Modified Retrace for Off-Policy Temporal Difference Learning with Linear Function Approximation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Off-policy learning is key to extend reinforcement learning as it allows to learn about a target policy from a different behavior policy that generates the data. However, it is well known as "the deadly triad" when combined with bootstrapping and function approximation. Retrace is an efficient and convergent off-policy algorithm with tabular value functions which employs truncated importance sampling ratios. Unfortunately, Retrace is known to be unstable with linear function approximation. In this paper, we propose modified Retrace to measure the "off-policyness" between the behavior policy and the target policy, derive a new off-policy temporal difference learning algorithm with linear function approximation, and obtain a convergence guarantee under standard assumptions. Experimental results on counterexamples and control tasks validate the effectiveness of the proposed algorithm compared with traditional algorithms.

1 Positive definite matrix in off-policy learning methods

Positive definite (p.d.) matrix plays an important role in convergence analysis of reinforcement learning algorithms with linear function approximation. The convergence of on-policy TD(0) is established by Sutton [1988], where the key is the p.d. matrix $\mathbf{A}_{\text{on}}$ based on the invariance of the on-policy state distribution. Off-policy learning seeks to learn a target policy while exploring actions according to a behavior policy to avoid getting stuck in local optima. However, due to the inconsistency between the behavior policy μ and the target policy π, off-policy learning is prone to be instable, especially when combined with function approximation and bootstrapping, known as "the deadly triad" (Sutton and Barto [2018]). The fundamental reason is that the positive definiteness of the matrix $\mathbf{A}_{\text{off}}$ is not guaranteed (Sutton *et al.* [2016]).

Baird and others [1995] proposed residual algorithms by minimizing mean squared Bellman errors to solve the residual fixed point in closed-form. The key matrix is p.d., thus ensuring the stability of the algorithms. However, residual methods require double sampling in non-deterministic environments to remove dependencies between successor states. More importantly, the residual fixed point is in most cases worse than the TD fixed point (Scherrer [2010]).

Stable algorithms to solve the TD fixed point mainly include two approaches. Gradient based methods guarantee the positive definiteness of the correlation matrix by constructing different objective functions. Sutton *et al.* [2008] proposed the first convergent off-policy temporal difference learning algorithm, gradient TD (GTD), which minimizes the norm of the expected TD update (NEU) and involves a p.d. matrix $\mathbf{A}_{\text{GTD}} = \begin{pmatrix} \sqrt{\eta}\mathbf{I} & \mathbf{A}_{\text{off}} \\ -\mathbf{A}_{\text{off}}^\top & 0 \end{pmatrix}$, where $\mathbf{I}$ is the identity matrix and η is the

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

stepsize ratio of the auxiliary parameter to the learning parameter. Sutton *et al.* [2009] proposed GTD2 algorithm with p.d. matrix $\mathbf{A}_{\text{GTD2}} = \begin{pmatrix} \sqrt{\eta}\mathbf{C} & \mathbf{A}_{\text{off}} \\ -\mathbf{A}_{\text{off}}^\top & 0 \end{pmatrix}$ and TD with gradient correction (TDC) algorithm with p.d. matrix $\mathbf{A}_{\text{TDC}} = \mathbf{A}_{\text{off}}^\top \mathbf{C}^{-1} \mathbf{A}_{\text{off}}$, both of which minimize the mean square projected Bellman error (MSPBE), where $\mathbf{C} = \mathbb{E}[\phi\phi^\top]$ and ϕ is feature of state or state-action pair. Hackman [2012] proposed Hybrid TD (HTD) algorithm with a p.d. matrix $\mathbf{A}_{\text{HTD}} = \mathbf{A}_{\text{off}}^\top \mathbf{A}_{\text{on}}^{-1} \mathbf{A}_{\text{off}}$, which replaces $\mathbf{C}^{-1}$ in $\mathbf{A}_{\text{TDC}}$ as $\mathbf{A}_{\text{on}}^{-1}$ to accelerate the learning rate. Liu *et al.* [2015, 2016, 2018] proposed accelerated GTD-MP and GTD2-MP algorithm via writing the objective functions, NEU and MSPBE, in the form of a convex-concave saddle-point formulation. Zhang *et al.* [2021] proposed Diff-GQ1 algorithm w.r.t saddle-point formulation of GTD2 and Diff-GQ2 algorithm w.r.t two-stage gradient evaluation, both of which minimize MSPBE in the average-reward setting. Pan *et al.* [2017] proposed accelerated TD (ATD) algorithm based on a second order update, which is a Hessian matrix derived from MSPBE and approximated in a low rank. The main disadvantage of gradient based methods is that they need to update an additional parameter to correct the gradient and tend to converge slowly (Hallak and Mannor [2017]).

The other approach utilizes importance sampling (IS) ratios to correct the state distribution between on-policy and off-policy updates. Precup *et al.* [2001] proposed per-episode updating version of stable off-policy linear TD learning algorithm, which attempts to completely correct the state distribution using IS ratios from behavior policy to target policy but turned out to have extremely high variance, where the p.d. matrix is $\mathbf{A}_{\text{on}}$. Sutton *et al.* [2016] proposed emphatic TD (ETD) algorithm with followon trace to correct from beginning of the excursion based on IS ratios, where p.d. matrix is $\mathbf{A}_{\text{ETD}} = \Phi^\top \mathbf{D}_f (\mathbf{I} - \gamma \mathbf{P}_\pi)\Phi$, $\mathbf{D}_f$ is a diagonal matrix with diagonal element approximated to $f = (\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} d_\mu$. Zhang *et al.* [2020] proposed convergent off-policy actor-critic algorithm in which the followon trace's variance is reduced by emphasis approximation. Zhang and Whiteson [2021] proposed truncated emphatic TD (TETD), where the p.d. matrix is $\mathbf{A}_{\text{TETD}} = \Phi^\top \mathbf{D}_{f_k} (\mathbf{I} - \gamma \mathbf{P}_\pi)\Phi$, f_k is a truncated followon trace of length k. The main disadvantage of ETD and TETD is that the followon trace may be of very high variance. Munos *et al.* [2016] proposed Retrace algorithm with a safe and efficient IS ratios truncated at 1, which guarantees convergence with a contraction mapping in the case of look-up table. However, based on an action-value extension to Baird's counterexample, Retrace was pointed out that it is not guaranteed to be stable when combined with function approximation (Touati *et al.* [2018]). Then, a convergent gradient-based Retrace (GRetrace) was proposed based on a quadratic convex-concave saddle-point formulation, which minimizes MSPBE (Touati *et al.* [2018]). However, this returns to the disadvantage of slow convergence of the gradient TD learning families.

