Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2026, 14(6), 2384-2408
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

A Unified Gradient Temporal Difference Learning Algorithm for Off-Policy Learning

Yafei Zhao[1*], Long Yang[2]

[1]Department of Mathematics, Zhejiang International Studies University, Hangzhou, China
[2]Chinese Academy of Sciences, Beijing, China
Email: *yfzhao@zisu.edu.cn, yanglong@ict.ac.cn

How to cite this paper: Zhao, Y.F. and Yang, L. (2026) A Unified Gradient Temporal Difference Learning Algorithm for Off-Policy Learning. *Journal of Applied Mathematics and Physics*, **14**, 2384-2408.
https://doi.org/10.4236/jamp.2026.146117

Received: May 28, 2026
Accepted: June 21, 2026
Published: June 24, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

In this paper, we propose a unification of gradient temporal difference (GTD) learning algorithm $\mathrm{GQ}(\sigma,\lambda)$ for off-policy learning. The proposed $\mathrm{GQ}(\sigma,\lambda)$ ranges from gradient Tree Backup(λ) to $\mathrm{GQ}(\lambda)$ when σ ranges from 0 to 1. We investigate the structure of TD fixed-point of $\mathrm{GQ}(\sigma,\lambda)$, and prove $\mathrm{GQ}(\sigma,\lambda)$ converges to its TD fixed-point with probability one. Furthermore, we prove that $\mathrm{GQ}(\sigma,\lambda)$ converges to an arbitrarily small neighborhood of the optimal solution with probability one. Empirical results show the $\mathrm{GQ}(\sigma,\lambda)$ with a value $\sigma\in(0,1)$ that creates a mixture of $\mathrm{GQ}(\lambda)$ and gradient Tree Backup(λ) achieves a better performance than both the extreme end $\sigma=0$ and $\sigma=1$.

Keywords

Reinforcement Learning, Off-Policy Learning

1. Introduction

In reinforcement learning (RL), unifying some disparate ideas not only providing a better understanding of existing algorithms but also creating better performing algorithms.

For example, $\mathrm{TD}(\lambda)$ [1] unifies one-step temporal difference learning (if $\lambda=0$) and Monte Carlo method (if $\lambda=1$) through the trace-decay parameter λ. Results show that the unified algorithm $\mathrm{TD}(\lambda)$ performs best at an intermediate value $\lambda\in(0,1)$ rather than the extreme cases of $\lambda=0$ and $\lambda=1$.

The work [2] [3] propose a multi-step $\mathrm{Q}(\sigma)$ that unifies n-step Sarsa [4] (if $\sigma=1$, *full-sampling*) and n-step Tree-Backup [5] ($\sigma=0$, *pure-expectation*),

where the parameter $\sigma \in [0,1]$ denotes the degree of the sampling. The work [2] have conducted experiments to show that for some value $\sigma \in (0,1)$, $\mathrm{Q}(\sigma)$ creates a mixture of full-sampling and pure-expectation approach, which performs better than the extreme case $\sigma = 0$ and $\sigma = 1$. Later, the work [6] [7] inherit the key idea of unification of $\mathrm{TD}(\lambda)$ and $\mathrm{Q}(\sigma)$, they propose $\mathrm{Q}(\sigma,\lambda)$ unifies $\mathrm{Sarsa}(\lambda)$ [4] and $\text{Tree-Backup}(\lambda)$ [5]. The previous works [6] [7] show that $\mathrm{Q}(\sigma,\lambda)$ performs best at an intermediate value $\sigma \in (0,1)$.

It is noteworthy that the theoretical analysis of the work [2] [6] [7] only consider the tabular learning, which requires a very large table to store the estimated value function when the state space is huge. That implies the previous methods of [2] [6] [7] are considerably expensive for high-dimensional RL, which is the main focus of this work.

Our Main Works

A practical way to address the high-dimensional curse is using a parametric function to estimate the value function. In this paper, we focus on extending $\mathrm{Q}(\sigma,\lambda)$ with linear function approximation. Since the divergence of semi-gradient with multi-step bootstrapping for off-policy learning are well-documented in the existing literature (e.g., [3] [8]), which could also happen in semi-gradient $\mathrm{Q}(\sigma,\lambda)$. To propose a convergent gradient-based algorithm, we derive the $\mathrm{GQ}(\sigma,\lambda)$ algorithm via the mean square projected Bellman error (MSPBE) objective function [9], that inspired by weight-duplication trick (also known as "two-timescale stochastic approximation") [9] [10]. When σ ranges from 0 to 1, $\mathrm{GQ}(\sigma,\lambda)$ ranges from gradient $\text{Tree Backup}(\lambda)$ ($\mathrm{TB}(\lambda)$) to $\mathrm{GQ}(\lambda)$ [11], *i.e.*, our $\mathrm{GQ}(\sigma,\lambda)$ unifies gradient $\text{Tree Backup}(\lambda)$ and $\mathrm{GQ}(\lambda)$.

Although $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ is a natural algorithm to extend $\mathrm{TB}(\lambda)$ with linear function approximation, to the best of our knowledge, the update rule of $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ has not been proposed in the existing literatures. It is worth to notice that Touati *et al.*, [12] have proposed another version of gradient $\mathrm{TB}(\lambda)$ ($\mathrm{GTB}(\lambda)$), which is different from the proposed $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$, we have clarified this point in Remark 3.

Then, we provide the convergence analysis of the proposed $\mathrm{GQ}(\sigma,\lambda)$. Theorem 1 shows that $\mathrm{GQ}(\sigma,\lambda)$ converges to its TD fixed-point with probability one. Additionally, Theorem 1 illustrates the structure of such TD fixed-point: it is the global asymptotically stable equilibrium of its corresponding ordinary differential equation (ODE). For more discussion, see Remark 4. Furthermore, Theorem 2 shows that $\mathrm{GQ}(\sigma,\lambda)$ converges to an arbitrarily small neighborhood of the optimal solution with probability one.

Finally, our experiments show that when σ ranges from 0 to 1, $\mathrm{GQ}(\sigma,\lambda)$ achieves the best performance of off-policy evaluation or control within a certain $\sigma \in (0,1)$, neither $\sigma = 0$, nor $\sigma = 1$, which implies that with a certain value

$\sigma \in (0,1)$, $\mathrm{GQ}(\sigma,\lambda)$ creates a mixture between $\mathrm{GQ}(\lambda)$ and gradient $\mathrm{TB}(\lambda)$ reaches a better performance than the extreme ends ($\sigma = 0$ and $\sigma = 1$).

2. Preliminary

In this section, we briefly review the basics of reinforcement learning and off-policy evaluation.

2.1. Reinforcement Learning

Reinforcement learning (RL) [3] is often formalized as Markov decision processes (MDP) that considers a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$; $\mathcal{S}$ is a set with finite states, $\mathcal{A}$ is a set with finite actions; $P: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$, $p_{ss'}^{a} = P(S_t = s' \mid S_{t-1} = s, A_{t-1} = a)$ is the probability of state transition from s to s' under playing the action a; $R(\cdot,\cdot): \mathcal{S} \times \mathcal{A} \to \mathbb{R}^1$ is the expected reward function; $\gamma \in (0,1)$.

A policy π is a probability distribution on $\mathcal{S} \times \mathcal{A}$, and $\pi(a \mid s)$ denotes the probability of playing a in state s. Let $\{S_t, A_t, R_{t+1}\}_{t \geq 0}$ be generated by π, its *state-action value function* is: $q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \mid S_0 = s, A_0 = a\right]$, where $\mathbb{E}_{\pi}[\cdot \mid \cdot]$ is conditional expectation on the actions selected according to π. Let $\mathcal{B}^{\pi}: \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ denote Bellman operator with respect to policy π:

$$\mathcal{B}^{\pi}: q \mapsto R^{\pi} + \gamma P^{\pi} q, \tag{1}$$

where $P^{\pi} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$, $R^{\pi} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$, their corresponding elements are: $\left[P^{\pi}\right]_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a \mid s) p_{ss'}^{a}$, $\left[R^{\pi}\right]_{s,a} = R(s,a)$. It is well-known that q^{π} is the unique fixed point of $\mathcal{B}^{\pi}$, *i.e.*, $\mathcal{B}^{\pi} q^{\pi} = q^{\pi}$, which is known as Bellman equation.

2.2. Off-Policy Evaluation

Let us consider the trajectory $\tau =: \{S_t, A_t, R_{t+1}\}_{t \geq 0}$ generated by the behavior policy μ, where $A_t \sim \mu(\cdot \mid S_t)$, $S_{t+1} \sim P(\cdot \mid S_t, A_t)$. Off-policy evaluation is the task to estimate the value function of the target policy π via the data that is generated by the behavior policy μ, where $\mu \neq \pi$.

Assumption 1 (Ergodicity). *The Markov chain induced by behavior policy μ is ergodic, i.e., there exists a stationary distribution $\xi(\cdot,\cdot)$ over $\mathcal{S} \times \mathcal{A}$: for $\forall (S_0, A_0)$,*

$$\frac{1}{n} \sum_{k=1}^{n} \mathbb{P}(S_k = s, A_k = a \mid S_0, A_0) \xrightarrow{n \to \infty} \xi(s,a) > 0. \tag{2}$$

The ergodicity of behavior policy μ is a standard assumption in off-policy learning [3], and it implies each-action pair is visited under this behavior policy μ. We use Ξ to denote a diagonal matrix whose diagonal element is $\xi(s,a)$, *i.e.*, $\Xi = \mathrm{diag}\{\cdots, \xi(s,a), \cdots\}$.

2.3. Temporal Difference Learning and λ-Return

TD learning updates value function as follows, $\forall t \geq 0$,

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_t \delta_t, \tag{3}$$

where $Q(\cdot,\cdot)$ is an estimator of q^π, α_t is step-size and δ_t is TD error. Let $Q_t =: Q(S_t, A_t)$, if

$$\delta_t = \delta_t^{\mathrm{S}} =: R_{t+1} + \gamma Q_{t+1} - Q_t, \tag{4}$$

then update (3) is Sarsa [4]. If δ_t is expected TD error:

$$\delta_t^{\mathrm{ES}} = R_{t+1} + \gamma \sum_{a \in \mathcal{A}} \pi(a \mid S_{t+1}) Q(S_{t+1}, a) - Q_t, \tag{5}$$

then update (3) is Expected Sarsa [4].

The λ-return is an average contains all the n-step returns by weighting proportionally to λ^{n-1}, where $\lambda \in [0,1]$. In this paper, we mainly consider two classic λ-return: Tree Backup(λ) (TB(λ)) and Expected Sarsa(λ).

Tree Backup(λ). For each pair (S_t, A_t) in the trajectory τ, TB(λ) [5] estimates $q^\pi(S_t, A_t)$ by

$$G_t^\lambda = Q_t + \sum_{k=t}^{\infty} (\gamma\lambda)^{k-t} \delta_k^{\mathrm{ES}} \prod_{i=t+1}^{k} \pi(A_i \mid S_i), \tag{6}$$

where δ_k^{ES} is expected TD error. Precup *et al.*, [5] have proved the iteration (6) converges to q^π with probability one under some certain conditions.

