Expected Sarsa(λ) with Control Variate for Variance Reduction

Long Yang, Yu Zhang, Jun Wen, Qian Zheng, Pengfei Li, Gang Pan

Department of Computer Science, Zhejiang University

{yanglong,hzzhangyu,junwen,qianzheng,pfl,gpan}@zju.edu.cn

Abstract

Off-policy learning is powerful for reinforcement learning. However, the high variance of off-policy evaluation is a critical challenge, which causes off-policy learning falls into an uncontrolled instability. In this paper, for reducing the variance, we introduce control variate technique to Expected Sarsa(λ) and propose a tabular ES(λ)-CV algorithm. We prove that if a proper estimator of value function reaches, the proposed ES(λ)-CV enjoys a lower variance than Expected Sarsa(λ). Furthermore, to extend ES(λ)-CV to be a convergent algorithm with linear function approximation, we propose the GES(λ) algorithm under the convex-concave saddle-point formulation. We prove that the convergence rate of GES(λ) achieves $\mathcal{O}(1/T)$, which matches or outperforms lots of state-of-art gradient-based algorithms, but we use a more relaxed condition. Numerical experiments show that the proposed algorithm performs better with lower variance than several state-of-art gradient-based TD learning algorithms: GQ(λ), GTB(λ) and ABQ(ζ).

Introduction

Off-policy learning is powerful for reinforcement learning due to it learns the target policy from the data generated by another policy (Sutton and Barto 1998). However, suffering high variance is a critical challenge for off-policy learning (A. Tamar and Mannor. 2016), which roots in the discrepancy of distribution between target policy and behavior policy. The resources of high variance of off-policy learning can be divided into two parts, **(I)** one is tabular case which has to do with the target of the update, **(II)** one is with function approximation which has to do with the distribution of the update (Sutton and Barto 2018).

In this paper, we mainly focus on the variance reduce technique to an important off-policy algorithm: Expected Sarsa(λ). We introduce control variate to Expected Sarsa(λ) and propose Expected Sarsa(λ) with control variate (ES(λ)-CV) for the tabular case. The control variate method is one of the most effective variance reduction techniques in statistical inference (Rubinstein and Kroese 2016). Control variate is an additional term that has zero expectation, which implies introducing control variate does not

change the expectation of update. Thus, learning with control variate does not introduce any biases, but it is potential to enjoy much lower variance (Thomas and Brunskill 2016; De Asis and Sutton 2018; Liu et al. 2018). Sutton and Barto (2018) (section 12.9) firstly introduces control variate to Expected Sarsa(λ), but their analysis is limited in linear function approximation. Later, De Asis and Sutton (2018) further introduce control variate to multi-step TD learning, but it constrains on off-line learning (which is extremely expensive for training).

Despite being easy to implement, competitive to the state of the art methods, and being used in practice, in RL, the TD learning with control variate technique lacks a robust theoretical analysis. In this paper, we focus on the theoretical analysis of ES(λ)-CV. We prove that the tabular ES(λ)-CV converges at an exponential fast for off-policy evaluation without biases. Furthermore, we analyze all the random sources lead to the variance of ES(λ)-CV, and we prove that if a proper estimator of value function reaches, ES(λ)-CV enjoys a lower variance than Expected Sarsa(λ).

Furthermore, we show the variance reduction way presented by (Sutton and Barto 2018) (section 12.9) to extend ES(λ)-CV with function approximation is unstable. Although this instability has been realized by Sutton and Barto (2018), it is only an intuitive guess inspired previous works (Maei 2011; Mahmood 2017a). In this paper, we provide a simple but rigorous theoretical analysis to illustrate the instability appears in (Sutton and Barto 2018). We also demonstrate this instability by a typical example.

To extend the ES(λ)-CV with function approximation be a convergent and stable algorithm, we propose GES(λ) algorithm under the the convex-concave saddle-point formulation (Liu et al. 2015). We prove the convergence rate of GES(λ) achieves $\mathcal{O}(1/T)$, where T is the number of iterations. Our $\mathcal{O}(1/T)$ matches or outperforms extensive state-of-art works (Nathaniel and Prashanth 2015; Liu et al. 2015; Wang et al. 2017; Dalal et al. 2018a; Dalal et al. 2018b; Touati et al. 2018), with a more relaxed condition than theirs. Besides, we prove the results of convergence rate without the assumption that the objective is strongly convex in the primal space and strongly concave in the dual space (Balamurugan and Bach 2016).

Finally, we conduct numerical experiments to show that the proposed algorithm is stable and converges faster with lower variance than lots of state-of-art gradient-based TD learning algorithms: GQ(λ) (Maei and Sutton 2010), GTB (λ) (Touati et al. 2018), and ABQ (ζ) (Mahmood, Yu, and Sutton 2017b).

Contributions

- We introduce control variate technique to Expected Sarsa(λ) and propose a tabular ES(λ)-CV algorithm. We prove that if a proper estimator of value function reaches, the proposed ES(λ)-CV enjoys a lower variance than Expected Sarsa(λ).

- We propose the GES(λ), which extends ES(λ)-CV to be a convergent algorithm with linear function approximation. We prove that the convergence rate of GES(λ) achieves $\mathcal{O}(1/T)$, which matches or outperforms lots of state-of-art gradient-based algorithms, but we use a more relaxed condition.

Preliminary and Some Notations

In this section, we introduce some necessary notations about reinforcement learning, temporal difference learning and λ-return. For the limitation of space, we more discussions about λ-return in Appendix A and B.

Reinforcement Learning The reinforcement learning (RL) is often formalized as *Markov decision processes* (MDP) (Sutton and Barto 1998) which considers 5-tuples form $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. $\mathcal{S}$ is the set contains all states, $\mathcal{A}$ is the set contains all actions. $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, $P_{ss'}^a = \mathcal{P}(S_t = s'|S_{t-1} = s, A_{t-1} = a)$ is the probability for the state transition from s to s' under taking the action a. $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^1$, $\mathcal{R}_s^a = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$. $\gamma \in (0, 1)$ is the discount factor.

A *policy* is a probability distribution on $\mathcal{S} \times \mathcal{A}$. *Target policy* π is the policy will be learned and *behavior policy* μ is used to generate behavior. $\tau = \{S_t, A_t, R_{t+1}\}_{t \geq 0}$ denotes a *trajectory*, where $A_t \sim \mu(\cdot|S_t)$ and $S_{t+1} \sim \mathcal{P}(\cdot|S_t, A_t)$. For a given policy π, its *state-action value function* $q^\pi(s, a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$, *state value function* $v^\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$, where $G_t = \sum_{k=0}^{\infty} \gamma^k R_{k+t+1}$ and $\mathbb{E}_\pi[\cdot|\cdot]$ denotes an conditional expectation on all actions which be selected according to π. It is known that $q^\pi(s, a)$ is the unique fixed point (Bertsekas 2012) of *Bellman operator* $\mathcal{B}^\pi$,

$$\mathcal{B}^\pi q^\pi = q^\pi, \tag{1}$$

which is known as *Bellman equation*, where

$$\mathcal{B}^\pi : q \mapsto R + \gamma P^\pi q,$$

$P^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ and $R \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, the corresponding elements of P^π and R are:

$$P_{ss'}^\pi = \sum_{a \in \mathcal{A}} \pi(a|s) P_{ss'}^a, R(s, a) = \mathcal{R}_s^a.$$

TD Learning Temporal difference (TD) learning (Sutton 1988) is one of the most important methods to solve model-free RL (in which, we cannot get $\mathcal{P}$). For the trajectory τ,

TD learning is defined as, $\forall\, t \geq 0$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t \delta_t, \tag{2}$$

where $Q(\cdot, \cdot)$ is an estimate of q^π, α_t is step-size and δ_t is TD error. Let $Q_t \overset{\text{def}}{=} Q(S_t, A_t)$, if δ_t is

$$\delta_t^{\text{S}} \overset{\text{def}}{=} R_{t+1} + \gamma Q_{t+1} - Q_t,$$

above update (2) is Sarsa algorithm (Rummery and Niranjan 1994). If δ_t is

$$\delta_t^{\text{ES}} = R_{t+1} + \mathbb{E}_\pi[Q(S_{t+1}, \cdot)] - Q_t, \tag{3}$$

update (2) is Expected Sarsa (Van Seijen et al. 2009), where $\mathbb{E}_\pi[Q(S_{t+1}, \cdot)] = \sum_{a \in \mathcal{A}} \pi(a|S_{t+1}) Q(S_{t+1}, a)$. If π is reduced to greedy policy, then Expected Sarsa reduces to Q-learning (Watkins 1989).

Expected Sarsa(λ) The standard *forward view* of λ-return (Sutton and Barto 1998) of on-policy Expected Sarsa is defined as follows,

$$G_t^{\lambda, \text{ES}} = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{t+n}, \tag{4}$$

where $G_t^{t+n} = \sum_{i=0}^{n-1} \gamma^i R_{t+i+1} + \gamma^n \bar{Q}_{t+n}$ is *n-step return* of Expected Sarsa, and $\bar{Q}_{t+n} = \mathbb{E}_\pi[Q(S_{t+n}, \cdot)]$. We can write $G_t^{\lambda, \text{ES}}$ recursively as follows (the detail is provided in Appendix A),

$$G_t^{\lambda, \text{ES}} = R_{t+1} + \gamma[(1 - \lambda)\bar{Q}_{t+1} + \lambda G_{t+1}^{\lambda, \text{ES}}]. \tag{5}$$

Now, we introduce an unbiased [1] recursive λ-return of Expected Sarsa for off-policy learning,

$$G_t^{\lambda\rho, \text{ES}} = R_{t+1} + \gamma[(1 - \lambda)\bar{Q}_{t+1} + \lambda \rho_{t+1} G_{t+1}^{\lambda\rho, \text{ES}}], \tag{6}$$

where $\rho_{t+1} = \pi(A_{t+1}|S_{t+1})/\mu(A_{t+1}|S_{t+1})$ is importance sampling. Eq.(6) firstly appears in (Maei and Sutton 2010; Maei 2011), but in which it is limited in function approximation. We develop (6) to be a general version which is conducive to the theoretical analysis of the following paragraph. The following Proposition 1 illustrates that $G_t^{\lambda\rho, \text{ES}}$ (6) is an unbiased estimate of q^π.

Proposition 1. *Let μ and π be the behavior and target policy, respectively. For the λ-return (6), we have*

$$\mathbb{E}_\mu[G_t^{\lambda\rho, \text{ES}}|(S_t, A_t) = (s, a)] = q^\pi(s, a).$$

For the limitation of space, more discussions about λ-return of Sarsa, Eq.(5)-(6), and the proof of Proposition 1 are provided in Appendix A and B.

---

[1] How to define the λ-return of Expected Sarsa for off-policy learning? Can we follow the way of (4) straightforwardly? Unfortunately, for the off-policy, the above idea cannot converge to q^π. In fact, n-step return of Expected Sarsa is sampled according to

$$R_{t:t+n} = \sum_{t=0}^{n} \gamma^t (P^\mu)^t R_{t+1} + \gamma^{n+1}(P^\mu)^n P^\pi Q.$$

Then according to (4), we define the λ-return of Expected Sarsa as follows,

$$(1 - \lambda) \sum_{n=0}^{\infty} \lambda^n R_{t:t+n} = ((1 - \lambda)\mathcal{B}^\pi + \lambda \mathcal{B}^\mu)Q,$$

which converges to $(1 - \lambda)q^\pi + \lambda q^\mu \neq q^\pi$. This is the fixed point of $(1 - \lambda)\mathcal{B}^\pi + \lambda \mathcal{B}^\mu \neq \mathcal{B}^\pi$ and it is a biased estimate of q^π.

Expected Sarsa(λ) with Control Variate

In this section, we firstly define Expected Sarsa(λ) with control variate (we use ES(λ)-CV for short). Then, prove its linear convergence rate of ES(λ)-CV for policy evaluation. Finally, we analyze the variance of ES(λ)-CV.