Our motivation: Another discipline treats the divergence of Q-learning as an overestimation problem and employs underestimation to enhance the stability. Double Q-learning (Hasselt [2010] and Double-DQN (Hasselt *et al.* [2016]) employ two estimators, using one estimator to obtain the optimal action and the other estimator to obtain the corresponding action value. Averaged-DQN (Anschel *et al.* [2017]) uses the K previously learned value estimates to produce the current action-value estimate. Soft operators underestimate of the maximum action value by combining all action values, e.g., tree backup (Precup *et al.* [2000]), mellowmax (Asadi and Littman [2017]; Kim *et al.* [2019]), and softmax (Song *et al.* [2019]). We argue that overestimation is actually a manifestation of the divergence of the Q-learning algorithm, but it may not be the essence of the divergence. As a solution, underestimation works well in practice, but it ignores the degree of off-policyness and thus lacks a direct connection to the proof of convergence.

In this paper, we first revisit the fundamental reason why Retrace with linear function approximation is not stable, then build a connection between the underestimation and the convergence proof, and propose the modified Retrace (MRetrace) learning algorithm with "off-policyness" ratios (see Table 1). Finally, we show that the key matrix of MRetrace is p.d., resulting in a convergence guarantee for MRetrace under standard conditions in the off-policy setting.

Table 1: Comparisons of off-policy learning algorithms with linear function approximation.

Name	definition	update rules	p.d.
Off-policy TD	$\rho_t = \frac{\pi(a_t\|s_t)}{\mu(a_t\|s_t)}$	$\alpha_t\rho_t\left(r_{t+1} + (\gamma\phi_{t+1} - \phi_t)^\top\theta_t\right)\phi_t$	no
Retrace	$c_t = \min\left(1, \frac{\pi(a_t\|s_t)}{\mu(a_t\|s_t)}\right)$	$\alpha_t c_t\left(\mathbb{E}_\pi[r_{t+1}] + (\gamma\mathbb{E}_\pi[\phi_{t+1}] - \phi_t)^\top\theta_t\right)\phi_t$	no
Modified Retrace	$x_t = \min_a\left\{\frac{\mu(a\|s_t)}{\pi(a\|s_t)}\right\}$	$\alpha_t\left(\rho_t r_{t+1} + (x_t\gamma\mathbb{E}_\pi[\phi_{t+1}] - \phi_t)^\top\theta_t\right)\phi_t$	yes

2 Notation and background

Reinforcement learning agent interacts with its environment which we modeled as a discounted Markov Decision Process $\langle S, A, R, T, \gamma\rangle$, where S is a finite state space, $|S| = n$, A is an action space, $T : S \times A \times S \to [0, 1]$ is a transition function, $R : S \times A \times S \to \mathbb{R}$ is a reward function, $\gamma \in [0, 1)$ is a discount factor. Policy $\pi : S \times A \to [0, 1]$ offers the probability $\pi(a|s)$ to choose action a in state s. The value function for policy π, denoted $V^\pi : S \to \mathbb{R}$, represents the expected sum of discounted rewards in the MDP under policy π: $V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t r_t | s_0 = s\right]$. $V^\pi(s)$ is the fixed point of the Bellman operator over the value function $\mathcal{T}^\pi V = r + \gamma\mathbf{P}_\pi V$, where r is the expected immediate reward and $\mathbf{P}_\pi$ denotes the $n \times n$ matrix of transition probabilities

$$[\mathbf{P}_\pi]_{ij} \doteq \sum_{a\in A}\pi(a|i)T(i, a, j). \tag{1}$$

Assume the state distribution d_π under policy π is steady and exists. Then the special property (Sutton *et al.* [2016]) of d_π is that

$$d_\pi = \mathbf{P}_\pi^\top d_\pi. \tag{2}$$

When the state space is too large to preserve $V^\pi(s)$, a linear function approximation is used to generalize between different states $V^\pi(s) \approx V_\theta(s) = \theta^\top\phi(s) = \sum_{i=1}^m \theta_i\phi_i(s)$, where θ is the weight vector, $\phi(s)$ is the feature vector of state s, and feature size is far less than state space $m \ll n$. Notably, equation $V_\theta = \mathcal{T}^\pi V_\theta$ no longer holds because the number of parameters is far less than the number of equations. A common and efficient solution is the TD fixed point $V_\theta = \Pi\mathcal{T}^\pi V_\theta$ with projection $\Pi = \Phi(\Phi^\top\mathbf{D}_\pi\Phi)^{-1}\Phi^\top\mathbf{D}_\pi$, where Φ is the $n \times m$ matrix with the $\phi(s)$ as its rows, $\mathbf{D}_\pi$ is the $n \times n$ diagonal matrix with d_π on its diagonal. It can be learned by the on-policy TD(0) algorithm:

$$\begin{aligned}\theta_{t+1} \doteq& \theta_t + \alpha_t\left(r_{t+1} + \gamma\theta_t^\top\phi_{t+1} - \theta_t^\top\phi_t\right)\phi_t\\ =& \theta_t + \alpha_t\left(r_{t+1}\phi_t - \phi_t(\phi_t - \gamma\phi_{t+1})^\top\theta_t\right),\end{aligned} \tag{3}$$

where $\alpha_t > 0$ is a step-size parameter, and we have used the shorthand $\phi_t \doteq \phi(s_t)$. Algorithms' convergence analysis with linear function approximation is mainly based on the ODE (Ordinary Differential Equations) approach (Borkar and Meyn [2000]), where the key relies on the matrix $\mathbf{A}$ being positive definite i.e. $\forall x \neq 0$, $x^\top\mathbf{A}x > 0$. Let $\mathbf{A}_{\text{on}}$ denote the key matrix of the expected update (3):