Expected Sarsa(λ). Sutton and Barto [3] have proposed a multi-step TD learning extends Expected Sarsa to λ-return version: for each $t \geq 0$,

$$G_t^\lambda = Q_t + \sum_{k=t}^{\infty} (\gamma\lambda)^{k-t} \delta_k^{\mathrm{ES}} \prod_{i=t+1}^{k} \frac{\pi(A_i \mid S_i)}{\mu(A_i \mid S_i)}. \tag{7}$$

For the convenience, in the following paragraph, we consider the following notations,

$$\rho_i =: \frac{\pi(A_i \mid S_i)}{\mu(A_i \mid S_i)}, \quad \prod_{i=t}^{k} \rho_i =: \rho_{t:k}, \quad \rho_{t:t+1} = 1.$$

2.4. A Unified View

In this section, we review an approach to unify TB(λ) and Expected Sarsa(λ).

Q(σ) **Algorithm.** Recently, De Asis *et al.*, [2] propose Q(σ) unifies multi-step Sarsa and multi-step TB(0). Concretely, according to a mixed TD error $\delta_t^{\pi,\sigma}$:

$$\delta_t^{\pi,\sigma} = \sigma\delta_t^{\mathrm{S}} + (1-\sigma)\delta_t^{\mathrm{ES}}, \tag{8}$$

De Asis *et al.*, [2] construct a multi-step estimator:

$$G_t^\sigma = Q_t + \sum_{k=t}^{\infty} \gamma^{k-t} \delta_k^{\pi,\sigma} \prod_{i=t+1}^{k} \left[(1-\sigma)\pi(A_i \mid S_i) + \sigma\right], \tag{9}$$

where $\sigma \in [0,1]$ is *sampling parameter*. When σ ranges from 0 to 1, the update (9) ranges from multi-step $\mathrm{TB}(0)$ to multi-step Sarsa. Experimental results show that a certain $\sigma \in (0,1)$ results in a mixture of $\mathrm{TB}(0)$ and Sarsa performs better than both $\sigma = 0$ and $\sigma = 1$ [2], which implies unifying some seemingly disparate algorithmic ideas can create better performing algorithms.

Off-Policy $\mathrm{Q}(\sigma,\lambda)$. Later, De Asis, [7] proposes a multi-step returns as follows,

$$\begin{aligned} G_t^{\sigma,\lambda} &= Q_t + \sum_{k=t}^{\infty} \delta_k^{\mathrm{ES}} \prod_{i=t+1}^{k} \gamma\lambda\left((1-\sigma)\pi(A_i \mid S_i) + \sigma\rho_i\right) \\ &= Q_t + \sum_{k=t}^{\infty} \delta_k^{\mathrm{ES}} \prod_{i=t+1}^{k} \gamma\lambda c_{i,\sigma}, \end{aligned} \tag{10}$$

where

$$c_{i,\sigma} = (1-\sigma)\pi(A_i \mid S_i) + \sigma\rho_i. \tag{11}$$

When σ ranges from 0 to 1, the estimator (10) ranges from $\mathrm{TB}(\lambda)$ (6) to Expected Sarsa(λ) (7).

We introduce a λ-operator $\mathcal{B}_{\sigma,\lambda}^{\pi,\mu}(\cdot): \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ that is a high level view of the λ-return (10),

$$\begin{aligned} \mathcal{B}_{\sigma,\lambda}^{\pi,\mu}(\cdot): q \mapsto q + \mathbb{E}_{\mu}\left[\sum_{k=0}^{\infty} \delta_k^{\mathrm{ES}} \prod_{i=1}^{k} \gamma\lambda c_{i,\sigma}\right] \\ = q + \sigma\left(I - \lambda\gamma P^{\pi}\right)^{-1}\left(\mathcal{B}^{\pi}q - q\right) \\ + (1-\sigma)\left(I - \lambda\gamma P^{\pi,\mu}\right)^{-1}\left(\mathcal{B}^{\pi}q - q\right), \end{aligned} \tag{12}$$

where $\mathcal{B}^{\pi}$ is Bellman operator (1), $P^{\pi,\mu} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, and whose elements are: $\left[P^{\pi,\mu}\right]_{s,s'} = \sum_{a\in\mathcal{A}} \pi(a \mid s)\mu(a \mid s)p_{ss'}^{a}$.

Remark 1. *Yang et al.* [6] *propose another version of* $\mathrm{Q}(\sigma,\lambda)$ *algorithm that extends* $\mathrm{Q}(\sigma)$ (9) *with eligibility trace*:

$$\tilde{G}_t^{\sigma,\lambda} = Q_t + \sum_{k=t}^{\infty} (\lambda\gamma)^{k-t} \delta_k^{\pi,\sigma}. \tag{13}$$

It is noteworthy that at one extreme end $\sigma = 1$, *both* $\mathrm{Q}(\sigma)$ (9) *and* $\tilde{G}_t^{\sigma,\lambda}$ (13) *reduce to on-policy learning. Particularly,* [6] *prove that the performance of* $\tilde{G}_t^{\sigma,\lambda}$ *for off-policy evaluation is determined by parameter* σ:

$$\left\| \mathbb{E}_{\mu}\left[\tilde{G}_t^{\sigma,\lambda} \mid (S_t, A_t) = (s,a)\right] - q^{\pi}(s,a) \right\|_{\infty} \le \sigma C, \tag{14}$$

where C *is a positive constant never reaches* 0 *no matter how we choose the starting time* t. *The upper error bound of* (14) *illustrates the capacity of* $\tilde{G}_t^{\sigma,\lambda}$ *for off-policy evaluation decays monotonously when* σ *ranges from* 0 *to* 1. *At the extreme end* $\sigma = 1$, $\tilde{G}_t^{\sigma,\lambda}$ *achieves the worst performance of off-policy evaluation. We think this is a natural result since* $\tilde{G}_t^{\sigma,\lambda}\big|_{\sigma=1}$ *is an on-policy algorithm exactly. Thus, in this paper, we mainly concern* (10) *for off-policy learning.*

2.5. Linear Function Approximation

TD learning (3) requires a very huge table to store the estimate value function $Q(\cdot,\cdot)$ when $|\mathcal{S}|$ is very large, which implies tabular TD learning is considerably expensive for high-dimensional RL. We often use a parametric function $Q_\theta(\cdot,\cdot)$ to approximate

$$q^\pi(s,a) \approx \phi^\top(s,a)\theta =: Q_\theta(s,a), \tag{15}$$

where $\theta \in \mathbb{R}^p$ is the parameter need to be learned, $\phi(s,a) = (\varphi_1(s,a), \varphi_2(s,a), \cdots, \varphi_p(s,a))^\top$, and each $\varphi_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Furthermore, Q_θ can be rewritten as a version of matrix

$$Q_\theta = \Phi\theta \approx q^\pi, \tag{16}$$

where Φ is a matrix whose rows are the state-action feature vectors $\phi^\top(s,a)$.

3. Divergence of $Q(\sigma,\lambda)$ with Semi-Gradient

In this section, firstly, we derive the semi-gradient $Q(\sigma,\lambda)$ algorithm; then we briefly analyze the divergence of extending $Q(\sigma,\lambda)$ (10) with semi-gradient method. In fact, the divergence of semi-gradient off-policy TD methods are well-documented in the literature (e.g., [8] [12] [13]), which are not specific to $Q(\sigma,\lambda)$.

3.1. Semi-Gradient $Q(\sigma,\lambda)$

Recall $\tau = \{(S_t, A_t, R_{t+1})\}_{t\geq 0}$ is generated by behavior policy μ, let $\phi_t = \phi(S_t, A_t)$, we define semi-gradient $Q(\sigma,\lambda)$ as follows:

$$\begin{aligned}\theta_{t+1} &= \theta_t - \alpha_t \nabla_\theta \left(G_t^\lambda(\theta) - \theta^\top \phi_t\right)^2\Big|_{\theta=\theta_t} \\ &= \theta_t - \alpha_t \left(G_t^\lambda(\theta_t) - \theta_t^\top \phi_t\right) \nabla_\theta \left(-Q_\theta(S_t, A_t)\right)\Big|_{\theta=\theta_t} \\ &= \theta_t + \alpha_t \left(\sum_{k=t}^{\infty} \delta_k^{\mathrm{ES}}(\theta_t) \prod_{i=t+1}^{k} \gamma\lambda c_{i,\sigma}\right)\phi_t,\end{aligned} \tag{17}$$

where

$$G_t^\lambda(\theta_t) = \theta_t^\top \phi_t + \sum_{k=t}^{\infty} \delta_k^{\mathrm{ES}}(\theta_t) \prod_{i=t+1}^{k} \gamma\lambda c_{i,\sigma}$$

is an off-line estimator of value function according to Equation (10), TD error $\delta_k^{\mathrm{ES}}(\theta_t) = R_{k+1} + \gamma \mathbb{E}_\pi\left[\theta_t^\top \phi(S_{k+1}, \cdot)\right] - \theta_t^\top \phi_k$, and α_t is step-size. Let

$$A_t = \phi_t \sum_{k=t}^{\infty} \left(\gamma \mathbb{E}_\pi\left[\phi(S_{k+1}, \cdot)\right] - \phi_k\right)^\top \prod_{i=t+1}^{k} \gamma\lambda c_{i,\sigma}, \tag{18}$$

$$b_t = \phi_t \sum_{k=t}^{\infty} R_{k+1} \prod_{i=t+1}^{k} \gamma\lambda c_{i,\sigma}. \tag{19}$$

Then update (17) can be rewritten as follows,

$$\theta_{t+1} = \theta_t + \alpha_t \left(A_t \theta_t + b_t \right). \tag{20}$$

Furthermore, we have

$$A_\sigma =: \mathbb{E}_\mu \left[A_t \right] = \Phi^\top \Xi P_\sigma^{\pi,\mu} \left(\gamma P^\pi - I \right) \Phi, \tag{21}$$

$$b_\sigma =: \mathbb{E}_\mu \left[b_t \right] = \Phi^\top \Xi P_\sigma^{\pi,\mu} R, \tag{22}$$

where $P_\sigma^{\pi,\mu} = \sigma \left(I - \lambda \gamma P^\pi \right)^{-1} + \left(1 - \sigma \right) \left(I - \lambda \gamma P^{\pi,\mu} \right)^{-1}$.

3.2. Divergence Analysis

Now, we only briefly discuss the divergence lies in the iteration (17). Under the conditions of Proposition 4.8 presented by Bertsekas and Tsitsiklis [14], $\{\theta_t\}_{t \geq 0}$ (17) converges to a certain point if and only if A_σ is a negative matrix. Furthermore, if iteration (17) converges, then it converges to its unique *TD fixed point* $\theta_\star$ that satisfes

$$A_\sigma \theta_\star + b_\sigma = 0. \tag{23}$$

Unfortunately, since the steady state-action distribution doesn't match the transition probability during off-policy learning, we can not guarantee the negative definiteness of A_σ, thus $\{\theta_t\}_{t \geq 0}$ may diverge. To clarify this point, we use the classic counterexample [12] to show the divergence of semi-gradient TD algorithms (17) for off-policy learning.

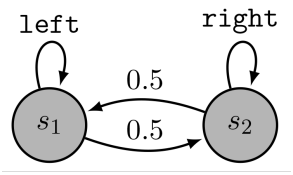


Figure 1. Counterexample from [12]: Two-State MDP.