ES(λ)-CV Algorithm

We define Expected Sarsa(λ) with control variate $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ as follows

$$\widetilde{G}_t^{\lambda\rho,\text{ES}} = R_{t+1} + \gamma\Big[(1-\lambda)\bar{Q}_{t+1} + \lambda(\rho_{t+1}\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}}$$
$$+ \underbrace{\bar{Q}_{t+1} - \rho_{t+1}Q_{t+1}}_{\text{control variate}})\Big], \qquad (7)$$

where the additional term $\bar{Q}_{t+1} - \rho_{t+1}Q_{t+1}$ is called control variate (CV). The following fact

$$\mathbb{E}_\mu[\bar{Q}_{t+1} - \rho_{t+1}Q_{t+1}] = 0$$

implies that $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ (7) extends $G_t^{\lambda\rho,\text{ES}}$ (6) without introducing biases.

Theorem 1 (Forward View of ES(λ)-CV). *Let $\rho_{t:k} = \prod_{i=t}^{k}\rho_i$ denote the cumulated importance sampling from time t to k, and we use $\rho_{t+1:t} = 1$ for convention. The recursive λ-return in Eq.(7) is equivalent to the following forward view: let δ_l^{ES} be the TD error defined in (3), $G_t^t = Q_t$, $G_t^{t+n} = R_{t+1} + \gamma(\rho_{t+1}G_{t+1}^{t+n} + \bar{Q}_{t+1} - \rho_{t+1}Q_{t+1})$*

$$\widetilde{G}_t^{\lambda\rho,\text{ES}} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n}$$
$$= Q_t + \sum_{l=t}^{\infty}(\gamma\lambda)^{l-t}\delta_l^{\text{ES}}\rho_{t+1:l}. \qquad (8)$$

Proof. See Appendix C. $\qquad\square$

Remark 1. *Eq.(8) illustrates that for a given finite horizon trajectory $\{S_t, A_t, R_{t+1}\}_{t=0}^{h}$, the total update (7) reaches*

$$\sum_{t=0}^{h}(\gamma\lambda)^t\delta_t^{ES}\rho_{1:t}, \qquad (9)$$

which is off-line update of ES(λ)-CV.

Policy Evaluation

For policy evaluation, our goal is to estimate q^π according to the trajectory collection $\mathcal{T} = \{\tau_k\}_{k\in\mathbb{N}}$, where $\tau_k = \{S_t, A_t, R_{t+1}\}_{t\geq 0} \sim \mu$, S_t, A_t, and R_{t+1} are dependent on the index k strictly, and we omit coefficient k to tight the expression without ambiguity.

The following λ-operator $\mathcal{B}_\lambda^\pi$ is a high level view of ES(λ)-CV (8), and it is helpful for us to introduce policy evaluation algorithm. $\forall q \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, t \geq 0$

$$\mathcal{B}_\lambda^\pi q \stackrel{\text{def}}{\mapsto} q + \mathbb{E}_\mu[\sum_{l=t}^{\infty}(\lambda\gamma)^{l-t}\delta_l^{\text{ES}}\rho_{t+1:l}] \qquad (10)$$

$$\stackrel{(a)}{=} q + (I - \lambda\gamma P^\pi)^{-1}(\mathcal{B}^\pi q - q), \qquad (11)$$

where $\mathcal{B}^\pi$ is defined in Eq.(1). We provide the equivalence (a) in Appendix D.

Theorem 2 (Policy Evaluation). *For any initial Q_0, consider the trajectory $\mathcal{T}$ generated by μ, and the following Q_k is generated according to the k-th trajectory $\tau_k \in \mathcal{T}$, $k \geq 1$,*

$$Q_{k+1} = \mathcal{B}_\lambda^\pi Q_k. \qquad (12)$$

By iterating over k trajectories, the upper-error of policy evaluation is bounded by

$$\|Q_k - q^\pi\| \leq \big(\frac{\gamma - \lambda\gamma}{1 - \lambda\gamma}\big)^k\|Q_0 - q^\pi\|. \qquad (13)$$

Proof. See Appendix E. $\qquad\square$

Remark 2. *The forward view (off-line update) of ES(λ)-CV (8) can be seen as sampled according to $Q_{t+1} = \mathcal{B}_\lambda^\pi Q_t$. For any $\gamma \in (0,1), \lambda \in [0,1]$, then $\frac{\gamma-\lambda\gamma}{1-\lambda\gamma} \in (0,1)$, thus Eq.(13) implies (8) converges to q^π at a linear convergence rate.*

Variance Analysis

Theorem 3 (Variance Analysis of ES(λ)-CV). *Consider a single trajectory τ_k with ffinite horizon $H + 1$, let $S_t = s, A_t = a, S_{t+1} = s', A_{t+1} = a', \mathbb{Var}[\widetilde{G}_{H+1}^{\lambda\rho,\text{ES}}] = 0$. The variance of $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ is given recursively as follows,*

$$\mathbb{Var}[\widetilde{G}_t^{\lambda\rho,\text{ES}}] = \mathbb{Var}[R_{t+1} + \gamma\bar{Q}_{t+1} - q^\pi(s,a)]$$
$$+ \gamma^2\lambda^2\mathbb{Var}[v^\pi(s') - \bar{Q}_{t+1}]$$
$$+ \gamma^2\lambda^2\mathbb{Var}[\Delta_{t+1}]$$
$$+ \gamma^2\lambda^2\mathbb{Var}[\rho_{t+1}\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}}], \qquad (14)$$

where $\Delta_{t+1} = \bar{Q}_{t+1} - \rho_{t+1}Q_{t+1} - v^\pi(s') + \rho_{t+1}q^\pi(s', a')$.

Proof. See Appendix F. $\qquad\square$

Now, let's illustrate the significance of Eq.(14).

(I) It demonstrates total random sources lead to the variance. The first 3 terms reveal the variance of $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ is cased by the following factors correspondingly: the error of one-step Expected Sarsa for policy evaluation, the error between $\bar{Q}_{t+1}$ and true value v^π, and state-action transition randomness. The last term in (14) is the variance of future time.

(II) Please notice that if the CV term $\bar{Q}_{t+1} - \rho_{t+1}Q_{t+1}$ (in Δ_{t+1}) vanishes, i.e. $\Delta_{t+1} = -v^\pi(s') + \rho_{t+1}q^\pi(s', a')$, Eq.(14) is reduced to the recursive variance of $G_t^{\lambda\rho,\text{ES}}$ (6). Thus, by Eq.(14), comparing the variance of $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ with $G_t^{\lambda\rho,\text{ES}}$ is equal to comparing the variance of Δ_{t+1}.

Furthermore, if a good estimator of q^π is available, the two following events happen:
1. For ES(λ)-CV, the term $\Delta_{t+1} \approx 0$. Since for a proper estimate of q^π, the following happens

$$\bar{Q}_{t+1} - \rho_{t+1}Q_{t+1} \approx 0, -v^\pi(s') + \rho_{t+1}q^\pi(s', a') \approx 0.$$

2. While, for ES(λ), $\Delta_{t+1} = -v^\pi(s') + \rho_{t+1}q^\pi(s', a')$, which is never be to 0, no matter how good an estimate of q^π we achieve.

Thus, if a good estimator of q^π is available, we have,

$$\underbrace{\mathbb{Var}[\Delta_{t+1}]}_{\text{for ES(λ)-CV iteration (7)}} \ll \underbrace{\mathbb{Var}[-v^\pi(s') + \rho_{t+1}q^\pi(s', a')]}_{\text{for ES(λ) iteration (6)}}.$$

Thus $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ enjoys a lower variance than $G_t^{\lambda\rho,\text{ES}}$.

Numerical Analysis

We use an experiment to verify that CV is efficient to reduce variance of ES(λ) for off-policy evaluation task. In this experiment, the target policy π is greedy policy, the value of π is selected by Q-learning with ϵ_k-greedy policy, where ϵ_k is decayed as $\epsilon_{k+1} = 0.95\epsilon_k$, $\epsilon_1 = 0.2$. After 150 episodes, $\epsilon_{150} \approx 0$, and the value of target policy π comes around -20. We use 0.2-greedy policy as behavior policy μ. All algorithms use step-size $\alpha_k = 0.5$ and $\lambda = 0.95$.
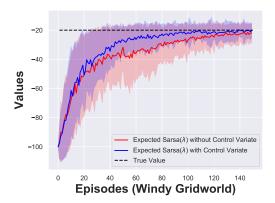

Figure 1: Comparison the performance between ES(λ)-CV and ES(λ) for off-policy evaluation task on windy gridworld. These unbroken lines are an average of 100 runs, and each run contains 150 episodes. To preferably show variance during the learning process, we show the shadow width as the standard deviation.

Gradient Expected Sarsa(λ)

In this section, we extend ES(λ)-CV with linear function approximation. Firstly, we prove the way to extend ES(λ)-CV with function approximation by (Sutton and Barto 2018) (section 12.9) is unstable. Then, we propose a convergent gradient Expected Sarsa(λ).

The Bellman equation (1) cannot be solved directly by tabular method for a large dimension of $\mathcal{S}$. We often use a parametric function to approximate $q^\pi(s, a) \approx \phi^\top(s, a)\theta = Q_\theta(s, a)$, where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^p$ is a *feature map*. Then Q_θ can be rewritten as a version of matrix $Q_\theta = \Phi\theta \approx q^\pi$, where Φ is a $|\mathcal{S}||\mathcal{A}| \times p$ matrix whose row is $\phi(s, a)$. We assume that Markov chain induced by behavior policy μ is ergodic (Bertsekas 2012), i.e. there exists a stationary distribution ξ such that $\forall (S_0, A_0) \in \mathcal{S} \times \mathcal{A}$, $\frac{1}{n} \sum_{k=1}^n P(S_k = s, A_k = a | S_0, A_0) \overset{n \to \infty}{\to} \xi(s, a)$. We denote Ξ as a $|\mathcal{S}| \times |\mathcal{A}|$ diagonal matrix whose diagonal element is $\xi(s, a)$.

Instability of ES(λ) with Function Approximation

A typical update to extend (8) has been presented in (Sutton and Barto 2018) (section 12.9),

$$\theta_{t+1} = \theta_t + \alpha_t(\widetilde{G}_{t,\theta}^{\lambda\rho,\text{ES}} - Q_\theta(S_t, A_t))\nabla Q_\theta(S_t, A_t)$$

$$= \theta_t + \alpha_t(\sum_{l=t}^\infty (\gamma\lambda)^{l-t}\delta_{l,\theta}^{\text{ES}}\rho_{t+1:l})\phi_t, \quad (15)$$
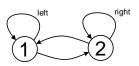

Figure 2: Two-state Example. We assign the features $\{(1,0)^\top, (2,0)^\top, (0,1)^\top, (0,2)^\top\}$ to the state-action pairs $\{(1, \text{right}), (2, \text{right}), (1, \text{left}), (2, \text{left})\}$, $\pi(\text{right}|\cdot) = 1$ and $\mu(\text{right}|\cdot) = 0.5$.

where α_t is step-size, $\delta_{l,\theta}^{\text{ES}} = R_{l+1} + \gamma\theta_l^\top \mathbb{E}_\pi[\phi(S_{l+1}, \cdot)] - \theta_l^\top \phi_l$, ϕ_l is short for $\phi(S_l, A_l)$. Once the system (15) has reached a stable state, for any θ_t, the expected parameter can been written as

$$\mathbb{E}[\theta_{t+1}|\theta_t] = \theta_t + \alpha_t(A\theta_t + b), \quad (16)$$

where

$$A = \Phi^\top\Xi(I - \gamma\lambda P^\mu)^{-1}(\gamma P^\pi - I)\Phi, \quad (17)$$

$$b = \Phi^\top\Xi(I - \gamma\lambda P^\mu)^{-1}r, r = \mathbb{E}[R_{t+1}|S_t, A_t]. \quad (18)$$

If the system (16) converges, then θ_t converges to the *TD fixed point* θ^* that satisfies $A\theta^* + b = 0$.