$$\begin{aligned}\mathbf{A}_{\text{on}} &= \lim_{t\to\infty}\mathbb{E}_\pi\left[\phi_t(\phi_t - \gamma\phi_{t+1})^\top\right]\\ &= \sum_s d_\pi(s)\mathbb{E}_\pi\left[\phi_t(\phi_t - \gamma\phi_{t+1})^\top | S_t = s\right]\\ &= \sum_s d_\pi(s)\sum_a\pi(a|s)\sum_{s'}T(s, a, s')\phi(s)\big(\phi(s) - \gamma\phi(s')\big)^\top\\ &= \sum_s d_\pi(s)\phi(s)\big(\phi(s) - \gamma\sum_a\pi(a|s)\sum_{s'}T(s, a, s')\phi(s')\big)^\top\\ &= \sum_s d_\pi(s)\phi(s)\big(\phi(s) - \gamma\sum_{s'}[\mathbf{P}_\pi]_{ss'}\phi(s')\big)^\top\\ &= \Phi^\top\mathbf{D}_\pi(\mathbf{I} - \gamma\mathbf{P}_\pi)\Phi.\end{aligned} \tag{4}$$

With property (2), $\mathbf{A}_{\text{on}}$ is proved to be p.d., thus the convergence of the on-policy TD algorithm is established (Sutton [1988]).

In this paper, we are concerned with the policy evaluation problem under off-policy learning. That is, the target policy π is different from the behavior policy μ which generates the samples.

3 Instability of off-policy learning and our motivation

Importance sampling ratios, $\rho_t = \frac{\pi(a_t|s_t)}{\mu(a_t|s_t)}$, are used to correct the TD updates. Off-policy TD(0) algorithm simply multiplies the whole on-policy TD update (3) by ρ_t:

$$\begin{aligned}\theta_{t+1} \doteq & \theta_t + \rho_t \alpha_t (r_{t+1} + \gamma \theta_t^\top \phi_{t+1} - \theta_t^\top \phi_t)\phi_t \\ = & \theta_t + \alpha_t \left(\rho_t r_{t+1}\phi_t - \rho_t \phi_t (\phi_t - \gamma \phi_{t+1})^\top \theta_t\right).\end{aligned} \tag{5}$$

Let $\mathbf{A}_{\text{off}}$ denote the key matrix of the expected update (5):

$$\begin{aligned}\mathbf{A}_{\text{off}} &= \lim_{t\to\infty} \mathbb{E}_\mu \left[\rho_t \phi_t (\phi_t - \gamma \phi_{t+1})^\top\right] \\ &= \sum_s d_\mu(s) \mathbb{E}_\mu \left[\rho_t \phi_t (\phi_t - \gamma \phi_{t+1})^\top | S_t = s\right] \\ &= \sum_s d_\mu(s) \sum_a \mu(a|s) \sum_{s'} T(s,a,s') \frac{\pi(a|s)}{\mu(a|s)} \phi(s) \big(\phi(s) - \gamma \phi(s')\big)^\top \\ &= \sum_s d_\mu(s) \phi(s) \big(\phi(s) - \gamma \sum_a \pi(a|s) \sum_{s'} T(s,a,s') \phi(s')\big)^\top \\ &= \sum_s d_\mu(s) \phi(s) \big(\phi(s) - \gamma \sum_{s'} [\mathbf{P}_\pi]_{ss'} \phi(s')\big)^\top \\ &= \Phi^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi) \Phi,\end{aligned} \tag{6}$$

where $\mathbf{D}_\mu$ is the $n \times n$ diagonal matrix with distribution d_μ of the behavior policy on its diagonal. Due to the difference between behavior policy and target policy, $\mathbf{P}_\pi^\top d_\mu \neq d_\mu$. The key matrix $\mathbf{A}_{\text{off}}$ is not guaranteed to be p.d. Thus, off-policy TD(0) algorithm may be unstable (Sutton *et al.* [2016]).

Retrace algorithm with look-up value function employs a truncated IS ratios $c_t = \min(1, \rho_t)$ and guarantees convergence (Munos *et al.* [2016]). We revisit Retrace(0) with linear function approximation by Touati *et al.* [2018], where the truncated IS ratios are multiplied to the whole TD error:

$$\begin{aligned}\theta_{t+1} \doteq & \theta_t + c_t \alpha_t \left(\mathbb{E}_\pi[r_{t+1}] + \gamma \theta_t^\top \mathbb{E}_\pi[\phi_{t+1}] - \theta_t^\top \phi_t\right) \phi_t \\ = & \theta_t + \alpha_t \left(c_t \mathbb{E}_\pi[r_{t+1}]\phi_t - c_t \phi_t (\phi_t - \gamma \mathbb{E}_\pi[\phi_{t+1}])^\top \theta_t\right)\end{aligned} \tag{7}$$

Let $\mathbf{A}_{\text{Retrace(0)}}$ denote the key matrix of the expected update (13):

$$\begin{aligned}\mathbf{A}_{\text{ Retrace(0)}} &= \lim_{t\to\infty} \mathbb{E}_\mu \left[c_t \phi_t (\phi_t - \gamma \mathbb{E}_\pi[\phi_{t+1}])^\top\right] \\ &= \sum_s d_\mu(s) \mathbb{E}_\mu \left[c_t \phi(s) (\phi(s) - \gamma \mathbb{E}_\pi[\phi(s')])^\top\right] \\ &= \sum_s d_\mu(s) \phi(s) \sum_a \mu(a|s) \min\left(1, \frac{\pi(a|s)}{\mu(a|s)}\right) (\phi(s) - \gamma \mathbb{E}_\pi[\phi(s')])^\top \\ &= \sum_s d_\mu(s) \phi(s) \sum_a \min\big(\mu(a|s), \pi(a|s)\big) \Big(\phi(s) - \gamma \sum_{s'} [\mathbf{P}_\pi]_{ss'} \phi(s')\Big)^\top \\ &= \Phi^\top \mathbf{D}_\mu \mathbf{D}_c (\mathbf{I} - \gamma \mathbf{P}_\pi) \Phi\end{aligned} \tag{8}$$

where $\mathbf{D}_c$ is the $n \times n$ diagonal matrix with d_c on its diagonal, each component of d_c is

$$d_c(s) = \sum_a \min(\mu(a|s), \pi(a|s)). \tag{9}$$

Counterexample (Tsitsiklis and Van Roy [1997]; Sutton *et al.* [2016]): The $\theta \to 2\theta$ problem has only two states. From each state, there are two actions, *left* and *right*, which take the agent to

the left or right state. All rewards are zeros. The features $\Phi = (1, 2)^\top$ are assigned to the left and the right state. The behavior policy takes the equal probability to *left* or *right* in both states, i.e., $\mathrm{P}_\mu = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$. The target policy only selects action right in both states, i.e., $\mathrm{P}_\pi = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$. The state distribution of the behavior policy is $d_\mu = (0.5, 0.5)^\top$. The discount factor is $\gamma = 0.9$.