For the MDP in **Figure 1**, the behavior policy $\mu(\text{right} \mid \cdot) = 0.5$, and target policy $\pi(\text{right} \mid \cdot) = 1$. We assign the features $\left\{ (1,0)^\top, (2,0)^\top, (0,1)^\top, (0,2)^\top \right\}$ to the state-action pairs $\left\{ (s_1, \text{right}), (s_2, \text{right}), (s_1, \text{left}), (s_2, \text{left}) \right\}$, From the dynamic transition shown in **Figure 1**, we have

$$P^\pi = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \Phi = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 2 \end{pmatrix}, \Xi = \frac{1}{2} I_{4\times 4}.$$

Then, according to (21), we have

$$\begin{aligned} A_\sigma &= \Phi^\top \Xi P_\sigma^{\pi,\mu} \left(\gamma P^\pi - I \right) \Phi \\ &= \begin{pmatrix} \dfrac{6(2-\sigma)\gamma - \gamma\lambda - 5(2-\sigma)}{2(1-\gamma\lambda)} & 0 \\ \dfrac{3\gamma}{2} & -5\left(1 - \dfrac{\sigma}{2}\right) \end{pmatrix}, \end{aligned}$$

and the eigenvalues of A are: $\frac{6(2-\sigma)\gamma-\gamma\lambda-5(2-\sigma)}{2(1-\gamma\lambda)}$ and $-5\left(1-\frac{\sigma}{2}\right)$. For any initial $\theta_0=(\theta_{0,1},\theta_{0,2})^\top$, let $\mathbb{E}[\theta_{t+1}]=:(\theta_{t+1,1},\theta_{t+1,2})^\top$, according to (20), the first component of the term $\mathbb{E}[\theta_{t+1}\mid\theta_t]$ is:

$$\theta_{t+1,1}=\theta_{0,1}\prod_{i=0}^{t}\left(1+\alpha_i\frac{6(2-\sigma)\gamma-\gamma\lambda-5(2-\sigma)}{2(1-\gamma\lambda)}\right). \tag{24}$$

For any $\lambda\in(0,1)$, if $\gamma\in\left(\frac{10-5\sigma}{12-6\sigma-\lambda},1\right)$, then $\frac{6(2-\sigma)\gamma-\gamma\lambda-5(2-\sigma)}{2(1-\gamma\lambda)}$ is a positive scalar, which implies A_σ can not be a negative matrix. Furthermore, if step size $\alpha_t:\sum_{i\geq 0}\alpha_t=\infty$, we have

$$\left|\theta_{t+1,1}\right|=\left|\theta_{0,1}\right|\prod_{i=0}^{t}\left(\frac{6(2-\sigma)\gamma-\gamma\lambda-5(2-\sigma)}{2(1-\gamma\lambda)}\right)\to+\infty, \tag{25}$$

Equation (25) is a direct result of the following conclusion that could be found in any calculus textbook. Let $p_i=1+a_i$, where $a_i>0$, if $\sum_{i=1}^{\infty}a_i=+\infty$, then $\prod_{i=1}^{\infty}p_i=\prod_{i=1}^{\infty}(1+a_i)=+\infty$, which implies the way (17) to extend $\mathrm{Q}(\sigma)$ with linear function approximation via off-line estimate is unstable for off-policy learning.

4. Gradient $\mathrm{Q}(\sigma,\lambda)$

In this section, we derive the gradient $\mathrm{Q}(\sigma,\lambda)$ ($\mathrm{GQ}(\sigma,\lambda)$) algorithm. The proposed $\mathrm{GQ}(\sigma,\lambda)$ unifies $\mathrm{GQ}(\lambda)$ [11] if $\sigma=1$. For more discussion, see Remark 2. At another extreme end $\sigma=0$, the proposed $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ can be seen as a new way to extend $\mathrm{TB}(\lambda)$ (6) with linear function approximation. Although $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ is a natural algorithm extends $\mathrm{TB}(\lambda)$ (6) with linear function approximation, to the best of our knowledge, the update rule of $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ has not been proposed in the existing literature. For more discussion, see Remark 3.

We derive the gradient the $\mathrm{GQ}(\sigma,\lambda)$ algorithm via mean square projected Bellman error (MSPBE) [9] objective function as follows,

$$J(\theta)=\frac{1}{2}\left\|\Phi\theta-\Pi\mathcal{B}_{\sigma,\lambda}^{\pi,\mu}(\Phi\theta)\right\|_{\Xi}^{2}, \tag{26}$$

where

$$\Pi=\Phi\left(\Phi^\top\Xi\Phi\right)^{-1}\Phi^\top\Xi$$

is an $|\mathcal{S}|\times|\mathcal{S}|$ *projection matrix*, $\|\cdot\|_{\Xi}^{2}$ is the weighted Euclidean norm: $\|x\|_{\Xi}^{2}=x^\top\Xi x$. Furthermore, we can rewrite $\min_\theta J(\theta)$ as follows,

$$\min_\theta J(\theta)=\min_\theta\frac{1}{2}\left\|A_\sigma\theta+b_\sigma\right\|_{M^{-1}}^{2}, \tag{27}$$

where $M=\mathbb{E}_{\mu}\left[\phi_t\phi_t^{\top}\right]=\Phi^{\top}\Xi\Phi$.

The gradient method is a natural approach to solve problem (27), however, it is worth to notice that the challenges are two-fold: (I) Firstly, since the invertible matrix M^{-1} is involved in $\nabla J(\theta)$, so it is too expensive to apply stochastic gradient to solve the problem (27) directly. (II) Since $\nabla J(\theta)=A_{\sigma}^{\top}M^{-1}\left(A_{\sigma}\theta+b_{\sigma}\right)$ involves the product of expectations, then the unbiased estimate of $\nabla J(\theta)$ cannot be obtained via a single sample. It needs to sample twice, so it is a double-sampling problem, which is the second bottleneck of applying gradient to solve the problem (27). Additionally, $M^{-1}=\mathbb{E}\left[\phi_t\phi_t^{\top}\right]^{-1}$ cannot also be estimated via a single sample. The above analysis pushes us to find a new practical way to solve the problem (27).

Let $e_{-1,\sigma}=0$, $e_{t,\sigma}=\lambda\gamma c_{t,\sigma}e_{t-1}+\phi_t$, $\bar{\phi}_{t+1}=\mathbb{E}_{\pi}\left[\phi\left(S_{t+1,\cdot}\right)\right]$, then we have the following equation:

$$-\frac{1}{2}\nabla J(\theta_t)=-\mathbb{E}_{\mu}\left[\left(\gamma\bar{\phi}_{t+1}-\phi_t\right)e_{t,\sigma}^{\top}\right]\varpi \tag{28}$$

$$-\frac{1}{2}\nabla J(\theta_t)=\mathbb{E}_{\mu}\left[\delta_t^{\mathrm{ES}}e_{t,\sigma}\right]-\mathbb{E}\left[\gamma(1-\lambda)\bar{\phi}_{t+1}e_{t,\sigma}^{\top}\right]\varpi, \tag{29}$$

where

$$\varpi=\mathbb{E}_{\mu}\left[\phi_t\phi_t^{\top}\right]^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}}e_{t,\sigma}\right],\delta_t^{\mathrm{ES}}=R_{t+1}+\gamma\theta_t^{\top}\bar{\phi}_{t+1}-\theta_t^{\top}\phi_t.$$

For the limitation of space, we provide the derivation of (28)-(29) in **Appendix A**. We use the sign convention that the mean update of θ follows the negative gradient $-\nabla J(\theta_t)$. The term (31) provides its unbiased stochastic approximation, while the auxiliary recursion (30) tracks $\varpi=M^{-1}\mathbb{E}_{\mu}\left[\delta_t^{\mathrm{ES}}e_{t,\sigma}\right]$, circumventing the double-sampling issue.

Let's consider the term $\varpi=\mathbb{E}_{\mu}\left[\phi_t\phi_t^{\top}\right]^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}}e_{t,\sigma}\right]$ that is a solution of a least-squares problem, and a typical least mean square (LMS) update rule to find the vector ϖ is:

$$\omega_{t+1}=\omega_t+\beta_t\left(\delta_t^{\mathrm{ES}}e_{t,\sigma}-\phi_t\omega_t^{\top}\phi_t\right), \tag{30}$$

where β_t is step-size. Then, by directly sampling from (29) with (30), we define the update rule of θ as follows,

$$\theta_{t+1}=\theta_t+\alpha_t\left(\delta_t^{\mathrm{ES}}e_{t,\sigma}-\gamma(1-\lambda)\bar{\phi}_{t+1}e_{t,\sigma}^{\top}\omega_t\right), \tag{31}$$

where α_t is step-size. We provide the details in **Algorithm 1**.

Remark 2 (Case of $\sigma=1$). *When $\sigma=1$, we regard $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=1}$ as an approach to extend Expected Sarsa(λ) (7) with linear function approximation. A more interesting result is that the proposed $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=1}$ is reduced to $\mathrm{GQ}(\lambda)$ [12] exactly. Now, we provide two fresh interpretations to the proposed $\mathrm{GQ}(\lambda)$:*

- *$\mathrm{GQ}(\lambda)$ is at one extreme end ($\sigma=1$) of the proposed $\mathrm{GQ}(\sigma,\lambda)$, i.e., the proposed $\mathrm{GQ}(\sigma,\lambda)$ contains $\mathrm{GQ}(\lambda)$.*

Algorithm 1 Gradient $Q(\sigma,\lambda)$

Require:Initialize parameter ω_0, θ_0 arbitrarily, $\gamma\in(0,1)$, $\lambda\in[0,1]$, $\sigma\in[0,1]$, $\alpha_t>0$, $\beta_t>0$.
Given: Target policy π, behavior policy μ. N is the total number of episodes.
for $i=0$ **to** N **do**
  $e_{-1}=0$.
  for $t=0$ **to** T_i **do**
    Observe $\{S_t,A_t,R_{t+1},S_{t+1}\}$ by μ.
    $c_{t,\sigma}=(1-\sigma)\pi(A_t|S_t)+\sigma\rho_t$,where $\rho_t=\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}$
    $e_t=\lambda\gamma c_{t,\sigma}e_{t-1}+\phi_t$
    $\delta_t=R_{t+1}+\gamma\theta_t^{\top}\mathbb{E}_{\pi}[\phi(S_{t+1},\cdot)]-\theta_t^{\top}\phi_t$
    $\omega_{t+1}=\omega_t+\beta_t(\delta_t e_t-\phi_t\omega_t^{\top}\phi_t)$
    $\theta_{t+1}=\theta_t+\alpha_t\left(\delta_t^{\mathrm{ES}}e_t-\gamma(1-\lambda)\mathbb{E}_{\pi}[\phi(S_{t+1},\cdot)]e_t^{\top}\omega_t\right)$
  end for
end for
Output:θ

- *Furthermore, just because* $\mathrm{GQ}(\lambda)$ *is same as* $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=1}$, $\mathrm{GQ}(\lambda)$ *can be seen as an extension of the tabular* Expected Sarsa(λ) (7) *with linear function approximation, while the original motivation of Maei and Sutton* [11] *to propose* $\mathrm{GQ}(\lambda)$ *is to introduce eligibility trace to gradient temporal difference learning for off-policy evaluation. Thus, the proposed* $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=1}$ *provides a fresh understanding for the* $\mathrm{GQ}(\lambda)$ [11].