What condition guarantees the convergence of the (15)/ (16)? Unfortunately, the instability of (15) for off-policy is firstly realized by Sutton and Barto(2018), but it is only an intuitive guess inspired by previous works. Now, we provide a simple but rigorous theoretical analysis to illustrate the divergence of Eq.(15). It is known that for on-policy learning $\mu = \pi$, A is a negative definite matrix (Tsitsiklis and Van Roy 1997). Thus, for on-policy learning, (15) converges to $-A^{-1}b$. However, for off-policy learning, since the steady state-action distribution does not match the transition probability and $P^\pi\xi \neq \xi$, which results in, there is no guarantee that A is a negative definite matrix (Tsitsiklis and Van Roy 1997). Thus (15) may diverge.

An Unstable Example Now, we use a typical example (Touati et al. 2018) to illustrate the instability of iteration (15). The state transition of the example is presented in Figure 2. After some simple algebra (the detail is provided in Appendix G), we have $A = \begin{pmatrix} \frac{6\gamma - \gamma\lambda - 5}{2(1-\gamma\lambda)} & 0 \\ \frac{3\gamma}{2} & -\frac{5}{2} \end{pmatrix}$. For any $\theta_0 = (\theta_{0,1}, \theta_{0,2})^\top$, a positive constant step-size α, according to (16), we have

$$\mathbb{E}[\theta_{t+1}|\theta_t] \overset{\text{def}}{=} (\theta_{t+1,1}, \theta_{t+1,2})^\top, \quad (19)$$

$$\theta_{t+1,1} = \theta_{0,1}\prod_{l=0}^t(1 + \alpha\frac{6\gamma - \gamma\lambda - 5}{2(1-\gamma\lambda)}), \quad (20)$$

$$\theta_{t+1,2} = \theta_{0,2}\prod_{l=0}^t(1 - \alpha\frac{5}{2})^\top \quad (21)$$

For any $\lambda \in (0, 1)$, $\gamma \in (\frac{5}{6-\lambda}, 1)$, $\frac{6\gamma-\gamma\lambda-5}{2(1-\gamma\lambda)}$ is a positive scalar. Since then A cannot be a negative matrix. Furthermore, according to (20),

$$|\theta_{t+1,1}| = |\theta_{0,1}||(1 + \alpha\frac{6\gamma - \gamma\lambda - 5}{2(1-\gamma\lambda)})^{t+1}| \to +\infty.$$

Convergent Algorithm

The above discussion of the instability for off-policy learning shows that we should abandon the way presented in (15). In this section, we propose a convergent gradient ES(λ) algorithm.

We solve the problem by mean square projected Bellman equation (MSPBE) (Sutton et al. 2009a),

$$\text{MSPBE}(\theta, \lambda) = \frac{1}{2}\|\Phi\theta - \Pi\mathcal{B}_\lambda^\pi(\Phi\theta)\|_\Xi^2,$$

where $\Pi = \Phi(\Phi^T\Xi\Phi)^{-1}\Phi^T\Xi$ is an $|\mathcal{S}| \times |\mathcal{S}|$ projection matrix. Furthermore, $\text{MSPBE}(\theta, \lambda)$ can be rewritten as,

$$\min_\theta \text{MSPBE}(\theta, \lambda) = \min_\theta \frac{1}{2}\|A\theta + b\|_{M^{-1}}^2, \qquad (22)$$

where $M = \mathbb{E}[\phi_t\phi_t^\top] = \Phi^T\Xi\Phi$. The derivation of (22) is provided in Appendix H.

The computational complexity of the invertible matrix M^{-1} is at least $\mathcal{O}(p^3)$ (Golub and Van Loan 2012), where p is the dimension of feature space. Thus, it is too expensive to use gradient updates to solve the problem (22) directly. Besides, as pointed out in (Szepesvári 2010; Liu et al. 2015), we cannot get an unbiased estimate of $\nabla_\theta\text{MSPBE}(\theta, \lambda) = A^\top M^{-1}(A\theta + b)$. In fact, since the update law of gradient involves the product of expectations, the unbiased estimate cannot be obtained via a single sample. It needs to sample twice, which is a double sampling problem. Secondly, $M^{-1} = \mathbb{E}[\phi_t\phi_t^T]^{-1}$ cannot also be estimated via a single sample, which is the second bottleneck of applying stochastic gradient method to solve problem (22).

A practical way is converting (22) to be a convex-concave saddle-point problem (Liu et al. 2015). For $f : \mathbb{R}^d \to \mathbb{R}$, its *convex conjugate* (Bertsekas 2009) function $f^* : \mathbb{R}^d \to \mathbb{R}$ is defined as

$$f^*(y) = \sup_{x\in\mathbb{R}^d}\{y^Tx - f(x)\}.$$

By $(\frac{1}{2}\|x\|_M^2)^* = \frac{1}{2}\|y\|_{M^{-1}}^2$, we have $\frac{1}{2}\|y\|_{M^{-1}}^2 = \max_\omega(y^T\omega - \frac{1}{2}\|\omega\|_M^2)$. Thus, (22) is equivalent to the next convex-concave saddle-point problem

$$\min_\theta \max_\omega \{(A\theta + b)^\top\omega - \frac{1}{2}\|\omega\|_M^2\}. \qquad (23)$$

It is easy to see that if (θ^*, ω^*) is the solution of problem (23), then $\theta^* = \arg\min_\theta \text{MSPBE}(\theta, \lambda)$. In fact, let $\omega^* = \arg\max_\omega (A\theta + b)^\top\omega - \frac{1}{2}\|\omega\|_M^2$, then $\omega^* = M^{-1}(A\theta + b)$. Taking ω^* into (23), then (23) is reduced to $\min_\theta \frac{1}{2}\|A\theta + b\|_{M^{-1}}^2$, which illustrates that the solution of (22) contained in (23). Gradient update is a natural way to solve problem (23) (ascending in ω and descending in θ) as follows,

$$\omega_{t+1} = \omega_t + \beta_t(A\theta_t + b - M\omega_t), \qquad (24)$$

$$\theta_{t+1} = \theta_t - \alpha_t A^\top\omega_t, \qquad (25)$$

where α_t, β_t is step-size, $t \geq 0$.

Stochastic On-line Implementation However, since A, b, and M are versions of expectations, for model-free RL, we can not get the probability of transition. A practical way is to find the unbiased estimators of them. Let $e_0 =$

Algorithm 1 Gradient Expected Sarsa(λ)

Initialization: $\omega_0 = 0, \theta_0 = 0, \alpha_0 > 0, \beta_0 > 0$
for $i = 0$ **to** n **do**
  $e_{-1} = 0$
  for $t = 0$ **to** T_i **do**
    Observe $\{S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}\} \sim \mu$
    $e_t = \lambda\gamma\rho_t e_{t-1} + \phi_t$, where $\rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
    $\delta_t = R_{t+1} + \gamma\theta_t^\top\mathbb{E}_\pi\phi(S_{t+1}, \cdot) - \theta_t^\top\phi_t$
    $\omega_{t+1} = \omega_t + \beta_t(e_t\delta_t - \phi_t\phi_t^\top\omega_t)$
    $\theta_{t+1} = \theta_t - \alpha_t(\gamma\mathbb{E}_\pi[\phi(S_{t+1}, .)] - \phi_t)e_t^\top\omega_t$
  end for
end for
Output:θ

$0, \rho_t = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}, e_t = \lambda\gamma\rho_t e_{t-1} + \phi_t, \hat{b}_t = R_{t+1}e_t, \hat{A}_t = e_t(\gamma\mathbb{E}_\pi[\phi(S_{t+1}, .)] - \phi_t)^\top, \hat{M}_t = \phi_t\phi_t^\top$. By Theorem 9 in (Maei 2011), we have

$$\mathbb{E}[\hat{A}_t] = A, \mathbb{E}[\hat{b}_t] = b, \mathbb{E}[\hat{M}_t] = M.$$

Replacing the expectations in (24) and (25) by corresponding unbiased estimates, we define the stochastic on-line implementation of (24) and (25) as follows,

$$\omega_{t+1} = \omega_t + \beta_t(\hat{A}_t\theta_t + \hat{b}_t - \hat{M}_t\omega_t), \qquad (26)$$

$$\theta_{t+1} = \theta_t - \alpha_t\hat{A}_t^\top\omega_t. \qquad (27)$$

More details are summarized in Algorithm 1.

Convergence Analysis

We measure the convergence rate of problem (23) by *primal-dual gap error* (Nemirovski et al. 2009). Let

$$\Psi(\theta, \omega) = (A\theta + b)^T\omega - \frac{1}{2}\|\omega\|_M^2,$$

the primal-dual gap error at each solution (ω, θ) is

$$\epsilon_\Psi(\theta, \omega) = \max_{\omega'}\Psi(\theta, \omega') - \min_{\theta'}\Psi(\theta', \omega).$$

Theorem 4 (Convergence of Algorithm 1). *Consider the sequence $\{(\theta_t, \omega_t)\}_{t=1}^T$ generated by (27), step-size α, β are positive constants. Let (θ^*, ω^*) be the optimal solution of (23), $\bar{\theta}_T = \frac{1}{T}(\sum_{t=1}^T\theta_t), \bar{\omega}_T = \frac{1}{T}(\sum_{t=1}^T\omega_t)$ and we choose the step-size α, β satisfy $1 - \sqrt{\alpha\beta}\|A\|_* > 0$, where $\|A\|_* = \sup_{\|x\|=1}\|Ax\|$ is operator norm. If parameter (θ, ω) is on a bounded $D_\theta \times D_\omega$, i.e diam $D_\theta = \sup\{\|\theta_1 - \theta_2\|; \theta_1, \theta_2 \in D_\theta\} \leq \infty$, diam $D_\omega \leq \infty$, $\mathbb{E}[\epsilon_\Psi(\bar{\theta}_T, \bar{\omega}_T)]$ is upper bounded by:*

$$\sup_{(\theta, \omega)}\{\frac{1}{T}(\frac{\|\theta - \theta_0\|^2}{2\alpha} + \frac{\|\omega - \omega_0\|^2}{2\beta})\}.$$

Proof. See Appendix I. $\qquad\square$

Remark 3. *Theorem 4 illustrates (I) when $\alpha = \beta = \mathcal{O}(\frac{1}{\sqrt{T}})$, then the overall convergence rate of $\mathbb{E}[\epsilon_\Psi(\bar{\theta}_T, \bar{\omega}_T)]$ is $\mathcal{O}(\frac{1}{\sqrt{T}})$, which reaches the worst rate of black box oriented sub-gradient methods (Nesterov 2004); (II) when $\alpha = \beta = \mathcal{O}(1)$, a positive scalar, then $\mathbb{E}[\epsilon_\Psi(\bar{\theta}_T, \bar{\omega}_T)] = \mathcal{O}(\frac{1}{T})$.*

Algorithm	Reference	Step-size	Convergence Rate
TD(0)	(Nathaniel and Prashanth 2015)	$\alpha_t = \mathcal{O}(\frac{1}{t^\eta}), \eta \in (0,1)$	$\mathcal{O}(1/\sqrt{T})$
TD(0)	(Dalal et al. 2018a)	$\sum_{t=1}^{\infty} \alpha_t = \infty$	$\mathcal{O}(e^{-\frac{\sigma}{2}T^{1-\eta}} + \frac{1}{T^\eta})$
GTD(0)	(Dalal et al. 2018b)	$\sum_{t=1}^{\infty} \alpha_t = \infty, \frac{\beta_t}{\alpha_t} \to 0$	$\mathcal{O}((1/T)^{\frac{1-\kappa}{3}})$
GTD	(Liu et al. 2015)	constant step-size	$\mathcal{O}(1/\sqrt{T})$
GTD	(Wang et al. 2017)	$\sum_{t=1}^{\infty} \alpha_t = \infty, \frac{\sum_{t=1}^{T}\alpha_t^2}{\sum_{t=1}^{T}\alpha_t} \le \infty$	$\mathcal{O}(1/\sqrt{T})$
GTB/GRetrace	(Touati et al. 2018)	$\alpha_t, \beta_t = \mathcal{O}(\frac{1}{t})$	$\mathcal{O}(1/T)$
Ours		constant step-size	$\mathcal{O}(1/T)$

Table 1: Convergence Rate of Gradient Temporal Difference Learning

Related Works and Comparison

Liu et al.(2015) firstly derives GTD via convex-concave saddle-point formulation, and they prove the convergence rate reaches $\mathbb{E}[\epsilon_\Psi(\tilde{\theta}_T, \tilde{\omega}_T)] = \mathcal{O}(\frac{1}{\sqrt{T}})$, where $\tilde{\theta}_T$ is Polyak-average: $\tilde{\theta}_T = \frac{\sum_{t=1}^{T}\alpha_t\theta_t}{\sum_{t=1}^{T}\alpha_t}, \tilde{\omega}_T = \frac{\sum_{t=1}^{T}\alpha_t\omega_t}{\sum_{t=1}^{T}\alpha_t}$. Their GTD requires each θ_t, ω_t is projected into the space D_θ, D_ω. Later, Wang et al.(2017) extends the work of Liu et al.(2015), they suppose the data is generated from Markov processes rather than I.I.D assumption. Wang et al.(2017) prove the convergence rate $\mathbb{E}[\epsilon_\Psi(\tilde{\theta}_T, \tilde{\omega}_T)] = \mathcal{O}(\frac{\sum_{t=1}^{T}\alpha_t^2}{\sum_{t=1}^{T}\alpha_t})$, the best convergence rate reaches $\mathcal{O}(\frac{1}{\sqrt{T}})$, where the step-size satisfies $\sum_{t=1}^{\infty}\alpha_t = \infty, \frac{\sum_{t=1}^{T}\alpha_t^2}{\sum_{t=1}^{T}\alpha_t} \le \infty$ and $(\tilde{\theta}_T, \tilde{\omega}_T)$ is also Polyak-average, the same as (Liu et al. 2015). Besides, the GTD of Wang et al.(2017) also require projecting the parameter into the space D_θ, D_ω.