The key matrix of the off-policy TD(0) algorithm for this example is:

$$\begin{aligned} \mathbf{A}_{\text{off}} &= \Phi^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{P}_\pi)\Phi \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \times \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= -0.2. \end{aligned} \tag{10}$$

Next, let us turn to Retrace(0) algorithm. According to (9), in this example $\mathrm{D}_c = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$. Then, the key matrix of Retrace(0) algorithm for this example is:

$$\begin{aligned} \mathbf{A}_{\text{ Retrace(0)}} &= \Phi^\top \mathbf{D}_\mu \mathbf{D}_c (\mathbf{I} - \gamma \mathbf{P}_\pi)\Phi \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \times \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= -0.1. \end{aligned} \tag{11}$$

That both $\mathbf{A}_{\text{off}}$ and $\mathbf{A}_{\text{ Retrace(0)}}$ are negative means that the key matrix are not positive definite. Thus, off-policy TD(0) and Retrace(0) with linear function approximation are not stable.

A connection between Retrace and Underestimation: Define a general but straightforward implementation of the underestimation as follows:

$$\theta_{t+1} \doteq \theta_t + \alpha_t \left(\mathbb{E}_\pi[r_{t+1}] + u_t \gamma \theta_t^\top \mathbb{E}_\pi[\phi_{t+1}] - \theta_t^\top \phi_t \right) \phi_t, \tag{12}$$

where $u_t \in (0, 1]$ is a parameter to control the degree of the underestimation. Consider an underestimated version of Retrace (URetrace), i.e., $u_t = c_t$. The update rule of URetrace(0) is as follows:

$$\theta_{t+1} \doteq \theta_t + \alpha_t \left(\mathbb{E}_\pi[r_{t+1}] + c_t \gamma \theta_t^\top \mathbb{E}_\pi[\phi_{t+1}] - \theta_t^\top \phi_t \right) \phi_t. \tag{13}$$

Then, its key matrix is $\mathbf{A}_{\text{ URetrace}} = \Phi^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{D}_c \mathbf{P}_\pi)\Phi$. For the counterexample, the value is:

$$\begin{aligned} \mathbf{A}_{\text{ URetrace}} &= \Phi^\top \mathbf{D}_\mu (\mathbf{I} - \gamma \mathbf{D}_c \mathbf{P}_\pi)\Phi \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \times \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= 1.15. \end{aligned} \tag{14}$$

Nevertheless, the positive definiteness of the matrix $\mathbf{A}_{\text{ URetrace}}$ is unknown. Nonetheless, this finding encouraged us to propose the following new algorithm.

4 The Modified Retrace(0) algorithm

Importance sampling ratios, $\rho_t = \frac{\pi(s_t, a_t)}{\mu(s_t, a_t)}$, represent the "off-policyness" of the current state and action between the target policy and the behavior policy. The farther the target policy deviates, the more unstable the learning algorithm will be. The maximum of the "off-policyness", $\max_a \rho_t = \max_a \{ \frac{\pi(s_t, a_t)}{\mu(s_t, a_t)} \}$, is the key to the stability of off-policy learning algorithms.

In order to reduce the impact of the deviation of the target policy, we introduce an "off-policyness" ratio that takes the reciprocal of the above maximum degree as follows:

$$x(s_t) \doteq \frac{1}{\max_a \rho_t} = \min_a \Big\{ \frac{1}{\rho_t} \Big\} = \min_a \Big\{ \frac{\mu(a|s_t)}{\pi(a|s_t)} \Big\}. \tag{15}$$

Obviously, $x(s_t) \leq 1$[1], and $x(s_t) = 1$ only when $\forall a, \pi(a|s_t) = \mu(a|s_t)$.

To establish the connection between underestimation and convergence, "off-policyness" ratio $x(s_t)$, x_t for short, is used to control the underestimation degree in our modified Retrace(0) algorithm:

$$\begin{aligned}\theta_{t+1} \doteq & \theta_t + \alpha_t\big(\rho_t r_{t+1} + x_t\gamma\theta_t^\top \mathbb{E}_\pi[\phi_{t+1}] - \theta_t^\top\phi_t\big)\phi_t \\ = & \theta_t + \alpha_t\big(\rho_t r_{t+1}\phi_t - \phi_t(\phi_t - x_t\gamma\mathbb{E}_\pi[\phi_{t+1}])^\top\theta_t\big) \\ = & \theta_t + \alpha_t\big(\mathbf{b}_t - \mathbf{A}_t\theta_t\big),\end{aligned} \tag{16}$$

where $\mathbf{b}_t = \rho_t r_{t+1}\phi_t$, $\mathbf{A}_t = \phi_t\left(\phi_t - x_t\gamma\mathbb{E}_\pi[\phi_{t+1}]\right)^\top$.

$$\begin{aligned}\mathbf{b} = \lim_{t\to\infty}\mathbb{E}[\mathbf{b}_t] &= \lim_{t\to\infty}\mathbb{E}_\mu\left[\rho_t r_{t+1}\phi_t\right] \\ &= \sum_s d_\mu(s)\mathbb{E}_\mu\left[\rho_t r_{t+1}\phi_t | S_t = s\right] \\ &= \sum_s d_\mu(s)\sum_a \mu(a|s)\frac{\pi(a|s)}{\mu(a|s)} r_{t+1}\phi(s) \\ &= \sum_s d_\mu(s)\phi(s)\sum_a \pi(a|s)\sum_{s'} T(s,a,s')R(s,a,s') \\ &= \Phi^\top\mathbf{D}_\mu r_\pi,\end{aligned} \tag{17}$$

where r_π is expected reward vector under policy π with each component $r_\pi(s) = \sum_a\sum_{s'}\pi(a|s)R(s,a,s')$.