Computational complexity. The computational complexity of **Algorithm 1** is $O(|\mathcal{A}|p)$ per step in time, and $O(p)$ in memory, where p is the feature dimension and $|\mathcal{A}|$ is the number of actions. This maintains the same asymptotic efficiency as the baseline $\mathrm{GQ}(\lambda)$ and gradient $\mathrm{TB}(\lambda)$ methods.

Remark 3 (Case of $\sigma=0$). *When* $\sigma=0$, *the algorithm* $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ *is reduced to a version of extending* $\mathrm{TB}(\lambda)$ (6) *with linear function approximation. Although* $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ *is a natural algorithm extends* $\mathrm{TB}(\lambda)$ (6) *with linear function approximation, to the best of our knowledge, the update rule of* $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ *has not been proposed in the existing literatures. It is worth to notice that* [12] *have also proposed another version of gradient* $\mathrm{TB}(\lambda)$ ($\mathrm{GTB}(\lambda)$), *while the difference between the proposed* $\mathrm{GQ}(\sigma,\lambda)\big|_{\sigma=0}$ *and* $\mathrm{GTB}(\lambda)$ [12] *is reflected at least two aspects*:

- *Firstly, the proposed gradient* $\mathrm{Q}(\sigma,\lambda)\big|_{\sigma=0}$ *and* $\mathrm{GTB}(\lambda)$ [12] *share the same update rule of* e_t *and* ω_t, *but instead of* (31), $\mathrm{GTB}(\lambda)$ *updates the parameter* θ *as follows*,

$$\theta_{t+1}=\theta_t+\alpha_t\left(-\gamma\bar{\phi}_{t+1}+\phi_t\right)e_t^{\top}\omega_t, \tag{32}$$

where $e_t=e_{t,\sigma}\big|_{\sigma=0}=\lambda\gamma\pi\left(A_t\,|\,S_t\right)e_{t-1}$.

- *Secondly, Touati et al.,* [12] *derive the update rule* (32) *of their* $\mathrm{GTB}(\lambda)$ *via the convex-concave saddle-point framework* [15], *while our* $\mathrm{GQ}(\sigma,\lambda)$ *is based on the weight-duplication trick*[1] *of* (29)-(31).

[1]The term "weight-duplication trick" we use here is coming from [10] [11], while some other literatures may call it "two-timescale stochastic approximation", e.g., [18].

5. Convergence Analysis

In this section, we prove the convergence of the proposed $\mathrm{GQ}(\sigma,\lambda)$. Theorem 1 shows that $\mathrm{GQ}(\sigma,\lambda)$ converges to its TD fixed-point (23) with probability one. Besides, Theorem 1 illustrates the structure of this TD fixed-point: it is the global asymptotically stable equilibrium of its corresponding ordinary differential equation (ODE). For more discussion, we provide in Remark 4. Theorem 2 shows that $\mathrm{GQ}(\sigma,\lambda)$ converges to an arbitrarily small neighborhood of the optimal solution with probability one. While we rely on the standard two-timescale ODE method for the general convergence mechanism, our novel theoretical step is verifying that the proposed σ-dependent trace coefficient $c_{t,\sigma}$ yields a stable equilibrium structure for the ODEs.

5.1. Additional Assumptions

We need some additional assumptions to present the convergent of $\mathrm{GQ}(\sigma,\lambda)$, those assumptions are widely used in reinforcement learning [13] [17]-[19].

Assumption 2 (Diminishing Step-size). *The positive sequences* $\{\alpha_t\}_{t\geq 0}$, $\{\beta_t\}_{t\geq 0}$ *satisfy the following conditions with probability one,*

$$\sum_{t=0}^{\infty}\alpha_t=\sum_{t=0}^{\infty}\beta_t=\infty,\ \sum_{t=0}^{\infty}\alpha_t^2<\infty,\sum_{t=0}^{\infty}\beta_t^2<\infty.$$

Assumption 3. *The features* $\{\phi_t\}_{t\geq 0}$ *is uniformly bounded by* $\phi_{\max}$. *The reward function is uniformly bounded by* $R_{\max}$. *The importance sampling* $\rho_t=\dfrac{\pi(A_t\mid S_t)}{\mu(A_t\mid S_t)}$ *is uniformly bounded by* $\rho_{\max}$.

Assumption 4 (Solvability of Problem). *The matrix* A_σ *is non-singular and* $\mathrm{rank}(\Phi)=p$.

Assumption 4 requires the non-singular matrix A_σ, which implies the optimal parameter $\theta^\star=-A^{-1}b$ is well defined. The feature matrix Φ has linearly independent columns implies the matrix $M=\Phi^\top\Xi\Phi$ is positive defined.

5.2. Main Results and Discussion

Theorem 1 (Convergence of Algorithm 1). *Under Assumption* 1-4, *we consider the iteration* $\{(\theta_t,\omega_t)\}_{t\geq 0}$ *that is generated according to* (30)-(31). *The step-size* α_t, β_t *satisfy Assumption* 2 *and* $\eta_t=\dfrac{\beta_t}{\alpha_t}\to 0$, *as* $t\to\infty$. *We define two functions* $G(\theta)$, $H(\omega,\theta)$ *as follows,*

$$G(\theta)=-A_\sigma^\top M^{-1}(A_\sigma\theta+b_\sigma),\tag{33}$$

$$H(\omega,\theta)=A_\sigma\theta+b_\sigma-M\omega.\tag{34}$$

Then $(\theta_t,\omega_t)\xrightarrow{w.p.1}(\theta_\star,\omega_\star)$, as $t\to\infty$, where $(\theta_\star,\omega_\star)$ is the unique global asymptotically stable equilibrium with respect to the following ODE correspondingly:

$$\begin{cases} \dot{\theta}(t) =: \dfrac{\mathrm{d}}{\mathrm{d}t}\theta(t) = G(\theta(t)) \\ \dot{\omega}(t) =: \dfrac{\mathrm{d}}{\mathrm{d}t}\omega(t) = H(\omega(t), \theta(t)), \end{cases} \tag{35}$$

where $\theta(t), \omega(t) \in \mathbb{R}^p$ are the functions are defined on continuous time $(0, \infty)$.

Proof. We provide its proof in **Appendix B**. □

Remark 4 (TD-Fixed Point of $\mathrm{GQ}(\sigma, \lambda)$). *Theorem* 1 *illustrates that the sequence* $\{(\theta_t, \omega_t)\}_{t \geq 0}$ *generated by* $\mathrm{GQ}(\sigma, \lambda)$ *converges to the global asymptotically stable equilibrium of its corresponding ODE* (35). *Furthermore, from the details of the proof in* ***Appendix B****, we know since* A_σ *is invertible and* M^{-1} *is positive definite, then* $A_\sigma^\top M^{-1} A_\sigma$ *is also positive defined. Because the ODE uses the negative MSPBE-gradient direction, the Jacobian at the equilibrium is* $-A_\sigma^\top M^{-1} A_\sigma$, *whose eigenvalues have negative real parts. So the following ODE*

$$\dot{\theta}(t) = -A_\sigma^\top M^{-1}\left(A_\sigma \theta(t) + b_\sigma\right) = G(\theta(t)), \tag{36}$$

has a unique global asymptotically stable equilibrium $\theta^\star$ satisfies the equation

$$A_\sigma \theta_\star + b_\sigma = 0, \tag{37}$$

which implies the global asymptotically stable equilibrium $\theta^\star$ of the ODE $\dot{\theta}(t) = G(\theta(t))$ is also the TD fixed point of (23). That means θ_t converges to its TD fixed point:

$$\theta_t \to \theta_\star = -A_\sigma^{-1} b_\sigma, \ \ w.p.1.$$

Remark 5 (Unification of TD-Fixed Point). *If* $\sigma = 0$, *the matrix* A_σ *is reduced to*

$$A_\sigma\big|_{\sigma=0} = \Phi^\top \Xi \left(I - \lambda\gamma P^{\pi,\mu}\right)^{-1}\left(\gamma P^\pi - I\right)\Phi,$$

which implies $\mathrm{GQ}(\sigma, \lambda)\big|_{\sigma=0}$ converges to the TD-fixed point of $\mathrm{GTB}(\lambda)$ [12], *i.e.*, $\mathrm{GQ}(\sigma, \lambda)\big|_{\sigma=0}$ shares the same TD-fixed point of $\mathrm{GTB}(\lambda)$ [12] as follows

$$\theta_\star = -A_\sigma^{-1} b_\sigma\big|_{\sigma=0}.$$

If $\sigma = 1$, then the key matrix A_σ is reduced to

$$A_\sigma\big|_{\sigma=1} = \Phi^\top \Xi \left(I - \lambda\gamma P^\pi\right)^{-1}\left(\gamma P^\pi - I\right)\Phi,$$

which illustrates the TD-fixed point of $\mathrm{GQ}(\sigma, \lambda)\big|_{\sigma=1}$ (*i.e.*, $\mathrm{GQ}(\lambda)$) algorithm as follows

$$\theta_\star = -A_\sigma^{-1} b_\sigma\big|_{\sigma=1}.$$

Theorem 1 presents an asymptotic result, which holds only in the limit as the number of iterations increases to infinity. Now, we present a result shows the distance between θ_t and the optimal solution $\theta^\star$ convergence to 0 in probability.

For any $T>0$, and $t\geq 0$, we introduce a notation

$$\kappa(t;T)=\min\left\{k\geq t\,|\,\sum_{i=t}^{k+1}\alpha_i>T\right\}$$

to denote the last iteration before the sum of step-size α_i between it and the t-th iteration exceeds T. Since we consider the Assumption 2, the notation $\kappa(t;T)$ is well-defined.

Theorem 2. *Under Assumption* 1-4, *we consider the iteration* $\{(\theta_t,\omega_t)\}_{t\geq 0}$ *generated by* (30)-(31). *The step-size* α_t, β_t *satisfy Assumption* 2. *Moreover, for the integers* $m_t\to\infty$, $m_t'\to\infty$, *as* $t\to\infty$,

$$\lim_{t\to\infty}\sup_{0<j\leq m_t}\left|\frac{\alpha_{t+j}}{\alpha_t}-1\right|=0,\lim_{t\to\infty}\sup_{0<j\leq m_t'}\left|\frac{\beta_{t+j}}{\beta_t}-1\right|=0.$$

Then there exists a sequence of positive numbers $T_t\to\infty$ (as $t\to\infty$) such that for any $\epsilon>0$,

$$\lim_{t\to\infty}\mathbb{P}\left[\theta_i\notin N_\epsilon(\theta^\star),\exists\, i\in\left[t,\kappa(t,T_t)\right]\right]=0, \tag{38}$$

where we use $N_\epsilon(\theta^\star)=\left\{\theta:\left\|\theta-\theta^\star\right\|_2\leq\epsilon\right\}$ to denote the ϵ-neighborhood of $\theta^\star$.