Both Polyak-averaging and projection make the implementation of gradient TD learning more difficult. Comparing with (Liu et al. 2015; Wang et al. 2017), our GES(λ) removes Polyak-averaging and projection, while reaches a faster convergence rate.

Recently, (Dalal et al. 2018b) proves GTD(0) family (Sutton et al. 2009a; Sutton, Maei, and Szepesvári 2009b) converges at $\mathcal{O}((\frac{1}{T})^{\frac{1-\kappa}{3}})$, but nerve reach $\mathcal{O}(\frac{1}{T})$, where $\kappa \in (0,1)$. Nathaniel and Prashanth (2015) proves TD(0) (Sutton 1988) converges at $\mathcal{O}(\frac{1}{\sqrt{T}})$ with step-size $\alpha_t = \mathcal{O}(\frac{1}{t^\eta})$, where $\eta \in (0,1)$. Then, Dalal et al.(2018a) further explores the property of TD(0), and they prove the convergence rate achieves $\mathcal{O}(e^{-\frac{\sigma}{2}T^{1-\eta}} + \frac{1}{T^\eta})$, but never reach $\mathcal{O}(\frac{1}{T})$, where $\eta \in (0,1)$, σ is the minimum eigenvalue of the matrix $A^\top + A$.

Comparing to the all above works, we improve the optimal convergence rate to $\mathcal{O}(\frac{1}{T})$ with a more relaxed step-size than theirs. Besides, although the GTB(λ)/GRetrace(λ) (Touati et al. 2018) reaches the same convergence rate as ours, their result depends on a decay step-size.

More details of the convergence rate of gradient temporal difference learning are summarized in Table 1.

Experiments

In this section, we employ three typical domains to test the capacity of GES(λ) for off-policy evaluation, *Mountaincar*, *Baird Star* (Baird 1995), and Two-state MDP (Touati et al. 2018). We compare GES(λ) with the three state-of-art algorithms: GQ(λ) (Maei and Sutton 2010), ABQ(ζ) (Mahmood, Yu, and Sutton 2017b), GTB(λ) (Touati et al. 2018). We choose the above three methods as baselines due to they are all learning by expected TD-error δ_t^{ES}, which is the same as GES(λ). For the limitation of space, we present some details of the experiments in Appendix J.

The Effect of Step-size

In this section, we verify the convergence result presented in Theorem 4/Remark 3. We use the empirical

$$\text{MSPBE} = \frac{1}{2}\|\hat{b} - \hat{A}\theta\|_{\hat{M}^{-1}}^2$$

to evaluate the performance of all the algorithms, where we evaluate $\hat{A}$, $\hat{b}$, and $\hat{M}$ according to their unbiased estimates by Monte Carlo method with 5000 episodes. Particular, for Mountaincar, to collect the samples, we run Sarsa with $p = 128$ features to obtain a stable policy. Then, we use this policy to collect trajectories that comprise the samples.

Figure 3 shows the comparison of the empirical MSPBE performance between a constant step-size and the decay step-size $\frac{1}{\sqrt{t}}$. Result (in Figure 3) illustrates that the GES(λ) with a proper constant step-size converges significantly faster than the learning with step-size $\frac{1}{\sqrt{t}}$, which support our theory analysis in Remark 3.

Comparison of Empirical MSPBE

The MSPBE distribution is computed over the combination of step-size, $(\alpha_k, \frac{\beta_k}{\alpha_k}) \in [0.1 \times 2^j | j = -10, -9, \cdots, -1, 0]^2$, and we set $\lambda = 0.99, \zeta = 0.95$ for ABQ(ζ).. All the result showed in Figure 4 is an average of 100 runs.

Result in Figure 4 shows that our GES(λ) learns significantly faster with better performance than GQ(λ), ABQ(ζ) and GTB(λ) in all domains. Besides, GES(λ) converges with a lower variance. We also notice that although Touati et al(2018) claim their GTB(λ) reaches the same convergence rate as our GES(λ), result in Figure 4 shows that our GES(λ) outperforms their GTB(λ) siginificantly.
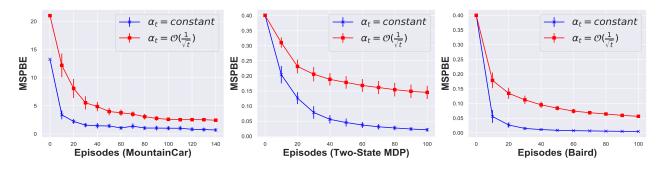
Figure 3: Comparison between the constant step-size and the decay step-size $\frac{1}{\sqrt{t}}$ for GES(λ).
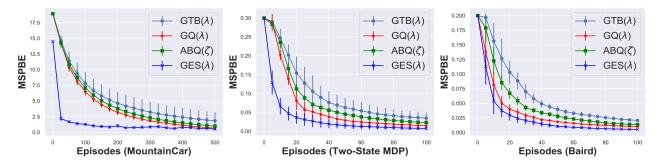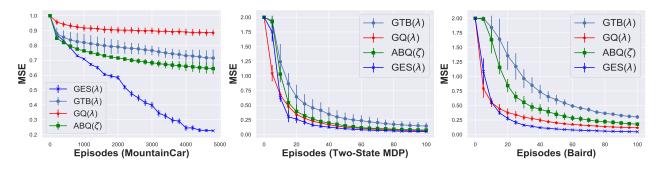

Figure 4: MSPBE comparison over episode.

Figure 5: MSE comparison over episode.

Comparison of Empirical MSE

We use the following empirical MSE according to (Adam and White 2016),

$$\text{MSE} = \|\Phi\theta - q^\pi\|_\Xi,$$

where q^π is estimated by simulating the target policy and averaging the discounted cumulative rewards overs trajectories. The combination of step-size for MSE is the same as previous empirical MSPBE. All the result showed in Figure 5 is an average of 100 runs and we set $\lambda = 0.99$, $\zeta = 0.95$ for ABQ(ζ).

The result in Figure 5 shows that GES(λ) converges significantly faster than all the three baselines with lower variance in Mountaincar domain. For the Two-state MDP and Baird domain, GES(λ) also achieves a better performance.

This conclusion further verifies the effectiveness of the proposed GES(λ).

Conclusion

In this paper, we introduce control variate technique to Expected Sarsa(λ) and propose ES(λ)-CV algorithm. We analyze all the random sources lead to the variance of ES(λ)-CV. We prove that if a good estimator of value function achieves, the ES(λ)-CV enjoys a lower variance than Expected Sarsa(λ) without control variate. Then, we extend ES(λ)-CV to be a convergent algorithm with function approximation and propose GES(λ) algorithm. We prove that the convergence rate of GES(λ) achieves $\mathcal{O}(1/T)$, which matches or outperforms several state-of-art gradient-based algorithms, but we use a more relaxed step-size. Finally,

we use numerical experiments to demonstrate the effectiveness of the proposed algorithm. Results show that the proposed algorithm converges faster and with lower variance than three typical algorithms GQ(λ), GTB(λ) and ABQ(ζ).

References

[A. Tamar and Mannor. 2016] A. Tamar, D. D., and Mannor., S. 2016. Learning the variance of the reward-to-go. *The Journal of Machine Learning Research* 17(13):1–36.

[Adam and White 2016] Adam, A., and White, M. 2016. Investigating practical linear temporal difference learning. In *International Conference on Autonomous Agents & Multiagent Systems*, 494–502.

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

[Balamurugan and Bach 2016] Balamurugan, P., and Bach, F. 2016. Stochastic variance re- duction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, 1416–1424.

[Bertsekas 2009] Bertsekas, D. P. 2009. *Convex optimization theory*. Athena Scientific Belmont.

[Bertsekas 2012] Bertsekas, D. P. 2012. *Dynamic Programming and Optimal Control*, volume 2. Athena scientific Belmont, MA.

[Dalal et al. 2018a] Dalal, G.; Szorenyi, B.; Thoppe, G.; and Mannor, S. 2018a. Finite sample analyses for td(0) with function approximation. In *AAAI2018*.

[Dalal et al. 2018b] Dalal, G.; Szorenyi, B.; Thoppe, G.; and Mannor, S. 2018b. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Annual Conference on Learning Theory (COLT)*.

[De Asis and Sutton 2018] De Asis, K., and Sutton, R. S. 2018. Per-decision multi-step temporal difference learning with control variates. In *Association for Uncertainty in Artificial Intelligence (UAI)*.

[Golub and Van Loan 2012] Golub, G. H., and Van Loan, C. F. 2012. *Matrix computations*, volume 3.

[Liu et al. 2015] Liu, B.; Liu, J.; Ghavamzadeh, M.; Mahadevan, S.; and Petrik, M. 2015. Finite-sample analysis of proximal gradient td algorithms. In *UAI*, 504–513. Citeseer.

[Liu et al. 2018] Liu, H.; Zhou, D.; Feng, Y.; Peng, J.; Mao, Y.; and Liu, Q. 2018. Action-dependent control variates for pol- icy optimization via steins identity. In *International Conference of Learning Representation*.

[Maei and Sutton 2010] Maei, H. R., and Sutton, R. S. 2010. Gq(λ): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the third conference on artificial general intelligence*, volume 1, 91–96.

[Maei 2011] Maei, H. R. 2011. *Gradient temporal-difference learning algorithms*. Ph.D. Dissertation, University of Alberta Edmonton, Alberta.

[Mahmood, Yu, and Sutton 2017b] Mahmood, A. R.; Yu, H.; and Sutton, R. S. 2017b. Multi-step off-policy learning without importance sampling ratios.

[Mahmood 2017a] Mahmood, A. 2017a. *Incremental off-policy reinforcement learning algorithms*. Ph.D. Dissertation, University of Alberta Edmonton, Alberta.

[Nathaniel and Prashanth 2015] Nathaniel, K., and Prashanth, L. 2015. On td(0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *International Conference on Machine Learning*, 626–634.

[Nemirovski et al. 2009] Nemirovski, A.; Juditsky, A.; Lan, G.; and Shapiro, A. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization* 19(4):1574–1609.

[Nesterov 2004] Nesterov, Y. 2004. *Introductory lectures on convex optimization: a basic course.* Kluwer Academic Publishers, Dordrecht.

[Rubinstein and Kroese 2016] Rubinstein, R. Y., and Kroese, D. P. 2016. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons.

[Rummery and Niranjan 1994] Rummery, G. A., and Niranjan, M. 1994. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering.

[Sutton and Barto 1998] Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*. MIT press Cambridge.

