015. Quasi newton temporal difference learning. In *Asian Conference on Machine Learning*, 159–172.

Hackman, L. 2012. *Faster Gradient-TD Algorithms*. Ph.D. thesis, University of Alberta.

Hallak, A.; Tamar, A.; Munos, R.; and Mannor, S. 2016. Generalized emphatic temporal difference learning: bias-variance analysis. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 1631–1637.

Hirsch, M. W. 1989. Convergent activation dynamics in continuous time networks. *Neural Netw.*, 2(5): 331–349.

Johnson, R.; and Zhang, T. 2013. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 315–323.

Korda, N.; and La, P. 2015. On TD (0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International conference on machine learning*, 626–634. PMLR.

Liu, B.; Gemp, I.; Ghavamzadeh, M.; Liu, J.; Mahadevan, S.; and Petrik, M. 2018. Proximal gradient temporal difference learning: Stable reinforcement learning with polynomial sample complexity. *Journal of Artificial Intelligence Research*, 63: 461–494.

Liu, B.; Liu, J.; Ghavamzadeh, M.; Mahadevan, S.; and Petrik, M. 2015. Finite-sample analysis of proximal gradient TD algorithms. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, 504–513.

Liu, B.; Liu, J.; Ghavamzadeh, M.; Mahadevan, S.; and Petrik, M. 2016. Proximal Gradient Temporal Difference Learning Algorithms. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 4195–4199.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proc. 16th Int. Conf. Mach. Learn.*, 278–287.

Pan, Y.; White, A.; and White, M. 2017. Accelerated gradient temporal difference learning. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, 2464–2470.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *International Conference on Machine Learning*, 1889–1897.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Schwartz, A. 1993. A reinforcement learning method for maximizing undiscounted rewards. In *Proc. 10th Int. Conf. Mach. Learn.*, volume 298, 298–305.

Sutton, R.; Maei, H.; Precup, D.; Bhatnagar, S.; Silver, D.; Szepesvári, C.; and Wiewiora, E. 2009. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proc. 26th Int. Conf. Mach. Learn.*, 993–1000.

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

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Sutton, R. S.; Maei, H. R.; and Szepesvári, C. 2008. A Convergent $O(n)$ Temporal-difference Algorithm for Off-policy Learning with Linear Function Approximation. In *Advances in Neural Information Processing Systems*, 1609–1616. Cambridge, MA: MIT Press.

Sutton, R. S.; Mahmood, A. R.; and White, M. 2016. An emphatic approach to the problem of off-policy temporal-difference learning. *The Journal of Machine Learning Research*, 17(1): 2603–2631.

Tsitsiklis, J. N.; and Van Roy, B. 1997. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, 1075–1081.

Xu, T.; Wang, Z.; Zhou, Y.; and Liang, Y. 2019. Reanalysis of Variance Reduced Temporal Difference Learning. In *International Conference on Learning Representations*.

Xu, T.; Wang, Z.; Zhou, Y.; and Liang, Y. 2020. Reanalysis of variance reduced temporal difference learning. *arXiv preprint arXiv:2001.01898*.

Zhang, S.; and Whiteson, S. 2022. Truncated emphatic temporal difference methods for prediction and control. *The Journal of Machine Learning Research*, 23(1): 6859–6917.