Proof. We provide its proof in **Appendix C**. □

Remark 6 *Theorem* 2 *shows that as* $t\to\infty$, *the sequence* $\{\theta_i\}_{i=t}^{\kappa(t,T_t)}$ *collects to an arbitrarily small neighborhood of the optimal solution* $\theta^\star$ *with probability one, i.e.,*

$$\lim_{t\to\infty}\mathbb{P}\left(\bigcap_{i=t}^{\kappa(t,T_t)}\left\{\theta_i\in N_\epsilon(\theta^\star)\right\}\right)=1.$$

In fact, the results (38) implies

$$\lim_{t\to\infty}\mathbb{P}\left(\theta_i\notin\mathcal{N}(\theta_\star,\epsilon),\exists\, i\in\left[t,\kappa(t,T_t)\right]\right)=0,$$

then we know

$$\begin{aligned}
&1-\lim_{t\to\infty}\mathbb{P}\left(\theta_i\notin\mathcal{N}(\theta_\star,\epsilon),\exists\, i\in\left[t,\kappa(t,T_t)\right]\right)\\
&=\lim_{t\to\infty}\mathbb{P}\left(\theta_i\notin\mathcal{N}(\theta_\star,\epsilon),\exists\, i\in\left[t,\kappa(t,T_t)\right]\right)\\
&=\lim_{t\to\infty}\mathbb{P}\left(\overline{\bigcup_{i=t}^{\kappa(t,T_t)}\left\{\theta_i\notin\mathcal{N}(\theta_\star,\epsilon)\right\}}\right)\\
&=\lim_{t\to\infty}\mathbb{P}\left(\bigcap_{i=t}^{\kappa(t,T_t)}\left\{\theta_i\in\mathcal{N}(\theta_\star,\epsilon)\right\}\right).
\end{aligned}$$

That is $\lim_{t\to\infty}\mathbb{P}\left(\bigcap_{i=t}^{\kappa(t,T_t)}\left\{\theta_i\in\mathcal{N}(\theta_\star,\epsilon)\right\}\right)=1$.

6. Experiments

In this section, we test the capacity of $\mathrm{GQ}(\sigma,\lambda)$ for two typical tasks: off-policy evaluation and control.

- For off-policy evaluation, we compare $\mathrm{GQ}(\sigma,\lambda)$ with four state-of-art algorithms: $\mathrm{GQ}(\lambda)$ [11], $\mathrm{ABQ}(\zeta)$ [20], $\mathrm{GTB}(\lambda)$, $\mathrm{GRetrace}(\lambda)$ [12] over

two typical measurements: MSPBE and mean square error (MSE). It is worth noting that if $\sigma=0$, $\mathrm{GQ}(\sigma,\lambda)$ is reduced to a new version of $\mathrm{GTB}(\lambda)$ (we denote it as $\mathrm{GTB}(\lambda)$-v0), thus, we also show its comparison to $\mathrm{GTB}(\lambda)$ [12] (we denote it as $\mathrm{GTB}(\lambda)$-v1).

- For the control task, our goal is to show the the trade-off between $\sigma=0$ and $\sigma=1$. Empirical results show the $\mathrm{GQ}(\sigma,\lambda)$ with a value $\sigma\in(0,1)$ that creates a mixture of $\mathrm{GQ}(\lambda)$ and gradient Tree Backup(λ) achieves a better performance than both the extreme end $\sigma=0$ and $\sigma=1$.

These domains are standard diagnostic benchmarks intended to isolate the off-policy instability under linear function approximation. Scaling to very high-dimensional environments is left as future work.

6.1. Off-Policy Evaluation Experiments

We present the domains in the experiments as follows.

(1) **Two State MDP** [12]. This MDP is shown in **Figure 1**, the behavior policy $\mu(\text{right}|\cdot)=0.5$, and target policy $\pi(\text{right}|\cdot)=1$. We assign the features $\left\{(1,0)^{\top},(2,0)^{\top},(0,1)^{\top},(0,2)^{\top}\right\}$ to the state-action pairs $\left\{(s_1,\text{right}),(s_2,\text{right}),(s_1,\text{left}),(s_2,\text{left})\right\}$, *i.e.*,

$$\Phi=\begin{pmatrix}1&2&0&0\\0&0&1&2\end{pmatrix}^{\top}.$$

(2) **Baird Star** [21]. The Baird Star is an episodic seven states MDP with two actions: dashed action and solid action. In this example, the behavior policy $\mu(\cdot|\text{dashed})=\frac{6}{7}$, $\mu(\cdot|\text{solid})=\frac{1}{7}$ and target policy $\pi(\cdot|\text{solid})=1$. We choose the feature map matrix as follows

$$\Phi=\begin{pmatrix}2\boldsymbol{I}_{7\times7}&\mathbf{1}_{7\times1}&\mathbf{0}_{7\times8}\\\mathbf{0}_{7\times8}&2\boldsymbol{I}_{7\times7}&\mathbf{1}_{7\times1}\end{pmatrix},$$

where $\boldsymbol{I}$ denotes the identity matrix, $\mathbf{0}$ denotes a matrix whose elements are all 0, and $\mathbf{1}_{7\times1}$ denotes a vector whose elements are all 1. We used $\theta_0=(1,1,1,1,1,1,10,1)^{\mathrm{T}}$ as initial parameter vector for the methods that allow specifying a start estimate, TD-learning is known to diverge for this initialization of the parameter-vector [3] [22].

(3) **Cliff Walking**. This is a standard undiscounted, episodic task, with start and goal states, and the usual actions causing movement up, down, right, and left.

(4) **Grid World**. This environment from ([3], Chapter 4), where the agent on an 4 × 4 grid and your goal is to reach the terminal state at the top left or the bottom right corner.

(5) **Windy Gridworld**. This environment from ([3], Chapter 6). Windy Gridworld problem for reinforcement learning. Actions include going up, down, right, and left. In each column the wind pushes you up a specific number of steps (for the next action). If an action would take you off the grid, you remain in the previ-

ous state. For each step you get a reward of −1, until the agent reach into a terminal state.

We summarize the domains, feature settings, target policy and behavior policy below.

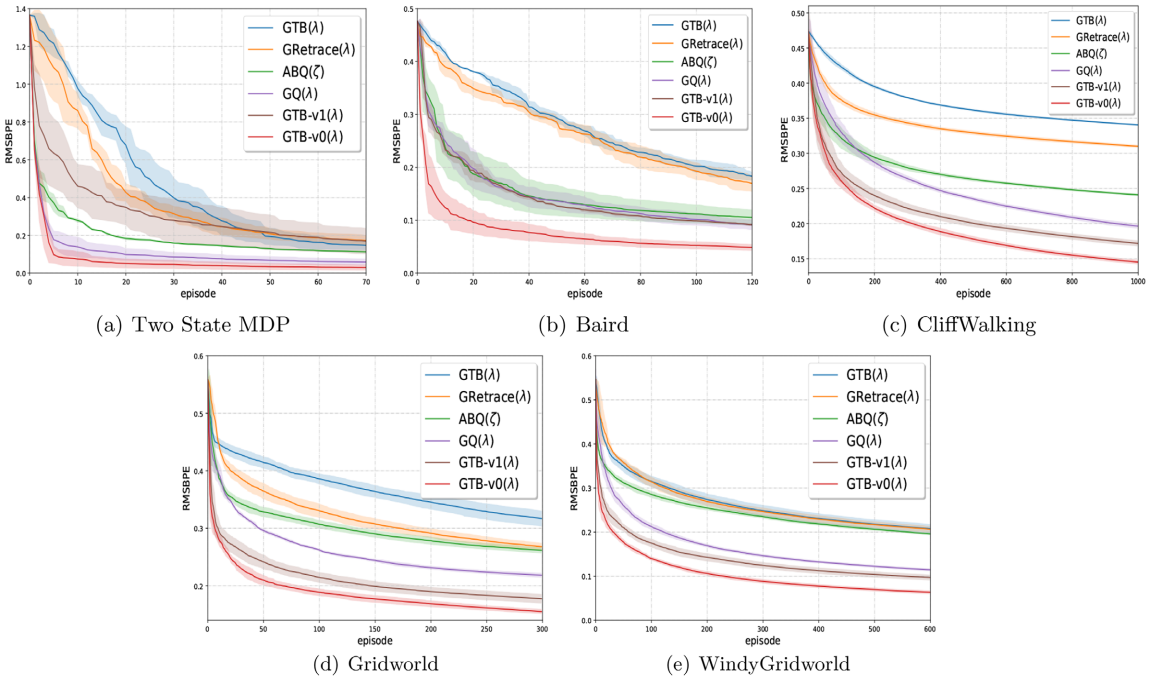
Two Measurements for Off-Policy Evaluation. In this section, we use empirical $\text{RMSPBE} = \frac{1}{2}\left\|\hat{b} + \hat{A}\theta\right\|^2_{\hat{M}^{-1}}$ to evaluate the performance, where we evaluate $\hat{A}$, $\hat{b}$, and $\hat{M}$ according to their unbiased estimators. Additionally, we also compare the performance over a common measurement empirical MSE: $\text{MSE} = \left\|\Phi\theta - q^{\pi}\right\|^2_{\Xi}$, where q^{π} is estimated by simulating the target policy π and averaging the discounted cumulative rewards over trajectories.

Hyper-parameter Setting. We run the hyper parameter σ as follows: σ ranges from 0 to 1 with step of 0.02, *i.e.*,

$$\Sigma = \{0, 0.02, 0.04, \cdots, 0.98, 1.0\}.$$

It collects 51 results w.r.t. $\text{GQ}(\sigma, \lambda)$. We set $\lambda = 0.99$, $\gamma = 0.99$, and run the step-size $\alpha_t = \{10^{-2}, 2\times10^{-2}, 10^{-3}, 2\times10^{-3}\}$, $\eta_t = \beta_t / \alpha_t = \{2^0, 2^{-1}, \cdots, 2^{-10}\}$.

Results Report. All the results shown in Figure 2 and Figure 3 are the average of 5 runs, where we choose the best σ among the 51 results w.r.t. the space Σ The results the proposed $\text{GQ}(\sigma, \lambda)$ outperforms the the baseline algorithms. The results of Figure 2 and Figure 3 also show $\text{GQ}(\sigma, \lambda)$ with an intermediate σ (between 0 and 1) has a better performance than the extreme case ($\sigma = 0$ and 1). This experiment further validates that unifying some existing algorithms can create a better algorithm.



(a) Two State MDP (b) Baird (c) CliffWalking

(d) Gridworld (e) WindyGridworld

Figure 2. RMSPBE comparison with different baseline algorithms.