[Sutton and Barto 2018] Sutton, R. S., and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

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

[Sutton, Maei, and Szepesvári 2009b] Sutton, R. S.; Maei, H. R.; and Szepesvári, C. 2009b. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, 1609–1616.

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

[Szepesvári 2010] Szepesvári, C. 2010. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning* 4(1):1–103.

[Thomas and Brunskill 2016] Thomas, P., and Brunskill, E. 2016. Data-efficient off- policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2139–2148.

[Touati et al. 2018] Touati, A.; Bacon, P.-L.; Precup, D.; and Vincent, P. 2018. Convergent tree-backup and retrace with function approximation. In *International Conference on Machine Learning*.

[Tsitsiklis and Van Roy 1997] 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.

[Van Seijen et al. 2009] Van Seijen, H.; Van Hasselt, H.; Whiteson, S.; and Wiering, M. 2009. A theoretical and empirical analysis of expected sarsa. In *Adaptive Dynamic Programming and Reinforcement Learning, 2009.*, 177–184.

[Wang et al. 2017] Wang, Y.; Chen, W.; Liu, Y.; Ma, Z.-M.; and Liu, T.-Y. 2017. Finite sample analysis of the gtd policy evaluation algorithms in markov setting. In *Advances in Neural Information Processing Systems*, 5504–5513.

[Watkins 1989] Watkins, C. J. C. H. 1989. *Learning from delayed rewards*. Ph.D. Dissertation, King's College, Cambridge.

Appendix A: λ-Return of Sarsa for Off-policy Learning

For the discussion of off-policy learning, we need the background of importance sampling. Thus, the basic common conclusion about importance sampling (IS) and pre-decision importance sampling (PDIS) (**?**) is necessary.

Off-Policy Learning via Importance Sampling

Usually, we require that every action taken by π is also taken by μ, which is often called *coverage* (Sutton and Barto 2018) in reinforcement learning.

Assumption 1 (Coverage). *$\forall\,(s,a) \in \mathcal{S} \times \mathcal{A}$, we require that $\pi(a|s) > 0 \Rightarrow \mu(a|s) > 0$.*

The difficulty of off-policy roots in the discrepancy between target policy π and behavior policy μ —-we want to learn the target policy while we only get the data generated by behavior policy. One technique to hand this discrepancy is *importance sampling* (IS) (Rubinstein and Kroese 2016). Let $\tau_t^h = \{S_t, A_t, R_{t+1}\}_{t \geq 0}^h$ be a trajectory with finite horizon $h < \infty$. Let $\rho_{t:k} = \prod_{i=t}^k \rho_i$ denote the *cumulated importance sampling ratio*, where $\rho_i = \frac{\pi(A_i|S_i)}{\mu(A_i|S_i)}$ and $k \leq h$. Let $G_t^h = \sum_{k=0}^{h-t-1} \gamma^k R_{k+t+1}$, under Assumption 1 the IS estimator $G_t^{\text{IS}} = \rho_{t:h-1}G_t^h$ is a unbiased estimation of q^π. However, it is known that IS estimator suffers from large variance of the product $\rho_{t:h-1}$ (Sutton and Barto 1998). Pre-decision importance sampling (PDIS) (**?**) $G_t^{\text{PDIS}} = \sum_{k=0}^{h-t-1} \gamma^k \rho_{t:t+k} R_{t+k+1}$ is a practical variance reduction method without introducing bias, i.e. $\mathbb{E}_\mu[G_t^{\text{PDIS}}|S_t = s, A_t = a] = q^\pi(s,a)$.

$$\mathbb{E}_\mu[\rho_{t:h-1}G_t^h] = \mathbb{E}_\mu[\underbrace{\rho_{t:h-1}R_{t+1} + \rho_{t:h-1}\gamma R_{t+2} + \cdots + \rho_{t:h-1}\gamma^{h-t-1}R_h}_{\overset{\text{def}}{=}G_t^{\text{IS}}\ \text{IS-return}}]$$

$$= \mathbb{E}_\mu[\underbrace{\rho_t R_{t+1} + \rho_{t:t+1}\gamma R_{t+2} + \cdots + \rho_{t:h-1}\gamma^{h-t-1}R_h}_{\overset{\text{def}}{=}G_t^{\text{PDIS}}\ \text{PDIS-return}}] = \mathbb{E}_\mu[\sum_{k=0}^{h-t-1} \gamma^k \rho_{t:t+k} R_{t+k+1}].$$

For the equation $\mathbb{E}_\mu[G_t^{\text{IS}}] = \mathbb{E}_\mu[G_t^{\text{PDIS}}]$, please see(**?**) or section 5.9 in (Sutton and Barto 2018).

Lemma 1 (Section 3.10, (**?**); Section 5.9, (Sutton and Barto 2018)). *Let $\tau_t^h = \{S_k, A_k, R_{k+1}\}_{k=t}^h$ be the trajectory generated by behavior policy μ, for a given policy π and under Assumption 1, the following holds,*

$$\mathbb{E}_\mu[\rho_{t:h-1}R_{t+k}] = \mathbb{E}_\mu[\rho_{t:t+k-1}R_{t+k}]. \tag{28}$$

Lemma 1 implies that for any time $t+k$ $(k \geq 0)$, the importance sampling factors after $t+k$ have no effect in the expectation, thus the following holds: for all $k \geq 0$,

$$\mathbb{E}_\mu[\rho_{t:h-1}R_{t+k}] = \mathbb{E}_\mu[\rho_{t:t+k-1}R_{t+k}] = \mathbb{E}_\pi[R_{t+k}]. \tag{29}$$

λ-Return of Sarsa

The λ-*return* (Sutton and Barto 1998) is an average contains all the n-*step return* by weighting proportionally to λ^{n-1}, $\lambda \in [0,1]$. For example, let $G_t^{t+n} = \sum_{i=0}^{n-1} \gamma^i R_{t+i+1} + \gamma^n Q_{t+n}$ be n-step return, then the *standard forward view* of Sarsa(λ) is $G_t^{\lambda,\text{S}} = (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1} G_t^{t+n}$, which is equivalent to the following recursive version

$$G_t^{\lambda,\text{S}} = R_{t+1} + \gamma[(1-\lambda)Q_{t+1} + \lambda G_{t+1}^{\lambda,\text{S}}].$$

We only discuss the case of off-policy learning. On-Policy is a particular case of off-policy learning if $\rho_t = 1$. One version of λ-return of off-policy Sarsa(λ) via importance sampling is defined as the following recursive iteration (Section 12.8, (Sutton and Barto 2018)):

$$G_t^{\lambda\rho,\text{S}} = \rho_t(R_{t+1} + \gamma[(1-\lambda)Q_{t+1} + \lambda G_{t+1}^{\lambda\rho,\text{S}}]). \tag{30}$$

The next Proposition 2 gives a forward view of Eq.(30), and $G_t^{\lambda\rho,\text{S}}$ is an unbiased estimate of q^π.

Proposition 2. *Let μ be behavior policy and π be the target policy. $G_t^t = Q_t$, $G_t^{t+n} = \rho_t(R_{t+1} + \gamma G_{t+1}^{t+n})$, and $G_t^{\lambda\rho} = (1-\lambda)\sum_{n=1}^\infty \lambda^{n-1} G_t^{t+n}$, then $G_t^{\lambda\rho}$ is equivalent to $G_t^{\lambda\rho,\text{S}}$ defined in Eq.(30). Furthermore, $\mathbb{E}_\mu[G_t^{\lambda\rho}|(S_t, A_t) = (s,a)] = q^\pi(s,a)$.*

Proof. We restate the complete calculation process of off-policy λ-return $G_t^{\lambda\rho}$ as belowing

$$G_t^t = Q_t, G_t^{t+n} = \rho_t(R_{t+1} + \gamma G_{t+1}^{t+n}), \tag{31}$$

$$G_t^{\lambda\rho} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n} \tag{32}$$

$$= (1-\lambda)G_t^{t+1} + \lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n+1}$$

$$= (1-\lambda)\rho_t(R_{t+1} + \gamma Q_{t+1}) + \lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\Big(\underbrace{\rho_t(R_{t+1} + \gamma G_{t+1}^{t+n+1})}_{=G_t^{t+n+1};\text{Eq.(31)}|_{n\leftarrow n+1}}\Big)$$

$$= (1-\lambda)\rho_t(R_{t+1} + \gamma Q_{t+1}) + \lambda\rho_t R_{t+1} + \gamma\lambda\Big[\underbrace{(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t+1}^{t+n+1}}_{=G_{t+1}^{\lambda\rho};\text{Eq.(32)}|_{t\leftarrow t+1}}\Big]$$

$$= \rho_t\Big(R_{t+1} + \gamma[(1-\lambda)Q_{t+1} + \lambda G_{t+1}^{\lambda\rho}]\Big). \tag{33}$$

The last Eq.(33) implies that from the definition of standard λ-return Eq.(31) and Eq.(32), we can get the recursive form of Eq.(30).

Expanding Eq.(31), we get the complete n-step return as follows

$$G_t^{t+n} = \sum_{k=1}^{n}\gamma^{k-1}\rho_{t:t+k-1}R_{t+k} + \gamma^n\rho_{t:t+n}Q(S_{t+n}, A_{t+n}). \tag{34}$$

By Eq.(28) and Eq.(29), we have

$$\mathbb{E}_{\mu}[G_t^{t+n}|(S_t, A_t) = (s, a)]$$

$$= \mathbb{E}_{\mu}[\sum_{k=1}^{n}\gamma^{k-1}\rho_{t:t+k-1}R_{t+k} + \gamma^n\rho_{t:t+n}Q(S_{t+n}, A_{t+n})|(S_t, A_t) = (s, a)]$$

$$= \mathbb{E}_{\pi}[\sum_{k=1}^{n}\gamma^{k-1}R_{t+k} + \gamma^n Q(S_{t+n}, A_{t+n})|(S_t, A_t) = (s, a)] = q^{\pi}(s, a), \tag{35}$$

thus, $\mathbb{E}_{\mu}[G_t^{\lambda\rho}|(S_t, A_t) = (s, a)] = \mathbb{E}_{\mu}[(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n}|(S_t, A_t) = (s, a)] = q^{\pi}(s, a).$ $\square$

Appendix B: Proof of Eq.(5) and Proposition 1

Eq.(5): Recursive λ-Return of Expected Sarsa for On-policy Case

In this section, we prove **(I)** the forward view of Eq.(5); **(II)** Eq.(5) is an unbiased estimate of q^{π}.

Let $G_t^{\lambda,\text{ES}} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n}$, *where* $G_t^{t+n} = \sum_{i=0}^{n-1}\gamma^i R_{t+i+1} + \gamma^n\bar{Q}_{t+n}$ *is n-step return of Expected Sarsa and* $\bar{Q}_{t+n} = \mathbb{E}_{\pi}[Q(S_{t+n}, \cdot)]$, *then* $G_t^{\lambda,\text{ES}}$ *can be written recursively as:* $G_t^{\lambda,\text{ES}} = R_{t+1} + \gamma[(1-\lambda)\bar{Q}_{t+1} + \lambda G_{t+1}^{\lambda,\text{ES}}]$. *Besides,* $\mathbb{E}_{\pi}[G_t^{\lambda,\text{ES}}|(S_t, A_t) = (s, a))] = q^{\pi}(s, a).$

Proof. By the definition of n-step return of Expected Sarsa: $G_t^{t+n} = \sum_{i=0}^{n-1}\gamma^i R_{t+i+1} + \gamma^n\bar{Q}_{t+n}$, then G_t^{t+n} can be written as the following recursive form:

$$G_t^{t+n+1} = R_{t+1} + \gamma G_{t+1}^{t+n+1}. \tag{36}$$

Now, we turn to analyses $G_t^{\lambda,\text{ES}}$:

$$G_t^{\lambda,\text{ES}} = (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n}$$

$$= (1-\lambda)G_t^{t+1} + (1-\lambda)\sum_{n=2}^{\infty}\lambda^{n-1}G_t^{t+n}$$

$$= (1-\lambda)(R_{t+1}+\gamma\bar{Q}_{t+1}) + \lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n+1}$$

$$\overset{\text{Eq.(36)}}{=} (1-\lambda)(R_{t+1}+\gamma\bar{Q}_{t+1}) + \lambda(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}[R_{t+1}+\gamma G_{t+1}^{t+n+1}]$$

$$= (1-\lambda)(R_{t+1}+\gamma\bar{Q}_{t+1}) + \lambda R_{t+1} + \gamma\lambda\underbrace{\left[(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_{t+1}^{t+n+1}\right]}_{=G_{t+1}^{\lambda,\text{ES}}}$$

$$= R_{t+1} + \gamma[(1-\lambda)\bar{Q}_{t+1} + \lambda G_{t+1}^{\lambda,\text{ES}}],$$

which is the result in Eq.(5).