$$\begin{aligned}\mathbf{A} = \lim_{t\to\infty}\mathbb{E}\left[\mathbf{A}_t\right] &= \lim_{t\to\infty}\mathbb{E}_\mu\left[\phi_t(\phi_t - x_t\gamma\mathbb{E}_\pi[\phi_{t+1}])^\top\right] \\ &= \sum_s d_\mu(s)\mathbb{E}_\mu\left[\phi(s)(\phi(s) - x_t\gamma\mathbb{E}_\pi[\phi(s')])^\top\right] \\ &= \sum_s d_\mu(s)\phi(s)\big(\phi(s) - \mathbb{E}_\mu[x_t\gamma\mathbb{E}_\pi[\phi(s')]]\big)^\top \\ &= \sum_s d_\mu(s)\phi(s)\Big(\phi(s) - \gamma\sum_a \mu(a|s)\min_b\Big\{\frac{\mu(b|s)}{\pi(b|s)}\Big\}\mathbb{E}_\pi[\phi(s')]\Big)^\top \\ &= \sum_s d_\mu(s)\phi(s)\Big(\phi(s) - \gamma\min_b\Big\{\frac{\mu(b|s)}{\pi(b|s)}\Big\}\sum_{s'}[\mathbf{P}_\pi]_{ss'}\phi(s')\Big)^\top \\ &= \Phi^\top\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)\Phi,\end{aligned} \tag{18}$$

where $\mathbf{D}_x$ is the $n \times n$ diagonal matrix with d_x on its diagonal, each component of d_x is $d_x(s) = \min_b\Big\{\frac{\mu(b|s)}{\pi(b|s)}\Big\}$. The modified Retrace algorithm solve a TD fixed point as:

$$\mathbb{E}_\mu\left[(\rho r + x\gamma\theta^\top\mathbb{E}_\pi[\phi'] - \theta^\top\phi)\phi\right] = 0. \tag{19}$$

Note that according to Scherrer [2010], (19) is TD fixed point due to the projection direction, $\mathbf{D}\Phi$. But it is neither the TD fixed point of the behavior policy, nor the TD fixed point of the target policy. It is actually an underestimation version of the TD fixed point of the target policy because there is a factor x in the front of the target state value. The quality of these fixed points will be discussed in the experimental section.

For the counterexample, $d_x = (0.5, 0.5)^\top$, the value of the new matrix is:

$$\begin{aligned}\mathbf{A} &= \Phi^\top\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)\Phi \\ &= \begin{bmatrix} 1 & 2 \end{bmatrix} \times \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \times \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}\right)\begin{bmatrix} 1 \\ 2 \end{bmatrix} \\ &= 1.15.\end{aligned} \tag{20}$$

[1]Note that $\sum_a \mu(a|s_t) = 1, \sum_a \pi(a|s_t) = 1$.

Extension to Q-learning with linear function approximation: the action value function is defined as $Q_\theta(s,a) = \theta^\top \phi(s,a)$. Consider the ϵ-greedy policy as the behavior policy. For each state s, the probability of selecting an action under the behavior policy is

$$\mu(a|s) = \begin{cases} 1-\epsilon+\frac{\epsilon}{|A_s|}, & \text{if } a \text{ is the greedy action;} \\ \frac{\epsilon}{|A_s|}, & \text{otherwise.} \end{cases} \tag{21}$$

the probability of selecting an action under the target policy is

$$\pi(a|s) = \begin{cases} 1, & \text{if } a \text{ is the greedy action;} \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

Then, the ratios are

$$\frac{\mu(a|s)}{\pi(a|s)} = \begin{cases} 1-\epsilon+\frac{\epsilon}{|A_s|}, & \text{if } a \text{ is the greedy action;} \\ +\infty, & \text{otherwise.} \end{cases} \tag{23}$$

Therefore, the "off-policyness" ratio is

$$x(s) = \min_a \Big\{ \frac{\mu(a|s)}{\pi(a|s)} \Big\} = \mu(a^*|s), \tag{24}$$

where $a^* = \arg\max_a Q_\theta(s,a)$. Thus, the update rule of the modified Q-learning with ϵ-greedy policy and linear function approximation is as follows:

$$\theta_{t+1} \leftarrow \theta_t + \alpha_t \big(r_{t+1} + \mu(a^*|s')\gamma \max_b Q_{\theta_t}(s',b) - Q_{\theta_t}(s,a) \big) \phi(s,a). \tag{25}$$

Note that ϵ keeps decreasing until zero in the training phase. The "off-policyness" ratio keeps increasing: $\lim_{\epsilon\to 0} x(s) = \lim_{\epsilon\to 0} 1-\epsilon+\frac{\epsilon}{|A_s|} = 1$. Thus, when the modified Q-learning (25) converges, the value function converges to the TD fixed point of the target policy.

5 Convergence

The purpose of this section is to establish that the MRetrace(0) algorithm converges with probability one under standard assumptions when $\{\phi_t, r_t, \phi'_t\}$ is obtained by the off-policy subsampling process (Sutton *et al.* [2008]).