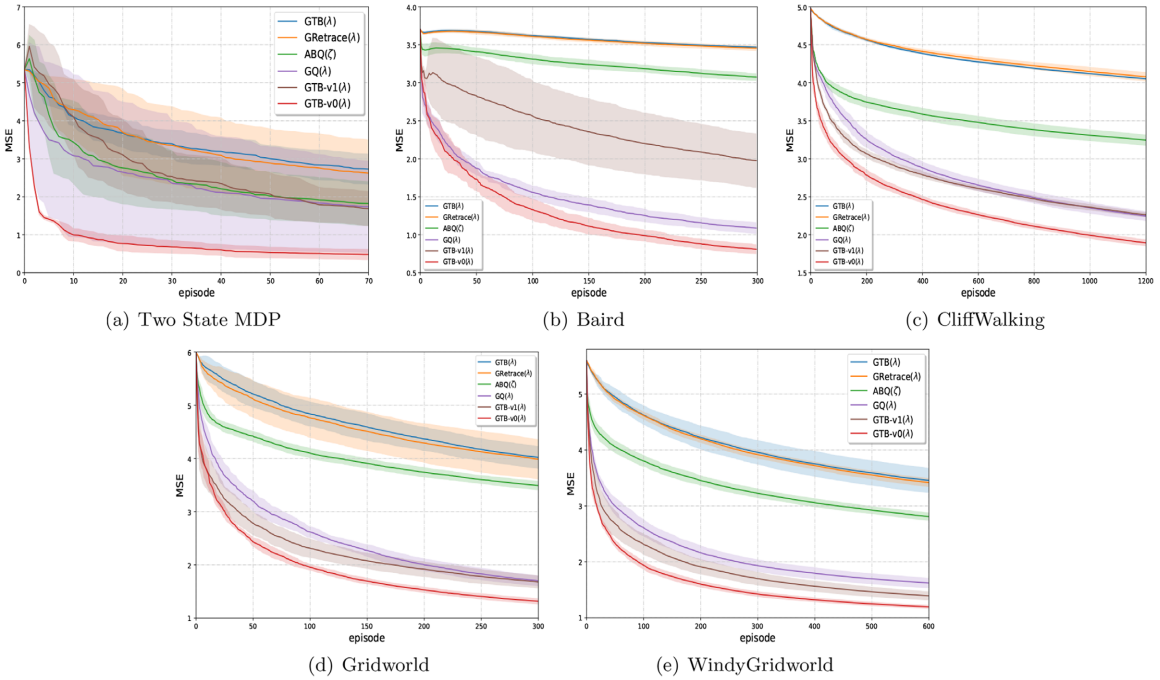


Figure 3. MSE comparison with with different baseline algorithms.

6.2. Control Domain: Off-Policy Evaluation

In this section, we test the off-policy evaluation behavior of $\mathrm{GQ}(\sigma,\lambda)$ algorithm on mountain car domain, where the agent considers the task of driving an under-powered car up a steep mountain road. The agent receives a reward of -1 at every step until it reaches the goal region at the top of the hill. Since the state space of this domain is continuous, we use the open tile coding software[2] to extract feature of states. Recall the states and actions of *MountainCar*:

$$\mathcal{S}=\{(\text{Velocity},\text{Position})\}=[-0.07,0.07]\times[-1.2,0.6],$$

$$\mathcal{A}=\{\text{left},\text{neutral},\text{right}\}.$$

In this experiment, if $\text{Velocity}>0$, we use behavior policy

$$\mu=\left(\frac{1}{100},\frac{1}{100},\frac{98}{100}\right),\pi=\left(\frac{1}{10},\frac{1}{10},\frac{8}{10}\right);$$

else

$$\mu=\left(\frac{98}{100},\frac{1}{100},\frac{1}{100}\right),\pi=\left(\frac{8}{10},\frac{1}{10},\frac{1}{10}\right).$$

Note that the target policy π is fixed throughout. Thus, this experiment acts as an off-policy evaluation within the Mountain Car domain. In this experiment, we set the number of tilings to be 4, and there are no white noise features. As suggested by Sutton and Barto [3], we set all the initial parameters to be 0, which is optimistic about causing extensive exploration.

[2] http://incompleteideas.net/rlai.cs.ualberta.ca/RLAI/RLtoolkit/tilecoding.html

As before, we run the hyper parameter σ as follows: σ ranges from 0 to 1 with step of 0.02. It collects 51 results w.r.t. $\mathrm{GQ}(\sigma,\lambda)$. We also set $\lambda=0.99$, $\gamma=0.99$, and run the step-size $\alpha_t=\{10^{-2},2\times10^{-2},10^{-3},2\times10^{-3}\}$, $\eta_t=\beta_t/\alpha_t=\{2^0,2^{-1},\cdots,2^{-10}\}$, the dimension of feature $p=\{512,1024,2048\}$.

Overall Presentation. We give more comprehensive results of the trade-off between $\sigma=0$ and $\sigma=1$. We statistic of the number of σ happens for the following three case:

- **(I)** $\mathrm{GQ}(\sigma,\lambda)$ performs better than both $\sigma=0$ and $\sigma=1$.
- **(II)** $\mathrm{GQ}(\sigma,\lambda)$ performs better than $\sigma=0$ or $\sigma=1$.
- **(III)** $\mathrm{GQ}(\sigma,\lambda)$ performs worse than $\sigma=0$ and $\sigma=1$.

The setting of σ is the same as the previous section, and the total number of σ reaches 51.

Table 1. Returns under various parameters.

Case		$\sigma=0$	$\sigma=1$	$\sigma=0.1$	$\sigma=0.2$	$\sigma=0.3$	$\sigma=0.4$	$\sigma=0.5$	$\sigma=0.6$	$\sigma=0.7$	$\sigma=0.8$	$\sigma=0.9$
$\alpha=0.001$	$p=512$	−122.0	−123.6	− 123.5	**−119.8**	**−117.4**	**−120.8**	**−119.1**	−120.5	**−119.2**	**−119.3**	**−120.1**
	$p=1024$	−119.9	−121.9	−127.8	−120.2	−122.4	−122.0	**−118.4**	−121.6	−121.6	−124.0	−120.1
	$p=2048$	−122.3	−121.4	−122.6	−121.9	**−120.1**	−122.3	−122.6	**−119.6**	**−120.9**	−121.4	−122.5
$\alpha=0.002$	$p=512$	−126.9	−124.2	**−124.2**	−127.5	−125.3	−125.2	**−124.0**	**−121.6**	−125.4	−125.8	**−120.2**
	$p=1024$	−124.1	−123.1	−122.4	−126.3	−122.3	−123.0	−121.4	−126.2	**−121.3**	−123.6	−126.0
	$p=2048$	−124.8	−124.0	−126.4	−124.6	**−122.7**	**−123.9**	−125.1	**−122.5**	**−123.6**	−126.4	**−122.7**

Table 2. Percentage under various parameters.

Case		$p=512$	$p=1024$	$p=2048$
$\alpha=0.001$	I	**69.8%**	20.4%	**55.1%**
	II	14.1%	36.7%	20.4%
	III	16.1%	**42.9%**	24.5%
$\alpha=0.002$	I	**53.1%**	6.1%	**49.0%**
	II	38.8%	16.3%	18.4%
	III	8.1%	**77.6%**	32.6%

Results Report. The results shown in Figure 2 and Table 1 are average of 5 runs, and each run contains 400 episodes. The result in Figure 2 shows that $\mathrm{GQ}(\sigma,\lambda)$ with an intermediate σ (between 0 and 1) has a better performance than the extreme case ($\sigma=0$ and 1). This experiment further validates that unifying some existing algorithms can create a better algorithm for reinforcement. Table 1 shows the returns mountaincar reaches the goal region at the top of the hill. As shown in Table 2, those results also show $\mathrm{GQ}(\sigma,\lambda)$ achieves the best performance at a $\sigma\in(0,1)$, which implies the trade-off between $\sigma=0$ and $\sigma=1$. That is to see the $\mathrm{GQ}(\sigma,\lambda)$ with a value $\sigma\in(0,1)$ that creates a mix-

ture of $GQ(\lambda)$ and gradient Tree Backup(λ) achieves a better performance than both the extreme end $\sigma=0$ and $\sigma=1$.

7. Conclusion

In this paper, we extend tabular $Q(\sigma,\lambda)$ with function approximation, and propose $GQ(\sigma,\lambda)$. We analyze the convergence of $GQ(\sigma,\lambda)$. Our theory analysis shows that $GQ(\sigma,\lambda)$ converges to its TD fixed-point with probability one. Then, we show $GQ(\sigma,\lambda)$ converges to the optimal solution of the minimizing MSPBE problem. Finally, we conduct experiments on some standard domains to confirm the effectiveness of the proposed $GQ(\sigma,\lambda)$. Result show that the best performance of $GQ(\sigma,\lambda)$ achieved with a $\sigma\in(0,1)$, neither $\sigma=0$, nor $\sigma=1$. Extending this framework to non-linear function approximation is an important future direction. Such an extension faces theoretical hurdles, such as the parameter-dependence of the Jacobian and the loss of a fixed linear least-squares structure, which may require techniques like target networks or compatible gradients to stabilize.

Funding

The project was partially supported by Scientific Research Fund of Zhejiang Provincial Education Department under Grant No. Y202456228.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sutton, R.S. (1988) Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, **3**, 9-44. https://doi.org/10.1023/a:1022633531479

[2] De Asis, K., Hernandez-Garcia, J., Holland, G. and Sutton, R. (2018) Multi-Step Reinforcement Learning: A Unifying Algorithm. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 2902-2909. https://doi.org/10.1609/aaai.v32i1.11631

[3] Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction. MIT Press.

[4] Rummery, G.A. and Niranjan, M. (1994) Online Q-Learning Using Connectionist Systems, Volume 37. University of Cambridge, Department of Engineering.

[5] Precup, D., Sutton, R.S., Singh, S.P., *et al.* (2000) Eligibility Traces for Off-Policy Policy Evaluation. *Proceedings of the Seventeenth International Conference on Machine Learning*, Standord, 29 June-2 July 2000, 759-766.

[6] Yang, L., Shi, M., Zheng, Q., Meng, W. and Pan, G. (2018) A Unified Approach for Multi-Step Temporal-Difference Learning with Eligibility Traces in Reinforcement Learning. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 2984-2990. https://doi.org/10.24963/ijcai.2018/414

[7] De Asis, K. (2018) A Unified View of Multi-Step Temporal Difference Learning. Ph.D. Thesis, University of Alberta Edmonton.

[8] Sutton, R.S., Mahmood, A.R. and White, M. (2016) An Emphatic Approach to the Problem of Off-Policy Temporal-Difference Learning. *Journal of Machine Learning Research*, **17**, 2603-2631.

[9] Sutton, R.S., Maei, H.R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, C., *et al.* (2009) Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. *Proceedings of the* 26*th Annual International Conference on Machine Learning*, Montreal, 14-18 June 2009, 993-1000. https://doi.org/10.1145/1553374.1553501

[10] Maei, H.R., Szepesvári, C., Bhatnagar, S., Precup, D., Silver, D. and Sutton, R.S. (2009) Convergent Temporal-Difference Learning with Arbitrary Smooth Function Approximation. *Proceedings of the* 23*rd International Conference on Neural Information Processing Systems*, Vancouver, 7-10 December 2009, 1204-1212.

[11] Maei, H.R. and Sutton, R.S. (2010) GQ(λ): A General Gradient Algorithm for Temporal-Difference Prediction Learning with Eligibility Traces. *Proceedings of the* 3*rd Conference on Artificial General Intelligence* (*AGI*-10), Lugano, 5-8 March 2010, 100-105. https://doi.org/10.2991/agi.2010.22

[12] Touati, A., Bacon, P.L., Precup, D. and Vincent, P. (2018) Convergent Tree-Backup and Retrace with Function Approximation. arXiv: 1705.09322.