For on-policy learning, the following is obvious

$$\mathbb{E}_\pi[G_t^{t+n}] = \mathbb{E}_\pi\left[\sum_{i=0}^{n-1}\gamma^i R_{t+i+1} + \gamma^n\bar{Q}_{t+n}\right] = \mathbb{E}_\pi\left[\sum_{i=0}^{n-1}\gamma^i R_{t+i+1} + \gamma^n Q_{t+n}\right]. \tag{37}$$

It is similar to the Eq.(35), we have

$$\mathbb{E}_\pi[G_t^{\lambda,\text{ES}}|(S_t,A_t)=(s,a)] = \mathbb{E}_\pi\left[(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{t+n}|(S_t,A_t)=(s,a)\right] \tag{38}$$

$$\overset{(37)}{=} \mathbb{E}_\pi\left[(1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\left(\sum_{i=0}^{n-1}\gamma^i R_{t+i+1} + \gamma^n Q_{t+n}\right)|(S_t,A_t)=(s,a)\right] \tag{39}$$

$$\overset{(35)}{=} q^\pi(s,a), \tag{40}$$

which implies $G_t^{\lambda,\text{ES}}$ is an unbiased estimate of q^π. $\qquad\square$

Proof of Proposition 1

Proposition 1 *Let μ and π be the behavior and target policy, respectively. Consider the λ-return of Sarsa and Eq.(6), then*
$\mathbb{E}_\mu[G_t^{\lambda\rho,\text{ES}}|(S_t,A_t)=(s,a)] = \mathbb{E}_\pi[G_t^{\lambda,\text{S}}|(S_t,A_t)=(s,a)] = q^\pi(s,a).$

Proof. We expand $\mathbb{E}_\mu[G_t^{\lambda\rho,\text{ES}}|(S_t,A_t)=(s,a)]$ as follows

$$\mathbb{E}_\mu[G_t^{\lambda\rho,\text{ES}}|(S_t,A_t)=(s,a)]$$

$$= \mathbb{E}_\mu\left[R_{t+1} + \gamma[(1-\lambda)\bar{Q}_{t+1} + \lambda\rho_{t+1}G_{t+1}^{\lambda\rho,\text{ES}}]|(S_t,A_t)=(s,a)\right]$$

$$= \mathbb{E}_\pi\left[R_{t+1}+\gamma[(1-\lambda)Q_{t+1}]|(S_t,A_t)=(s,a)\right] + \mathbb{E}_\mu\left[\gamma\lambda\rho_{t+1}G_{t+1}^{\lambda\rho,\text{ES}}|(S_{t+1},A_{t+1})=(s',a')\right] \tag{41}$$

$$= \mathbb{E}_\pi\left[R_{t+1}+\gamma[(1-\lambda)Q_{t+1}]|(S_t,A_t)=(s,a)\right]$$

$$\quad + \gamma\lambda\sum_{s'\in\mathcal{S}}P_{ss'}^a\sum_{a'\in\mathcal{A}}\mu(a'|s')\frac{\pi(a'|s')}{\mu(a'|s')}\mathbb{E}_\mu[G_{t+1}^{\lambda\rho,\text{ES}}|(S_{t+1},A_{t+1})=(s',a')]$$

$$= \mathbb{E}_\pi\left[R_{t+1}+\gamma(1-\lambda)Q_{t+1}+\gamma\lambda\mathbb{E}_\mu[G_{t+1}^{\lambda\rho,\text{ES}}|(S_{t+1},A_{t+1})=(s',a')]\Big|(S_t,A_t)=(s,a)\right], \tag{42}$$

where Eq.(41) holds by the following facts: recall $\bar{Q}_{t+1} = \sum_{a \in \mathcal{A}} \pi(a|S_{t+1})Q_{t+1}(S_{t+1}, a)$, thus

$$\mathbb{E}_\mu[\bar{Q}_{t+1}] = \sum_{a \in \mathcal{A}} \mu(a|S_{t+1})\bar{Q}_{t+1} = \bar{Q}_{t+1} \underbrace{\sum_{a \in \mathcal{A}} \mu(a|S_{t+1})}_{=1} = \mathbb{E}_\pi[\bar{Q}_{t+1}].$$

If we continue to expand Eq.(42), then we have

$$\mathbb{E}_\mu[G_t^{\lambda\rho,\text{ES}}|(S_t, A_t) = (s, a)] = \mathbb{E}_\pi[G_t^{\lambda,\text{S}}|(S_t, A_t) = (s, a)] = q^\pi(s, a).$$

$\square$

Appendix C: Proof of Theorem 1

Theorem 1 (Forward View and Variance Analysis of Expected Sarsa(λ) with Control Variate) *Let μ and π denote the behavior and target policy, respectively. The λ-return with control variate defined in Eq.(7) is equivalent to the following forward view: let $G_t^t = Q_t$,*

$$G_t^{t+n} = R_{t+1} + \gamma\bar{Q}_{t+1} + \gamma(\rho_{t+1}G_{t+1}^{t+n} - \rho_{t+1}Q_{t+1}), \tag{43}$$

$$\widetilde{G}_t^{\lambda\rho,\text{ES}} = (1 - \lambda)\sum_{n=1}^{\infty} \lambda^{n-1}G_t^{t+n}. \tag{44}$$

Proof. Firstly, we prove Eq.(43),(44) is equivalent to Eq.(7). Let's expand $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ (in Eq.(44)),

$$\widetilde{G}_t^{\lambda\rho,\text{ES}} = (1 - \lambda)G_t^{t+1} + (1 - \lambda)\sum_{n=2}^{\infty} \lambda^{n-1}G_t^{t+n} \tag{45}$$

$$= (1 - \lambda)(\underbrace{R_{t+1} + \gamma\bar{Q}_{t+1}}_{=G_t^{t+1};\text{Eq.}(43),n=1}) + (1 - \lambda)\lambda\sum_{n=1}^{\infty} \lambda^{n-1}G_t^{t+n+1}$$

$$= (1 - \lambda)(R_{t+1} + \gamma\bar{Q}_{t+1})$$

$$\quad + (1 - \lambda)\lambda\sum_{n=1}^{\infty} \lambda^{n-1}\Big(\underbrace{R_{t+1} + \gamma(\rho_{t+1}G_{t+1}^{t+n+1} + \bar{Q}_{t+1} - \rho_{t+1}Q_{t+1})}_{=G_t^{t+n+1};\text{Eq.}(43),n \leftarrow n+1}\Big)$$

$$= (1 - \lambda)(R_{t+1} + \gamma\bar{Q}_{t+1})$$

$$\quad + \lambda(R_{t+1} + \gamma\bar{Q}_{t+1} - \gamma\rho_{t+1}Q_{t+1})) + \gamma\lambda\rho_{t+1}\underbrace{(1 - \lambda)\sum_{n=1}^{\infty} \lambda^{n-1}G_{t+1}^{t+n+1}}_{=\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}};\text{Eq.}(44)t \leftarrow t+1}$$

$$= R_{t+1} + \gamma\Big(\bar{Q}_{t+1} + \lambda\rho_{t+1}(\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} - Q_{t+1})\Big), \tag{46}$$

the last Eq.(46) implies

$$\widetilde{G}_t^{\lambda\rho,\text{ES}} = R_{t+1} + \gamma\big[(1 - \lambda)\bar{Q}_{t+1} + \lambda\big(\rho_{t+1}\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} + \bar{Q}_{t+1} - \rho_{t+1}Q_{t+1}\big)\big], \tag{47}$$

which is the Eq.(7)

$\square$

Appendix D: Proof of Eq.(11)

The Equivalence (a) for Eq.(11)

Proof.

$$q + \mathbb{E}_\mu[\sum_{l=t}^{\infty}(\lambda\gamma)^{l-t}\delta_l^{\text{ES}}\rho_{t+1:l}] \overset{(29)}{=} q + \mathbb{E}_\pi[\sum_{l=t}^{\infty}(\lambda\gamma)^{l-t}\delta_l^{\text{ES}}]$$

$$= q + (I - \lambda\gamma P^\pi)^{-1}(\mathcal{B}^\pi q - q), \tag{48}$$

Eq. (48) is a common result in RL, the details of $\mathbb{E}_\mu[\sum_{l=t}^{\infty}(\lambda\gamma)^{l-t}\delta_l^{\text{ES}}\rho_{t+1:l}] = \mathbb{E}_\pi[\sum_{l=t}^{\infty}(\lambda\gamma)^{l-t}\delta_l^{\text{ES}}]$ please refer to (**?**) or Section 6.3.9 in (Bertsekas 2012).

$\square$

Appendix E: Proof of Theorem 2

Theorem 2 (Policy Evaluation) *For any initial Q_0, consider the sequential trajectory collection $\mathcal{T}$, and the following Q_k is learned according to the k-th trajectory τ_k, $k \geq 1$,*

$$Q_{k+1} = \mathcal{B}_\lambda^\pi Q_k.$$

By iterating over k trajectories, the error of policy evaluation is upper bounded by

$$\|Q_k - q^\pi\| \leq \left(\frac{\gamma - \lambda\gamma}{1 - \lambda\gamma}\right)^k \|Q_0 - q^\pi\|.$$

Proof. (Proof of Theorem 2) By Eq.(11), the following equation holds (**?**; **?**),

$$\mathcal{B}_\lambda^\pi = (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n (\mathcal{B}^\pi)^{n+1}. \tag{49}$$

It is known that Bellman operator $\mathcal{B}^\pi$ is a γ-contraction (**?**),

$$\|\mathcal{B}^\pi Q_1 - \mathcal{B}^\pi Q_2\| \leq \gamma \|Q_1 - Q_2\|.$$

Thus we have

$$\|\mathcal{B}_\lambda^\pi Q_1 - \mathcal{B}_\lambda^\pi Q_2\| \overset{(49)}{\leq} (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n \|(\mathcal{B}^\pi)^{n+1}(Q_1 - Q_2)\|$$

$$\leq (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n \gamma \|(\mathcal{B}^\pi)^n (Q_1 - Q_2)\|$$

$$\cdots$$

$$\leq (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n \gamma^{n+1} \|Q_1 - Q_2\|$$

$$= \frac{(1 - \lambda)\gamma}{1 - \lambda\gamma} \|Q_1 - Q_2\|. \tag{50}$$

Since $0 < \dfrac{(1 - \lambda)\gamma}{1 - \lambda\gamma} < 1$, Eq.(50) implies that $\mathcal{B}_\lambda^\pi$ is a $\dfrac{(1 - \lambda)\gamma}{1 - \lambda\gamma}$-contraction. By Banach fixed point theorem (**?**), $\{Q_k\}_{k \geq 0}$ generated by $Q_{k+1} = \mathcal{B}_\lambda^\pi Q_k$ converges to the fixed point of $\mathcal{B}_\lambda^\pi$.

By Eq.(11), q^π is the unique fixed point of $\mathcal{B}_\lambda^\pi$. Thus, Q_{k+1} converges to q^π.