Assumption 1. *The Markov chain (s_t) is aperiodic and irreducible, so that* $\lim_{t\to\infty} \mathbb{P}(s_t = s'|s_0 = s) = d_\mu(s')$ *exists and is unique. Let s be a state randomly drawn from d_μ, and let s' be a state obtained by following π for one time step in the MDP from s. Let the behavior policy μ select all actions of the target policy π with positive probability in every state, and the target policy is deterministic. Further, let $r(s,s')$ be the reward incurred.*

This assumption implies that the state distribution vector d_μ of the behavior policy μ is the fixed point of

$$d_\mu = \mathbf{P}_\mu^\top d_\mu, \tag{26}$$

where element of matrix $\mathbf{P}_\mu$ is as follows:

$$[\mathbf{P}_\mu]_{ss'} = \sum \mu(a|s) T(s,a,s'). \tag{27}$$

Assumption 2. *$\{\phi_t, r_t, \phi'_t\}$ is such that $\mathbb{E}[||\phi_t||^2|s_{t_1}]$, $\mathbb{E}[r_t^2|s_{t_1}]$, $\mathbb{E}[||\phi'_t||^2|s_{t_1}]$ are uniformly bounded.*

Assumption 3. *Step-size sequence α_t satisfies $\alpha_t \in (0,1]$, $\sum_{t=0}^\infty \alpha_t = \infty$, and $\sum_{t=0}^\infty \alpha_t^2 < \infty$.*

Theorem 1. *(Convergence of MRetrace(0) with an off-policy sub-sampled process). Assume Assumption 1, 2, and 3. Let the parameter θ_t be updated by iteration (16). Let $\boldsymbol{A} = \mathbb{E}_\mu \left[\phi_t (\phi_t - x_t \gamma \mathbb{E}_\pi[\phi_{t+1}])^\top \right]$, $\boldsymbol{b} = \mathbb{E}_\mu[\rho_t r_t \phi_t]$. Assume that matrix $\boldsymbol{A}$ is non-singular. Then the parameter vector θ_t converges with probability one to the TD fixed-point (19).*

Proof. The proof follows from the procedures of Sutton *et al.* [2008, 2009] for GTD and GTD2, which are based on the ordinary-differential-equation (ODE) approach (Borkar and Meyn [2000]). First, $\mathbf{A}$ and $\mathbf{b}$ are well-defined according to Assumption 1 and 2.

Now we apply Theorem 2.2 of Borkar and Meyn [2000]. We write $\theta_{t+1} = \theta_t + \alpha_t(-\mathbf{A}\theta_t + \mathbf{b} + (\mathbf{A} - \mathbf{A}_{t+1})\theta_t + (\mathbf{b}_{t+1} - \mathbf{b})) = \theta_t + \alpha_t(h(\theta_t) + M_{t+1})$, where $h(\theta) = \mathbf{b} - \mathbf{A}\theta$ and $M_{t+1} = (\mathbf{A} - \mathbf{A}_{t+1})\theta_t + \mathbf{b}_{t+1} - \mathbf{b}$. Let $\mathcal{F}_t = \sigma(\theta_1, M_1, \ldots, \theta_{t-1}, M_t)$. Theorem 2.2 requires the verification of the following conditions: (i) The function h is Lipschitz and $h_\infty(\theta) = \lim_{r\to\infty} h(r\theta)/r$ is well-defined for every $\theta \in \mathbb{R}^m$; (ii-a) The sequence $(M_t, \mathcal{F}_t)$ is a martingale difference sequence, and (ii-b) for some $C_0 > 0$, $\mathbb{E}[||M_{t+1}||^2|\mathcal{F}_t] \leq C_0(1 + ||\theta_t||^2)$ holds for any initial parameter vector θ_1; (iii) The sequence α_t satisfies $0 < \alpha_t \leq 1$, $\sum_{t=0}^{\infty} \alpha_t = \infty$, and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$; (iv) The ODE $\dot{\theta} = h_\infty(\theta)$ has the origin as a globally asymptotically stable equilibrium; and (v) The ODE $\dot{\theta} = h(\theta)$ has a unique globally asymptotically stable equilibrium.

Clearly, $h(\theta)$ is Lipschitz with coefficient $||\mathbf{A}||$ and $h_\infty(\theta) = -\mathbf{A}\theta$. By construction, $(M_t, \mathcal{F}_t)$ satisfies $\mathbb{E}[M_{t+1}|\mathcal{F}_t] = 0$ and $M_t \in \mathcal{F}_t$, i.e., it is a martingale difference sequence. Condition (ii-b) can be shown to hold by a simple application of the triangle inequality and the boundedness of the second moments of $\{\phi_t, r_t, \phi_t'\}_t$. Condition (iii) is satisfied by our conditions on the step-size sequences α_t.

For the last two conditions, we begin by showing that the matrix $\mathbf{A} = \mathbb{E}_\mu[\phi_t(\phi_t - x_t\gamma\mathbb{E}_\pi[\phi_{t+1}])^\top] = \Phi^\top\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)\Phi$ is p.d. Note that $\mathbf{A}$ consists of $\Phi^\top$ and Φ wrapped around an $n \times n$ matrix $\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)$. According to assumption that the matrix $\mathbf{A}$ is non-singular, then, $\mathbf{A}$ is p.d. whenever the key matrix $\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)$ is p.d.

Based on two theorems showed by Sutton [1988]; Sutton *et al.* [2016], positive definiteness of the key matrix is assured if all of its columns and rows sum to positive numbers. One theorem is that any matrix $\mathbf{M}$ is p.d. if and only if the symmetric matrix $\mathbf{S} = \mathbf{M} + \mathbf{M}^\top$ is p.d. Another theorem is that any symmetric real matrix $\mathbf{S}$ is p.d. if the absolute values of its diagonal entries are greater than the sum of the absolute values of the corresponding off-diagonal entries. For the key matrix, $\mathbf{M} = \mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)$, the diagonal entries are positive and the off-diagonal entries are negative, so all we have to show is that all components of both ($\mathbf{M1}$) and ($\mathbf{1}^\top\mathbf{M}$) are positive, where $\mathbf{1}$ is the column vector with all components equal to 1. They can be verified as follows:

$$\mathbf{M1} = \mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi)\mathbf{1} = \mathbf{D}_\mu(\mathbf{1} - \gamma\mathbf{D}_x\mathbf{P}_\pi\mathbf{1}) = \mathbf{D}_\mu(\mathbf{1} - \gamma\mathbf{D}_x\mathbf{1}) = \mathbf{D}_\mu(\mathbf{1} - \gamma d_x) \tag{28}$$