[13] Yang, L., Zheng, G., Zhang, Y., Zheng, Q., Li, P. and Pan, G. (2021) On Convergence of Gradient Expected Sarsa(λ). *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 10621-10629. https://doi.org/10.1609/aaai.v35i12.17270

[14] Bertsekas, D.P. and Tsitsiklis, J.N. (1996) Neuro-Dynamic Programming, Volume 5. Athena Scientific.

[15] Liu, B., Liu, J., Ghavamzadeh, M., Mahadevan, S. and Petrik, M. (2015) Finite-Sample Analysis of Proximal Gradient TD Algorithms. arXiv: 2006.14364.

[16] Dalal, G., Szorenyi, B., Thoppe, G. and Mannor, S. (2018) Finite Sample Analysis of Two-Timescale Stochastic Approximation with Applications to Reinforcement Learning. arXiv: 1703.05376.

[17] Wang, Y., Chen, W., Liu, Y.T., Ma, Z.M. and Liu, T.Y. (2017) Finite Sample Analysis of the GTD Policy Evaluation Algorithms in Markov Setting. arXiv: 1809.08926.

[18] Bhandari, J., Russo, D. and Singal, R. (2018) A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation. arXiv: 1806.02450.

[19] Xu, T.Y., Zou, S.F. and Liang, Y.B. (2019) Two Time-Scale Off-Policy TD Learning: Non-Asymptotic Analysis over Markovian Samples. arXiv: 1909.11907.

[20] Mahmood, A.R., Yu, H. and Sutton, R.S. (2017) Multi-Step Off-Policy Learning without Importance Sampling Ratios. arXiv: 1702.03006.

[21] Baird, L. (1995) Residual Algorithms: Reinforcement Learning with Function Approximation. In: Prieditis, A. and Russell, S., Eds., *Machine Learning Proceedings* 1995, Elsevier, 30-37. https://doi.org/10.1016/b978-1-55860-377-6.50013-x

[22] Dann, C., Neumann, G. and Peters, J. (2014) Policy Evaluation with Temporal Differences: A Survey and Comparison. *The Journal of Machine Learning Research*, **15**, 809-883.

[23] Borkar, V.S. (1997) Stochastic Approximation with Two Time Scales. *Systems & Control Letters*, **29**, 291-294. https://doi.org/10.1016/s0167-6911(97)90015-3

[24] Kushner, H. and Yin, G.G. (2003) Stochastic Approximation and Recursive Algorithms and Applications, Volume 35. Springer Science & Business Media.

[25] Yu, H.Z. (2016) Weak Convergence Properties of Constrained Emphatic Temporal-

Difference Learning with Constant and Slowly Diminishing Stepsize. *Journal of Machine Learning Research*, **17**, 7745-7802.

[26] Yu, H.Z. (2017) On Convergence of Some Gradient-Based Temporal-Differences Algorithms for Off-Policy Learning. arXiv: 1712.09652.

Appendix

A. Derivation of (28)-(29)

Proof. Let us calculate $\mathrm{MSPBE}(\theta,\lambda)$ directly,

$$\begin{aligned}
-\frac{1}{2}J(\theta_t) &= -\frac{1}{2}\nabla_\theta \mathrm{MSPBE}(\theta,\lambda)\big|_{\theta=\theta_t} \\
&= -\frac{1}{2}\nabla_\theta\left(\mathbb{E}\left[\delta_t e_{t,\sigma}\right]^\top \mathbb{E}\left[\phi_t\phi_t^\top\right]^{-1}\mathbb{E}\left[\delta_t e_{t,\sigma}\right]\right) \\
&= -\left(\nabla_\theta \mathbb{E}\left[\delta_t e_{t,\sigma}\right]^\top\right)\mathbb{E}\left[\phi_t\phi_t^\top\right]^{-1}\mathbb{E}\left[\delta_t e_{t,\sigma}\right] \\
&= -\mathbb{E}\left[\left(\gamma\mathbb{E}_\pi\phi(S_{t+1},\cdot)-\phi_t\right)e_{t,\sigma}^\top\right]\mathbb{E}\left[\phi_t\phi_t^\top\right]^{-1}\mathbb{E}\left[\delta_t e_{t,\sigma}\right] \\
&= -\mathbb{E}\left[\gamma\mathbb{E}_\pi\phi(S_{t+1},\cdot)e_{t,\sigma}^\top-\phi_t e_{t,\sigma}^\top\right]\mathbb{E}\left[\phi_t\phi_t^\top\right]^{-1}\mathbb{E}\left[\delta_t e_{t,\sigma}\right] \\
&= \mathbb{E}\left[\phi_t\phi_t^\top+\phi_{t+1}\gamma\lambda e_{t,\sigma}^\top-\gamma\mathbb{E}_\pi\phi(S_{t+1},\cdot)e_{t,\sigma}^\top\right]\underbrace{\mathbb{E}\left[\phi_t\phi_t^\top\right]^{-1}\mathbb{E}\left[\delta_t e_{t,\sigma}\right]}_{=:\varpi} \\
&= \mathbb{E}\left[\delta_t^{\mathrm{ES}} e_{t,\sigma}\right]-\mathbb{E}\left[\gamma(1-\lambda)\bar{\phi}_{t+1}e_{t,\sigma}^\top\right]\varpi.
\end{aligned} \tag{39}$$

□

B. Proof of Theorem 1

The ODE method (see Lemma 1) is our is our main tool to prove Theorem 1. We refer the reader to that reference for further technical details.

Lemma 1 ([25]). *For the stochastic recursion of* x_n, y_n *given by*

$$x_{n+1} = x_n + a_n\left[g(x_n,y_n)+M_{n+1}^{(1)}\right], \tag{40}$$

$$y_{n+1} = y_n + b_n\left[h(x_n,y_n)+M_{n+1}^{(2)}\right], n\in\mathbb{N} \tag{41}$$

if the following assumptions are satisfied:

- (A1) *Step-sizes* $\{a_n\},\{b_n\}$ *are positive, satisfying*

$$\sum_n a_n = \sum_n b_n = \infty, \sum_n a_n^2 + b_n^2 < \infty, \frac{b_n}{a_n}\to 0 \text{ as } n\to\infty.$$

- (A2) *The map* $g:\mathbb{R}^{d+k}\to\mathbb{R}^d, h:\mathbb{R}^{d+k}\to\mathbb{R}^k$ *are Lipschitz.*
- (A3) *The sequence* $\left\{M_{n+1}^{(1)}\right\}_{n\in\mathbb{N}}, \left\{M_{n+1}^{(2)}\right\}_{n\in\mathbb{N}}$ *are martingale difference sequences w.r.t. the increasing* σ *-fields* $\mathcal{F}_n \overset{\mathrm{def}}{=} \sigma\left(x_m, y_m, M_m^{(1)}, M_m^{(2)}, m\le n\right), n\in\mathbb{N}$, *satisfying*

$$\mathbb{E}\left[M_{n+1}^{(i)} \mid \mathcal{F}_n\right] = 0, i = 1,2, n\in\mathbb{N}.$$

Furthermore, $\left\{M_{n+1}^{(i)}\right\}_{n\in\mathbb{N}}, i=1,2$, *are square-integrable with*

$$\mathbb{E}\left[\left\|M_{n+1}^{(i)}\right\|^2 \mid \mathcal{F}_n\right] \le K\left(1+\|x_n\|^2+\|y_n\|^2\right),$$

for some constant $K>0$.

- (A4) *For each* $x\in\mathbb{R}^d$, *the o.d.e.*

$$\dot{y}(t)=h(x,y(t))$$

has a global asymptotically stable equilibrium $\Omega(x)$ *such that*:

$\Omega(x):\mathbb{R}^d\to\mathbb{R}^k$ *is Lipschitz.*

- (A5) *The o.d.e.*

$$\dot{x}(t)=g(x(t),\Omega(x(t)))$$

has a global asymptotically stable equilibrium x^*.

Then, the iterates (40), (41) *converge to* $(x^*,\Omega(x^*))$ *a.s. on the set*

$Q\overset{\text{def}}{=}\{\sup_n x_n<\infty,\sup_n y_n<\infty\}$.

Proof. Now, we apply Lemma 1 to prove results.

Step 1: On Convergence of ω_t.

We consider the following ODE:

$$\dot{\theta}(t)=0,\ \dot{\omega}(t)=\mathbb{E}\left[\delta_t^{\mathrm{ES}}e_t\mid\theta(t)\right]-M\omega(t).$$

The equation $\dot{\theta}(t)=0$ implies there exists a constant vector θ such that: $\theta(t)=\theta$, thus we can rewrite the above ODE associated $\omega(t)$ as follows,

$$\dot{\omega}(t)=\mathbb{E}\left[\delta_t^{\mathrm{ES}}e_t\mid\theta\right]-M\omega(t)=A_\sigma\theta+b_\sigma-M\omega(t).\tag{42}$$

Then, for any given θ,

$$\omega_\star=M^{-1}(A_\sigma\theta+b_\sigma)$$

is the unique globally asymptotically stable equilibrium for the ODE (42). Let

$$H(\omega,\theta)=A_\sigma\theta+b_\sigma-M\omega,$$

for a fixed θ, let

$$H_\infty(\omega,\theta)=\lim_{r\to\infty}\frac{H(r\omega(t),\theta)}{r}=-M\omega(t).$$

Since M is a positive definite matrix, then 0 is a globally asymptotically stable equilibrium for the following ODE

$$\dot{\omega}(t)=H_\infty(\omega(t),\theta).$$

Let the σ-field $\mathcal{F}_t=\sigma(\{\theta_k,\omega_k,\phi_k,R_k\}_{k<t})$ be generated by the set $\{\theta_k,\omega_k,\phi_k,R_k\}_{k<t}$, where $t\geq 1$. Let

$$M_{t+1}=\delta_t^{\mathrm{ES}}e_t-\phi_t\omega_t^{\top}\phi_t-\mathbb{E}\left[\delta_t^{\mathrm{ES}}e_t-\phi_t\omega_t^{\top}\phi_t\mid\mathcal{F}_t\right],$$

then for each $t\geq 0$, we have $\mathbb{E}[M_{t+1}\mid\mathcal{F}_t]=0$. Furthermore, since Assumption 3 holds, there exists a non-negative constant $K_1>0$, s.t. $\{M_t\}_{t\in\mathbb{N}}$ is square-integrable with

$$\mathbb{E}\left[\|M_t\|^2\mid\mathcal{F}_t\right]\leq K_1\left(1+\|\theta_t\|^2+\|\omega_t\|^2\right).$$

Since then, we have verified the conditions (A2)-(A5) of Lemma 1, thus

$$\omega_k-\omega_\star\to 0,\ w.p.1$$

where $w.p.1$ is short for with probability one.

Step 2: On Convergence of θ_k.