Now, we turn to consider the convergence rate. According to (50), it is easy to see $\forall k \in \mathbb{N}$, $\|Q_{k+1} - Q_k\| \leq \dfrac{(1 - \lambda)\gamma}{1 - \lambda\gamma} \|Q_k - Q_{k-1}\|$. Then, $\forall k, n \in \mathbb{N}$,

$$\|Q_{k+n} - Q_k\| \leq \frac{(1 - \lambda)\gamma}{1 - \lambda\gamma} \|Q_{k+n-1} - Q_{k-1}\|$$

$$\leq \left(\frac{(1 - \lambda)\gamma}{1 - \lambda\gamma}\right)^2 \|Q_{k+n-2} - Q_{k-2}\|$$

$$\cdots$$

$$\leq \left(\frac{(1 - \lambda)\gamma}{1 - \lambda\gamma}\right)^k \|Q_n - Q_0\|,$$

let $n \to \infty$, we have

$$\|Q_k - q^\pi\| \leq \left(\frac{\gamma - \lambda\gamma}{1 - \lambda\gamma}\right)^k \|Q_0 - q^\pi\|.$$

$\square$

Appendix F: Proof of Theorem 3

Theorem 3 $\widetilde{G}_t^{\lambda\rho,\text{ES}}$ *is an unbiased estimator of q^π, whose variance is given recursively as follows,*

$$\mathbb{V}\text{ar}\big[\widetilde{G}_t^{\lambda\rho,\text{ES}}\big] = \mathbb{V}\text{ar}\big[R_{t+1} + \gamma\bar{Q}_{t+1} - q^\pi(s,a)\big] + \gamma^2\lambda^2\mathbb{V}\text{ar}\big[v^\pi(s') - \bar{Q}_{t+1}\big]$$
$$+ \gamma^2\lambda^2\mathbb{V}\text{ar}[\Delta_{t+1}] + \gamma^2\lambda^2\mathbb{V}\text{ar}\big[\rho_{t+1}\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}}\big],$$

where $t \geq 0$, $\Delta_{t+1} = \bar{Q}_{t+1} - \rho_{t+1}Q_{t+1} - v^\pi(s') + \rho_{t+1}q^\pi(s',a')$.

Lemma 2. *The expectation of the cross-term between the TD error at t and the difference between the return and value at $t+1$ is zero: for any $q(s,a) = \mathbb{E}[G_{t+1}|S_t = s, A_t = a]$, i.e., satisfying the Bellman equation, for any bounded function $b : \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S} \to \mathbb{R}$,*

$$\mathbb{E}[b(S_t, A_t, R_{t+1}, S_{t+1})(G_{t+1} - q(S_{t+1}, A_{t+1}))|S_t = s, A_t = a] = 0. \tag{51}$$

A similar result of state value function appears in **(?)**, and Lemma 2 expends it to state-action value function. Thus,we omit its proof, and for the details please refer to **(?)**.

Remark 4. *If G_{t+1} is replaced by Expected Sarsa estimator $R_{t+1} + \gamma\bar{Q}_{t+1}$, Eq.(51) holds.*

Proof. **(Proof of Theorem 3)**

$$\mathbb{V}\text{ar}\big[\widetilde{G}_t^{\lambda\rho,\text{ES}}\big]$$
$$= \mathbb{E}\big[(\widetilde{G}_t^{\lambda\rho,\text{ES}})^2\big] - \underbrace{\big(\mathbb{E}\big[\widetilde{G}_t^{\lambda\rho,\text{ES}}\big]\big)^2}_{=(q^\pi(s,a))^2;\text{Proposition1},(7)}$$

$$\overset{(7)}{=} \mathbb{E}\Big[\Big(R_{t+1} + \gamma\big[(1-\lambda)\bar{Q}_{t+1} + \lambda\big(\rho_{t+1}\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} + \bar{Q}_{t+1} - \rho_{t+1}Q_{t+1}\big)\big]\Big)^2 - (q^\pi(s,a))^2\Big]$$

$$= \mathbb{E}\Big[\Big(R_{t+1} + \gamma\big[(1-\lambda)\bar{Q}_{t+1}$$
$$+ \lambda\big(\rho_{t+1}\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} + \underbrace{\bar{Q}_{t+1} - v^\pi(s') - \rho_{t+1}Q_{t+1} + \rho_{t+1}q^\pi(s',a')}_{\Delta_{t+1}} + v^\pi(s') - \rho_{t+1}q^\pi(s',a')\big)\big]\Big)^2$$
$$- (q^\pi(s,a))^2\Big]$$

$$= \mathbb{E}\Big[\Big(R_{t+1} + \gamma\big[(1-\lambda)\bar{Q}_{t+1} + \lambda\big(\rho_{t+1}\big(\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} - q^\pi(s',a')\big) + \Delta_{t+1} + v^\pi(s')\big)\big]\Big)^2$$
$$- (q^\pi(s,a))^2\Big]$$

$$= \mathbb{E}\Big[\Big(R_{t+1} + \gamma\bar{Q}_{t+1} + \gamma\lambda(v^\pi(s') - \bar{Q}_{t+1}) + \gamma\lambda\big(\rho_{t+1}\big(\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} - q^\pi(s',a')\big) + \Delta_{t+1}\big)\Big)^2$$
$$- (q^\pi(s,a))^2\Big]$$

$$= \mathbb{E}\Big[\Big(R_{t+1} + \gamma\bar{Q}_{t+1} - q^\pi(s,a) + \gamma\lambda(v^\pi(s') - \bar{Q}_{t+1})$$
$$+ \gamma\lambda\big(\rho_{t+1}\big(\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} - q^\pi(s',a')\big) + \Delta_{t+1}\big) + q^\pi(s,a)\Big)^2 - (q^\pi(s,a))^2\Big]$$

$$= \mathbb{E}\Big[\Big(R_{t+1} + \gamma\bar{Q}_{t+1} - q^\pi(s,a)\Big)^2\Big] + \gamma^2\lambda^2\mathbb{E}\Big[\big(v^\pi(s') - \bar{Q}_{t+1}\big)^2\Big] + \gamma^2\lambda^2\mathbb{E}[\Delta_{t+1}^2]$$
$$+ \gamma^2\lambda^2\mathbb{E}\Big[\rho_{t+1}^2\big(\widetilde{G}_{t+1}^{\lambda\rho,\text{ES}} - q^\pi(s',a')\big)^2\Big] \tag{52}$$

Eq.(52) holds due to Remark 4 and Lemma 1 in **(?)**. By the definition of variance, Eq.(52) is equivalent to Eq.(14), which is the result we want to prove. $\qquad\square$

Appendix G: Two-State MDP Example

$$P^\pi = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \implies (I - \gamma\lambda P^\pi) = \begin{pmatrix} 1 & -\gamma\lambda & 0 & 0 \\ 0 & 1-\gamma\lambda & 0 & 0 \\ -\gamma\lambda & 0 & 1 & 0 \\ -\gamma\lambda & 0 & 0 & 1 \end{pmatrix},$$

then, we have

$$(I - \gamma\lambda P^\pi)^{-1} = \begin{pmatrix} 1 & \frac{\gamma\lambda}{1-\gamma\lambda} & 0 & 0 \\ 0 & \frac{1}{1-\gamma\lambda} & 0 & 0 \\ \gamma\lambda & \frac{\gamma^2\lambda^2}{1-\gamma\lambda} & 1 & 0 \\ \gamma\lambda & \frac{\gamma^2\lambda^2}{1-\gamma\lambda} & 0 & 1 \end{pmatrix}.$$

$$A = \underbrace{\begin{pmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 \end{pmatrix}}_{=\Phi^\top} \underbrace{\frac{1}{2}I}_{=\Xi} \underbrace{\begin{pmatrix} 1 & \frac{\gamma\lambda}{1-\gamma\lambda} & 0 & 0 \\ 0 & \frac{1}{1-\gamma\lambda} & 0 & 0 \\ \gamma\lambda & \frac{\gamma^2\lambda^2}{1-\gamma\lambda} & 1 & 0 \\ \gamma\lambda & \frac{\gamma^2\lambda^2}{1-\gamma\lambda} & 0 & 1 \end{pmatrix}}_{=(I-\gamma\lambda P^\pi)^{-1}} \underbrace{\begin{pmatrix} -1 & \gamma & 0 & 0 \\ 0 & \gamma-1 & 0 & 0 \\ \gamma & 0 & -1 & 0 \\ \gamma & 0 & 0 & -1 \end{pmatrix}}_{=\gamma P^\pi - I} \underbrace{\begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}}_{=\Phi}$$

$$= \begin{pmatrix} \dfrac{6\gamma - \gamma\lambda - 5}{2(1-\gamma\lambda)} & 0 \\ \dfrac{3\gamma}{2} & -\dfrac{5}{2} \end{pmatrix}. \tag{53}$$

Appendix H: Proof of Eq.(22)

For a given policy π, $Q_\theta = \Phi\theta$, then by the definition of MSPBE objection function, we have,

$$\begin{aligned}
\text{MSPBE}(\theta, \lambda) &= \|Q_\theta - \Pi\mathcal{B}_\lambda^\pi Q_\theta\|_\Xi^2 \\
&= \|\Pi Q_\theta - \Pi\mathcal{B}_\lambda^\pi Q_\theta\|_\Xi^2 \\
&= \|\Phi^T\Xi(Q_\theta - \mathcal{B}_\lambda^\pi Q_\theta)\|_{(\Phi^T\Xi\Phi)^{-1}}^2 \\
&= \|\Phi^T\Xi(I - \lambda\gamma P^\pi)^{-1}(\Phi\theta - \gamma P^\pi\Phi\theta - R^\pi)\|_{(\Phi^T\Xi\Phi)^{-1}}^2 \\
&= \|\Phi^T\Xi(I - \lambda\gamma P^\pi)^{-1}\big((I - \gamma P^\pi)\Phi\theta - R^\pi\big)\|_{(\Phi^T\Xi\Phi)^{-1}}^2 \\
&= \|b + A\theta\|_{(\Phi^T\Xi\Phi)^{-1}}^2, \tag{54}
\end{aligned}$$

where $A = \Phi^T\Xi(I - \lambda\gamma P^\pi)^{-1}(\gamma P^\pi - I)\Phi, b = \Phi\Xi(I - \lambda\gamma P^\pi)^{-1}r$.

Appendix I: Proof of Theorem 4

Theorem 4 *Consider the sequence $\{(\theta_t, \omega_t)\}_{t=1}^T$ generated by (27), step-size α, β are positive constants. Let $\bar\theta_T = \frac{1}{T}(\sum_{t=1}^T \theta_t)$, $\bar\omega_T = \frac{1}{T}(\sum_{t=1}^T \omega_t)$ and we chose the step-size α, β satisfy $1 - \sqrt{\alpha\beta}\|A\|_* > 0$, where $\|A\|_* = \sup_{\|x\|=1}\|Ax\|$ is operator norm. If parameter (θ, ω) is on a bounded $D_\theta \times D_\omega$, i.e diam $D_\theta = \sup\{\|\theta_1 - \theta_2\|; \theta_1, \theta_2 \in D_\theta\} \le \infty$, diam $D_\omega \le \infty$, $\mathbb{E}[\epsilon_\Psi(\bar\theta_T, \bar\omega_T)]$ is upper bounded by:*

$$\sup_{(\theta,\omega)} \left\{ \frac{1}{T}\left(\frac{\|\theta - \theta_0\|^2}{2\alpha} + \frac{\|\omega - \omega_0\|^2}{2\beta}\right) \right\}.$$

The proof of Theorem 4 uses a inequality (in Eq.(55)) , we present it in the next Proposition 3.