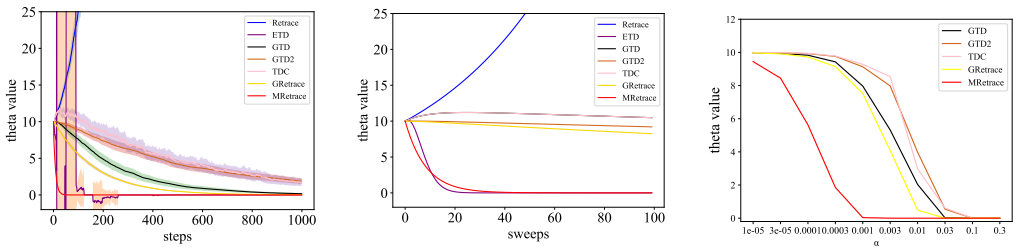
Each component of $\mathbf{M1}$ is $[\mathbf{D}_\mu(\mathbf{1} - \gamma d_x)](s) = d_\mu(s)(1 - \gamma\min_b\{\frac{\mu(b|s)}{\pi(b|s)}\}) \geq d_\mu(s)(1 - \gamma) > 0$.

$$\begin{aligned}
[\mathbf{D}_x\mathbf{P}_\pi]_{ij} = \min_b\Big\{\frac{\mu(b|i)}{\pi(b|i)}\Big\}\sum_a \pi(a|i)T(i,a,j) &= \sum_a \pi(a|i)\min_b\Big\{\frac{\mu(b|i)}{\pi(b|i)}\Big\}T(i,a,j) \\
&\leq \sum_a \pi(a|i)\frac{\mu(a|i)}{\pi(a|i)}T(i,a,j) \\
&= \sum_a \mu(a|i)T(i,a,j) \\
&= [\mathbf{P}_\mu]_{ij}.
\end{aligned} \tag{29}$$

$$\begin{aligned}
\mathbf{1}^\top\mathbf{M} = \mathbf{1}^\top\mathbf{D}_\mu(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi) = d_\mu^\top(\mathbf{I} - \gamma\mathbf{D}_x\mathbf{P}_\pi) &= d_\mu^\top - \gamma d_\mu^\top\mathbf{D}_x\mathbf{P}_\pi \\
&\geq d_\mu^\top - \gamma d_\mu^\top\mathbf{P}_\mu \\
&= d_\mu^\top - \gamma d_\mu^\top \\
&= (1 - \gamma)d_\mu^\top
\end{aligned} \tag{30}$$

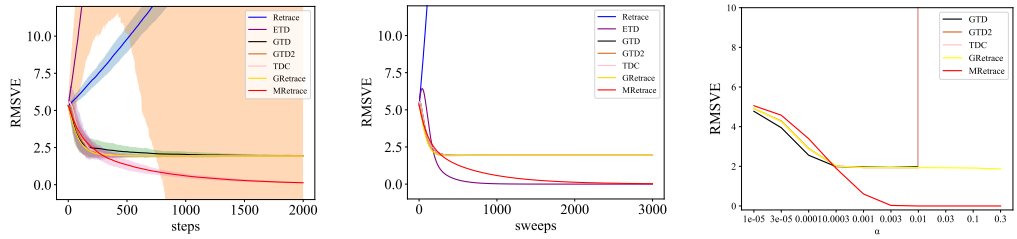
Each component of the vector $\mathbf{1}^\top\mathbf{M}$ is $[(1 - \gamma)d_\mu](s) = (1 - \gamma)d_\mu(s) > 0$. The row sums and the column sums are all positive. Thus, (iv) is satisfied.

Finally, for the ODE $\dot{\theta} = h(\theta)$, note that $\theta^* = A^{-1}b$ is the unique asymptotically stable equilibrium with $\overline{V}(\theta) = \frac{1}{2}||-A\theta + b||^2$ as its associated strict Liapunov function. The claim now follows. □



(a) Stochastic updates. (b) Sweep updates. (c) Sensitive testing.

Figure 1: Comparisons of various temporal difference updates in 2-states counterexample.



(a) Stochastic updates. (b) Sweep updates. (c) Sensitive testing.

Figure 2: Comparisons of various temporal difference updates in Baird's counterexample.

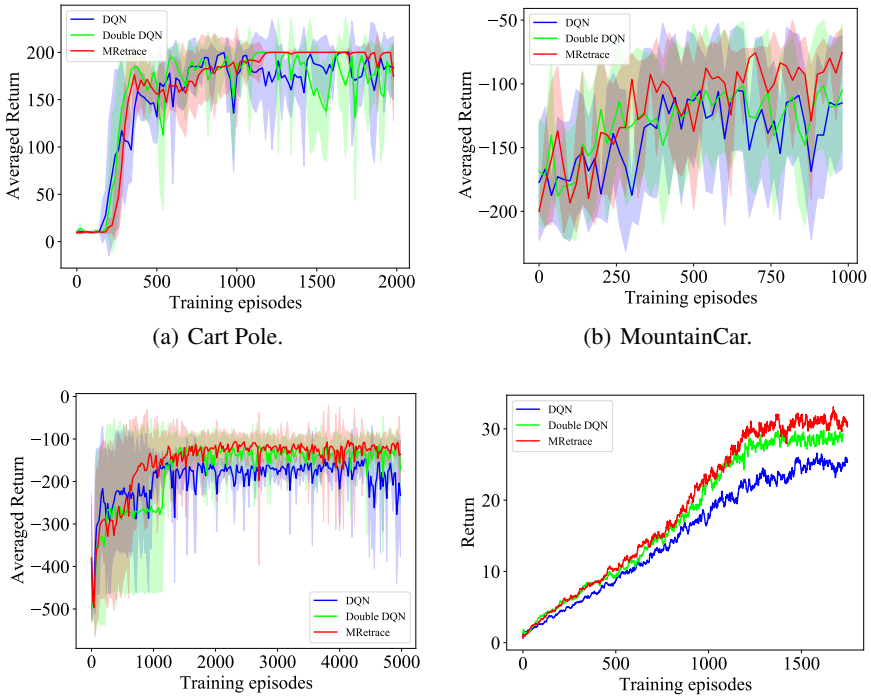
6 Experimental studies and discussions

In experiments, we care about two points about the proposed MRetrace algorithm: (1) Whether it converges experimentally, although it does converge in theory? (2) What is the quality of the TD fixed point it solves? We adopted two sets of experiments, i.e. counterexamples to test the stability and control tasks to test the quality.