Let the σ-field $\mathcal{G}_t = \sigma\left(\{\theta_k, \phi_k, R_k\}_{k<t}\right)$ be generated by the set $\{\omega_k, \phi_k, R_k\}_{k<t}$, where $t \geq 1$. The iteration (31) that can be rewritten as:

$$\theta_{t+1} = \theta_t + \alpha_t\left(\delta_t^{\mathrm{ES}} e_t - \gamma(1-\lambda)\bar{\phi}_{t+1} e_t^{\top} M^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \mid \theta_t\right]\right).$$

Furthermore, we define a random variable N_t as follows,

$$\begin{aligned} N_t = {} & \delta_t^{\mathrm{ES}} e_t - \gamma(1-\lambda)\bar{\phi}_{t+1} e_t^{\top} M^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \mid \theta_t\right] \\ & - \mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t - \gamma(1-\lambda)\bar{\phi}_{t+1} e_t^{\top} M^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \mid \theta_t\right] \mid \mathcal{G}_t\right], \end{aligned}$$

then for each $t \geq 0$, we have $\mathbb{E}\left[N_{t+1} \mid \mathcal{G}_t\right] = 0$. Now, we rewrite N_t as follows,

$$\begin{aligned} N_t = {} & \delta_t^{\mathrm{ES}} e_t - \gamma(1-\lambda)\bar{\phi}_{t+1} e_t^{\top} M^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \theta_t\right] - \mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \theta_t\right] \\ & + \gamma(1-\lambda)\mathbb{E}\left[\bar{\phi}_{t+1} e_t^{\top}\right] M^{-1}\mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \mid \theta_t\right]. \end{aligned}$$

Under Assumption 3, for each $t \geq 0$, there exists a non-negative constant $K_2 > 0$ such that,

$$\mathbb{E}\left[\|N_t\|^2 \mid \mathcal{G}_t\right] \leq K_2\left(1 + \|\theta_t\|^2\right).$$

We consider the iteration (31) associated with the ODE

$$\dot{\theta}(t) = \left(I - \gamma(1-\lambda)\mathbb{E}\left[\bar{\phi}_{t+1} e_t^{\top}\right] M^{-1}\right)\mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t \mid \theta(t)\right]. \tag{43}$$

Recall $M = \mathbb{E}\left[\phi_t \phi_t^{\top}\right]$, from Equation (28), the ODE (43) can be rewritten as follows,

$$\dot{\theta}(t) = -A_{\sigma}^{\top} M^{-1}\left(A_{\sigma}\theta(t) + b_{\sigma}\right) \overset{(33)}{=} G\left(\theta(t)\right). \tag{44}$$

Since A_{σ} is invertible, then $\theta_{\star} = -A_{\sigma}^{-1} b_{\sigma}$ is the unique global asymptotically stable equilibrium of ODE (44). Let

$$G_{\infty}(\theta) = \lim_{r\to\infty} \frac{G(r\theta, \omega)}{r} = -A_{\sigma}^{\top} M^{-1} A_{\sigma}\theta.$$

We consider the following ODE

$$\dot{\theta}(t) = G_{\infty}\left(\theta(t)\right) = -A_{\sigma}^{\top} M^{-1} A_{\sigma}\theta(t). \tag{45}$$

Since A_{σ} is invertible and M^{-1} is positive definite, then $A_{\sigma}^{\top} M^{-1} A_{\sigma}$ is a positive defined matrix. Equivalently, $-A_{\sigma}^{\top} M^{-1} A_{\sigma}$ is negative definite. Thus the vector 0 is the unique global asymptotically stable equilibrium of (45). According to Lemma 1,

$$\theta_k - \theta_{\star} \to 0, \ w.p.1.$$

Therefore the proof is completed. □

C. Proof of Theorem 2

Proof. Let $\zeta_t = \left(e_t, S_t, A_t, S_{t+1}\right) \in \mathbb{R}^p \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, then we rewrite the iteration (31)

and (30) as follows

$$\theta_{t+1} = \theta_t + \alpha_t \left(g\left(\theta_t, \omega_t, \zeta_t\right) + e_t \iota_t \right),$$

$$\omega_{t+1} = \omega_t + \beta_t \left(h\left(\theta_t, \omega_t, \zeta_t\right) + e_t \nu_t \right),$$

where

$$g\left(\theta_t, \omega_t, \zeta_t\right) = \mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t - \gamma\left(1-\lambda\right) \mathbb{E}_\pi \left[\phi\left(S_{t+1,\cdot}\right) \right] e_t^\top \omega_t \right],$$

$$e_t \iota_t = \delta_t^{\mathrm{ES}} e_t - \gamma\left(1-\lambda\right) \mathbb{E}_\pi \left[\phi\left(S_{t+1,\cdot}\right) \right] e_t^\top \omega_t - g\left(\theta_t, \omega_t, \zeta_t\right)$$

satisfies $\mathbb{E}\left[e_t \iota_t\right] = 0$; $h\left(\theta_t, \omega_t, \zeta_t\right) = \mathbb{E}\left[\delta_t^{\mathrm{ES}} e_t - \phi_t \omega_t^\top \phi_t \right]$ and $e_t \nu_t = \delta_t^{\mathrm{ES}} e_t - \phi_t \omega_t^\top \phi_t - h\left(\theta_t, \omega_t, \zeta_t\right)$.

Now, we apply ([26], Theorem 2.3 of Chapter 8) twice, once for each time-scale. We refer the reader to that reference for further technical details in ([26], Theorem 2.3 of Chapter 8), which requires us to verify the following conditions of **(i)**-**(iv)**:

(i) *The random variables* $g\left(\theta_t, \omega_t, \zeta_t\right)$ *and* $h\left(\theta_t, \omega_t, \zeta_t\right)$ *are uniformly integrable* (*UI*), *i.e.*,

$$\lim_{a\to\infty} \sup_{t\in\mathbb{N}} \mathbb{E}\left[\left\| g\left(\theta_t, \omega_t, \zeta_t\right) \right\| \mathbb{I}\left\{ \left\| g\left(\theta_t, \omega_t, \zeta_t\right) \right\| \geq a \right\} \right] = 0,$$

$$\lim_{a\to\infty} \sup_{t\in\mathbb{N}} \mathbb{E}\left[\left\| h\left(\theta_t, \omega_t, \zeta_t\right) \right\| \mathbb{I}\left\{ \left\| h\left(\theta_t, \omega_t, \zeta_t\right) \right\| \geq a \right\} \right] = 0,$$

where $\mathbb{I}\{\cdot\}$ is the indicator function.

In fact, the uniform integrability of both $g\left(\theta_t, \omega_t, \zeta_t\right)$ and $h\left(\theta_t, \omega_t, \zeta_t\right)$ are ensured by the the UI property of $e_t = \lambda\gamma c_{t,\sigma} e_{t-1} + \phi_t$, and according to the same analysis of Proposition 2 from [27], we have $g\left(\theta_t, \omega_t, \zeta_t\right)$ and $h\left(\theta_t, \omega_t, \zeta_t\right)$ are UI.

(ii) *The set of random variables* $\{\zeta_t\}$ *is tight, i.e, for each positive scalar* δ, *there exists a compact set* $D_\delta \in \mathbb{R}^p \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ *such that*

$$\inf_{t\in\mathbb{N}} \mathbb{P}\left[\zeta_t \in D_\delta \right] \geq 1 - \delta.$$

In fact, recall the Assumption 3 implies the trace vector e_t satisfies $\sup_{t\in\mathbb{N}} \left\| e_t \right\| < \infty$, *i.e.*,

$$\begin{aligned} \left\| e_t \right\|_2^2 &= \left\| \sum_{k=0}^{t} \left(\gamma\lambda\right)^{t-k} \prod_{i=k+1}^{t} c_{i,\sigma} \phi_k \right\|_2^2 \\ &\leq \frac{\phi_{\max}^2}{1 - \left(\gamma\lambda \left(\left(1-\sigma\right) + \sigma\rho_{\max} \right) \right)^2}. \end{aligned} \tag{46}$$

For each $a > 0$, according to Markov inequality, we have

$$\mathbb{P}\left[\left\| e_t \right\| \geq a \right] \leq \frac{\sup_{t\in\mathbb{N}} \left\| e_t \right\|}{a} \to 0, \ a \to \infty,$$

which implies $\{e_t\}$ is tight. Since we consider the MDPs with finite state space and action space, so the sequence $\left\{ \zeta_t = \left(e_t, S_t, A_t, S_{t+1}\right) \right\}_{t\geq 0}$ is also tight.

(iii) *Let* $D_\zeta \in \mathbb{R}^p \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ *be a compact set, for each* $\zeta \in D_\zeta$, *both* $g(\theta,\omega,\zeta_0)$ *and* $h(\theta,\omega,\zeta_0)$ *are continuous with respect to* (θ,ω).

Recall the definitions of $g(\theta,\omega,\zeta_0)$ and $h(\theta,\omega,\zeta_0)$, then we have

$$g(\theta,\omega,\zeta_0) = A_\sigma^\top M^{-1}(A_\sigma\theta + b_\sigma),$$

$$h(\theta,\omega,\zeta_0) = A_\sigma\theta + b_\sigma - M\omega.$$

The Assumption 3 implies A_σ, M, and b_σ are bounded, thus both $g(\theta,\omega,\zeta_0)$ and $h(\theta,\omega,\zeta_0)$ are continuous with respect to (θ,ω).

(iv) *Recall the notation* $\varpi = \mathbb{E}\left[\phi_t\phi_t^\top\right]^{-1}\mathbb{E}\left[\delta_t^{ES}e_t\right]$, *for each compact set* $D_\zeta \in \mathbb{R}^p \times \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, *let*

$$H_{m,n} = \frac{1}{m}\sum_{t=n}^{n+m-1}\left(h(\theta,\omega,\zeta_t) - H(\theta,\omega)\right)\mathbb{I}(\zeta_t \in D_\zeta),$$

$$G_{m,n} = \frac{1}{m}\sum_{t=n}^{n+m-1}\left(g(\theta,\varpi,\zeta_t) - G(\theta)\right)\mathbb{I}(\zeta_t \in D_\zeta),$$

$$X_{m,n} = \frac{1}{m}\sum_{t=n}^{n+m-1}\frac{\alpha_t}{\beta_t}g(\theta,\omega,\zeta_t),$$

then for each $(\theta,\omega) \in \mathbb{R}^p \times \mathbb{R}^p$, we have:

$$\lim_{m,n\to\infty}\mathbb{E}\left[H_{m,n}\right] = \lim_{m,n\to\infty}\mathbb{E}\left[G_{m,n}\right] = \lim_{m,n\to\infty}\mathbb{E}\left[X_{m,n}\right] = 0.$$

By the Jensen's inequality, it is sufficient that

$$H_{m,n} \to 0, G_{m,n} \to 0, X_{m,n} \to 0,$$

as $m,n \to \infty$. With the fact

$$\mathbb{E}\left[h(\theta,\omega,\zeta_0)\right] = A_\sigma\theta + b_\sigma - M\omega$$

and following the same analysis of Proposition 2.3 in [28], we have $H_{m,n} \to 0$, as $m,n \to \infty$. Similarly, we have $G_{m,n} \to 0$, as $m,n \to \infty$.

Note that $g(\theta,\omega,\zeta_t)$ is Lipschitz continuous with respect to the trace variable e_t, where $e_t \in \zeta_t$. Thus, there exists a positive scalar L such that $\|g(\theta,\omega,\zeta_t)\| \le L\|e_t\|$. From Assumption 3, we have $\sup_{t\in\mathbb{N}}\|e_t\| < \infty$ and $\lim_{t\to\infty}\alpha_t/\beta_t = 0$, then we have

$$X_{m,n} \to 0,$$

as $m,n \to \infty$. □