Proposition 3. *Consider the update of expection version in Eq.(25),*

$$\omega_{t+1} = \omega_t + \beta(A\theta_t + b - M\omega_t), \theta_{t+1} = \theta_t - \alpha A^\top\omega_t.$$

Let $F(\omega) = \frac{1}{2}\|\omega\|_M^2 - b^\top \omega$, then for any $(\theta, \omega) \in D_\theta \times D_\omega$, the following hlods

$$\frac{1}{2\alpha}\|\theta - \theta_t\|^2 + \frac{1}{2\beta}\|\omega - \omega_t\|^2$$

$$\geq \frac{1}{2\alpha}(\|\theta_t - \theta_{t+1}\|^2 + \|\theta_{t+1} - \theta\|^2) + \frac{1}{2\beta}(\|\omega_t - \omega_{t+1}\|^2 + \|\omega_{t+1} - \omega\|^2)$$

$$+ \Big(\langle -A\theta, \omega_{t+1}\rangle + F(\omega_{t+1})\Big) - \Big(\langle -A\theta_{t+1}, \omega\rangle + F(\omega)\Big)$$

$$+ \Big\langle A(\theta_{t+1} - \theta_t), \omega_{t+1} - \omega\Big\rangle. \tag{55}$$

Proof. (**Proof of Proposition 3**) Let sub-gradients of f at x be denoted as $\partial f(x)$, $\partial f(x) = \{g | f(x) - f(y) \leq g^T(x - y), \forall y \in \mathbf{dom}(f)\}$. By the definition of sub-gradient , we have $\frac{\omega_t - \omega_{t+1}}{\beta} + A\theta_t \in \partial F(\omega_{t+1})$. Since F is convex, then for any$(\theta, \omega) \in D_\theta \times D_\omega$ the following holds

$$F(\omega) \geq F(\omega_{t+1}) + \langle \frac{\omega_t - \omega_{t+1}}{\beta} + A\theta_t, \omega - \omega_{t+1}\rangle.$$

By the *law of cosines*: $2\langle a - b, c - b\rangle = \|a - b\|^2 + \|b - c\|^2 - \|a - c\|^2$, we have

$$0 = \frac{1}{2\alpha}\Big(\|\theta_t - \theta_{t+1}\|^2 + \|\theta_{t+1} - \theta\|^2 - \|\theta_t - \theta\|^2\Big) + \underbrace{\langle -A^\top \omega_{t+1}, \theta - \theta_{t+1}\rangle}_{=-\langle A(\theta - \theta_{t+1}), \omega_{t+1}\rangle},$$

$$F(\omega) \geq F(\omega_{t+1}) + \frac{1}{2\beta}\Big(\|\omega_t - \omega_{t+1}\|^2 + \|\omega_{t+1} - \omega\|^2 - \|\omega_t - \omega\|^2\Big) + \langle A\theta_t, \omega - \omega_{t+1}\rangle,$$

summing them implies the following inequality,

$$\frac{1}{2\alpha}\|\theta - \theta_t\|^2 + \frac{1}{2\beta}\|\omega - \omega_t\|^2$$

$$\geq \frac{1}{2\alpha}(\|\theta_t - \theta_{t+1}\|^2 + \|\theta_{t+1} - \theta\|^2) + \frac{1}{2\beta}(\|\omega_t - \omega_{t+1}\|^2 + \|\omega_{t+1} - \omega\|^2)$$

$$+ \Big(\langle -A\theta, \omega_{t+1}\rangle + F(\omega_{t+1})\Big) - \Big(\langle -A\theta_{t+1}, \omega\rangle + F(\omega)\Big)$$

$$- \Big\langle -A(\theta_{t+1} - \theta_t), \omega_{t+1} - \omega\Big\rangle,$$

which is we want to prove. $\qquad\square$

Proof. (**Proof of Theorem 4**) Let $\bar{\theta}_t = 2\theta_t - \theta_{t-1}, \epsilon = \sqrt{\frac{\beta}{\alpha}}$. then for any $(\theta, \omega) \in D_\theta \times D_\omega$:

$$\Big\langle A(\theta_{t+1} - \bar{\theta}_t), \omega_{t+1} - \omega\Big\rangle = \Big\langle A\big((\theta_{t+1} - \theta_t) - (\theta_t - \theta_{t-1})\big), \omega_{t+1} - \omega\Big\rangle$$

$$= \Big\langle A(\theta_{t+1} - \theta_t), \omega_{t+1} - \omega\Big\rangle - \Big\langle A(\theta_t - \theta_{t-1}), \omega_{t+1} - \omega_t\Big\rangle$$

$$- \Big\langle A(\theta_t - \theta_{t-1}), \omega_t - \omega\Big\rangle$$

$$\geq \Big\langle A(\theta_{t+1} - \theta_t), \omega_{t+1} - \omega\Big\rangle - \|A\|_*\|\theta_t - \theta_{t-1}\|\|\omega_{t+1} - \omega_t\|$$

$$- \Big\langle A(\theta_t - \theta_{t-1}), \omega_t - \omega\Big\rangle$$

$$\geq \Big\langle A(\theta_{t+1} - \theta_t), \omega_{t+1} - \omega\Big\rangle - \|A\|_*\Big(\frac{\epsilon}{2}\|\theta_t - \theta_{t-1}\|^2 + \frac{1}{2\epsilon}\|\omega_{t+1} - \omega_t\|^2\Big)$$

$$- \Big\langle A(\theta_t - \theta_{t-1}), \omega_t - \omega\Big\rangle.$$

By the inequality in Proposition 3, we have

$$
\frac{1}{2\alpha}\|\theta - \theta_t\|^2 + \frac{1}{2\beta}\|\omega - \omega_t\|^2 \geq \frac{1}{2\alpha}\left(\|\theta_{t+1} - \theta\|^2 + \|\theta_t - \theta_{t+1}\|^2\right) - \sqrt{\alpha\beta}\|A\|_* \frac{\|\theta_t - \theta_{t-1}\|^2}{2\alpha}
$$
$$
+ (1 - \sqrt{\alpha\beta}\|A\|_*)\frac{1}{2\beta}\|\omega_t - \omega_{t+1}\|^2 + \frac{1}{2\beta}\|\omega_{t+1} - \omega\|^2
$$
$$
+ \left(\langle -A\theta, \omega_{t+1}\rangle + F(\omega_{t+1})\right) - \left(\langle -A\theta_{t+1}, \omega\rangle + F(\omega)\right)
$$
$$
+ \left\langle -A(\theta_t - \theta_{t-1}), \omega_t - \omega \right\rangle - \left\langle -A(\theta_{t+1} - \theta_t), \omega_{t+1} - \omega \right\rangle. \tag{56}
$$

Summing the Eq.(56) from $t = 0 : T - 1$

$$
\frac{1}{2\alpha}\|\theta - \theta_0\|^2 + \frac{1}{2\beta}\|\omega - \omega_0\|^2 \geq \frac{1}{2\alpha}\left(\|\theta_T - \theta\|^2 + \|\theta_T - \theta_{T-1}\|^2\right) - \sqrt{\alpha\beta}\|A\|_* \sum_{t=1}^{T-1} \frac{\|\theta_t - \theta_{t-1}\|^2}{2\alpha}
$$
$$
+ (1 - \sqrt{\alpha\beta}\|A\|_*) \sum_{t=0}^{T-1} \frac{1}{2\beta}\|\omega_t - \omega_{t+1}\|^2 + \frac{1}{2\beta}\|\omega_T - \omega\|^2
$$
$$
+ \sum_{t=0}^{T-1}\left[\left(\langle -A\theta, \omega_{t+1}\rangle + F(\omega_{t+1})\right) - \left(\langle -A\theta_{t+1}, \omega\rangle + F(\omega)\right)\right]
$$
$$
- \left\langle A(\theta_T - \theta_{T-1}), \omega_T - \omega \right\rangle.
$$

By the Cauchy-Schwarz inequality $|\langle \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\| \leq \frac{1}{2}(\|\mathbf{u}\|^2 + \|\mathbf{v}\|^2)$, we have

$$
\left\langle A(\theta_T - \theta_{T-1}), \omega_T - \omega \right\rangle \leq \frac{1}{2\alpha}\|\theta_T - \theta_{T-1}\|^2 + \alpha\beta\|A\|_*^2 \frac{1}{2\beta}\|\omega_T - \omega\|^2,
$$

then the following holds, for any $(\theta, \omega) \in D_\theta \times D_\omega$:

$$
\frac{1}{2\alpha}\|\theta - \theta_0\|^2 + \frac{1}{2\beta}\|\omega - \omega_0\|^2
$$
$$
\geq \sum_{t=0}^{T-1}\left[\left(\langle -A\theta, \omega_{t+1}\rangle + F(\omega_{t+1})\right) - \left(\langle -A\theta_{t+1}, \omega\rangle + F(\omega)\right)\right]
$$
$$
+ (1 - \sqrt{\alpha\beta}\|A\|_*) \sum_{t=0}^{T-1} \frac{1}{2\beta}\|\omega_t - \omega_{t+1}\|^2 + (1 - \alpha\beta\|A\|_*^2)\frac{1}{2\beta}\|\omega_T - \omega\|^2
$$
$$
+ \frac{1}{2\alpha}\|\theta_T - \theta\|^2 + (1 - \sqrt{\alpha\beta}\|A\|_*) \sum_{t=1}^{T-1} \frac{\|\theta_t - \theta_{t-1}\|^2}{2\alpha}. \tag{57}
$$

Let $\bar{\theta}_T = \frac{\sum_{t=0}^{T-1}\theta_t}{T}, \bar{\omega}_T = \frac{\sum_{t=0}^{T-1}\omega_t}{T}$ and we chose the step-size α, β satisfy $1 - \sqrt{\alpha\beta}\|A\| > 0$. By the convexity of $F(\omega)$ and $G(\theta)$, then we deduce from (57):

$$
\Big(\underbrace{\langle -A\theta, \bar{\omega}_T\rangle + F(\bar{\omega}_T)}_{-\Psi(\theta, \bar{\omega}_T)}\Big) - \Big(\underbrace{\langle -A\bar{\theta}_T, \omega\rangle + F(\omega)}_{-\Psi(\bar{\theta}_T, \omega)}\Big) \leq \frac{1}{T}\left(\frac{\|\theta - \theta_0\|^2}{2\alpha} + \frac{\|\omega - \omega_0\|^2}{2\beta}\right). \tag{58}
$$

By Eq.(58), we have

$$
\mathbb{E}[\epsilon_\Psi(\bar{\theta}_T, \bar{\omega}_T)] \leq \sup_{(\theta, \omega)}\left\{\frac{1}{T}\left(\frac{\|\theta - \theta_0\|^2}{2\alpha} + \frac{\|\omega - \omega_0\|^2}{2\beta}\right)\right\}.
$$

$\square$

Appendix J: Details of Experiments

MountainCar Since the state space of mountaincar domain is continuous, we use the open tile coding software http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/RLtoolkit/tilecoding.html to extract feature of states.

In this experiment, we set the number of tilings to be 4 and there are no white noise features. The performance is an average 5 runs and each run contains 5000 episodes. We set $\lambda = 0.99$, $\gamma = 0.99$. The MSPBE/MSE distribution is computed over the combination of step-size, $(\alpha_k, \frac{\beta_k}{\alpha_k}) \in [0.1 \times 2^j | j = -10, -9, \cdots, -1, 0]^2$, and $\lambda = 0.99$. Following suggestions from Section 10.1 in (Sutton and Barto 2018), we set all the initial state-action values to be 0, which is optimistic to cause extensive exploration.

Baird Example The Baird example considers the episodic seven-state, two-action MDP. The `dashed` action takes the system to one of the six upper states with equal probability, whereas the `solid` action takes the system to the seventh state. The behavior policy b selects the `dashed` and `solid` actions with probabilities $\frac{6}{7}$ and $\frac{1}{7}$, so that the next-state distribution under it is uniform (the same for all nonterminal states), which is also the starting distribution for each episode. The target policy π always takes the solid action, and so the on-policy distribution (for π) is concentrated in the seventh state. The reward is zero on all transitions. The discount rate is $\gamma = 0.99$. The feature $\phi(\cdot, \texttt{dashed})$ and $\phi(\cdot, \texttt{solid})$ are defined as follows,

$$\phi(\texttt{s}_1, \texttt{dashed}) = (2, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, , 0, , 0, 0, 0, 0)$$
$$\phi(\texttt{s}_2, \texttt{dashed}) = (0, 2, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, , 0, , 0, 0, 0, 0)$$
$$\cdots$$
$$\phi(\texttt{s}_7, \texttt{dashed}) = (0, 0, 0, 0, 0, 0, 2, 1, 0, 0, 0, 0, , 0, , 0, 0, 0, 0), \tag{59}$$

$$\phi(\texttt{s}_1, \texttt{solid}) = (0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, , 0, , 0, 0, 0, 1)$$
$$\phi(\texttt{s}_2, \texttt{solid}) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, , 0, , 0, 0, 0, 1)$$
$$\cdots$$
$$\phi(\texttt{s}_7, \texttt{solid}) = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, , 0, , 0, 0, 2, 1). \tag{60}$$