In the first set of experiments, we used the 2-states counterexample and Baird's counterexample, implemented two update styles including stochastic updates and sweep updates, and finished parameter sensitivity test for converged algorithms. Compared algorithms include Retrace, ETD, GTD, GTD2, TDC, and GRetrace. Each algorithm was run 100 times independently.

Table 2: Hyperparameters in the control tasks.

Hyperparameter	Cart pole	Mountain car	Acrobot	Breakout
Training Episodes	2000	1000	5000	17000
Learning Rate	1e-3	1e-3	1e-5	1e-5
Learning Rate Decay	None	None	None	0.9995
Optimizer	Adam	Adam	Adam	Adam
Discount Factor γ	0.99	0.99	0.99	0.99
Max Epsilon	0.1	0.01	0.01	1
Min Epsilon	0.1	0.01	0.01	0.05
Epsilon Decay	None	None	None	0.99995
Minibatch Size	32	32	32	32
Replay Memory Size	10000	50000	1000000	6000
Network	MLP	MLP	MLP	CNN
CNN Output Feature Size	None	None	None	128
MLP Hidden Size	16×16	16×16	16×16	None
Target Network Update Frequency	500	500	500	500



(a) Cart Pole. (b) MountainCar.

(c) Acrobot. (d) Breakout.

Figure 3: Comparisons of DQN, Double DQN and MRetrace in control tasks.

Algorithms' learning curves are shown in Figure 1 and Figure 2, where the theta value is equal to the root of mean squared value error (RMSVE) since there is only one scalar parameter in the 2-state counterexample. We can see that (i) Retrace diverges in all cases. (ii) Expected ETD converges to zero the fastest. On the other hand, ETD converges with a high variance at the beginning in the 2-state counterexample, and diverges in Baird's counterexample which is consistent with results of computational experiments about ETD introduced in Page 282 of Sutton and Barto [2018]. (iii) MRetrace converges to zero relatively quickly in all cases. (iv) MRetrace performs best in parameter sensitivity tests.

This result is not surprising. MRetrace(0) updates only one parameter vector, and thus avoids the slow convergence problem compared to the gradient-based approaches.

In the second set of experiments, we used several typical control tasks provided by Gym: CartPole, MountainCar, Acrobot, and Breakout. Compared algorithms include DQN and Double DQN. Each algorithm employs the $\epsilon-$greedy strategy to explore. Hyperparameters of each task are shown in Table 2. Algorithms' learning curves are shown in Figure 3. We can see that MRetrace outperforms DQN and Double DQN in all tasks.

Note that if ϵ is set too high, MRetrace algorithm does not perform the best. This is because that the TD fixed point (19) solved by MRetrace is an underestimation version of the TD fixed point of the target policy. The larger ϵ is, the greater the degree of underestimation, and the farther away from the fixed point of the target policy. Therefore, MRetrace performs best when ϵ is set to a relatively small value, thanks to its close proximity to the fixed point of the target policy and fast convergence of our algorithm.

This provides more insight into reinforcement learning algorithms which suffer from a trade-off between bias and variance of the fixed point. Our algorithm and experimental results tend to favor stability factors, i.e., smaller variance, and proper consideration of bias.

References

Oron Anschel, Nir Baram, and Nahum Shimkin. Averaged-DQN: Variance reduction and stabilization for deep reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pages 176–185, 2017.

Kavosh Asadi and Michael L Littman. An alternative softmax operator for reinforcement learning. In *International Conference on Machine Learning*, pages 243–252, 2017.

Leemon Baird et al. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Machine Learning*, pages 30–37, 1995.

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Leah M Hackman. *Faster Gradient-TD Algorithms*. PhD thesis, University of Alberta, 2012.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1372–1383, 2017.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2094–2100, 2016.

Hado V Hasselt. Double q-learning. In *Advances in neural information processing systems*, pages 2613–2621, 2010.

Seungchan Kim, Kavosh Asadi, Michael Littman, and George Konidaris. Removing the target network from deep q-networks with the mellowmax operator. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 2060–2062, 2019.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 504–513, 2015.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Proximal gradient temporal difference learning algorithms. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4195–4199, 2016.

Bo Liu, Ian Gemp, Mohammad Ghavamzadeh, Ji Liu, Sridhar Mahadevan, and Marek Petrik. Proximal gradient temporal difference learning: Stable reinforcement learning with polynomial sample complexity. *Journal of Artificial Intelligence Research*, 63:461–494, 2018.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2016.

Yangchen Pan, Adam White, and Martha White. Accelerated gradient temporal difference learning. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, pages 2464–2470, 2017.

Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning*, pages 417–424, 2001.

Bruno Scherrer. Should one compute the temporal difference fix point or minimize the bellman residual? the unified oblique projection view. In *Proceedings of the 27th International Conference on Machine Learning*, pages 959–966, 2010.

Zhao Song, Ron Parr, and Lawrence Carin. Revisiting the softmax bellman operator: New benefits and new perspective. In *International Conference on Machine Learning*, pages 5916–5925. PMLR, 2019.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

Richard S Sutton, Hamid R Maei, and Csaba Szepesvári. A convergent $O(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 1609–1616, 2008.

R.S. Sutton, H.R. Maei, D. Precup, S. Bhatnagar, D. Silver, C. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, pages 993–1000, 2009.

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

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

Ahmed Touati, Pierre-Luc Bacon, Doina Precup, and Pascal Vincent. Convergent tree backup and retrace with function approximation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4955–4964, 2018.

John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081, 1997.

Shangtong Zhang and Shimon Whiteson. Truncated emphatic temporal difference methods for prediction and control. *arXiv:2108.05338*, 2021.

Shangtong Zhang, Bo Liu, Hengshuai Yao, and Shimon Whiteson. Provably convergent two-timescale off-policy actor-critic with function approximation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11204–11213, 2020.

Shangtong Zhang, Yi Wan, Richard S Sutton, and Shimon Whiteson. Average-reward off-policy policy evaluation with function approximation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12578–12588, 2021.