Finite-Sample Analysis of Off-Policy TD-Learning via Generalized Bellman Operators

Zaiwei Chen [1] **Siva Theja Maguluri** [1] **Sanjay Shakkottai** [2] **Karthikeyan Shanmugam** [3]

Abstract

Policy evaluation (including multi-step off-policy importance sampling) has the interpretation of solving a generalized Bellman equation. In this paper, we derive finite-sample bounds for any general off-policy TD-like stochastic approximation algorithm that solves for the fixed point of this generalized Bellman operator. Our key step is to show that the generalized Bellman operator is simultaneously a contraction mapping with respect to a weighted ℓ_p-norm for each p in $[1, \infty)$, with a common contraction factor. Our results immediately imply finite-sample bounds of variants of off-policy TD-learning algorithms in the literature (e.g. $Q^\pi(\lambda)$, Tree-Backup, Retrace, and Q-trace).

1. Introduction

Reinforcement learning (RL) demonstrated its success in learning effective policies for a variety of decision making problems. In RL, there is an important sub-problem – called the policy evaluation problem – of estimating the expected long term reward of a given policy.

The policy evaluation problem is usually solved with the TD-learning method (Sutton, 1988). A key ingredient in TD-learning is the policy used to collect samples (called the behavior policy). Ideally we want to generate samples from the target policy whose value function we want to estimate, and this is called on-policy sampling. However, in many cases such on-policy sampling is not possible due to practical reasons, and hence we need to work with historical data that is generated by a possibly different policy (aka off-policy sampling). For example, in high stake applications such as clinic trials (Zhao et al., 2011), it is not practically possible to re-collect data every time we need to evaluate the performance of a given policy.

[1]Georgia Institute of Technology [2]The University of Texas at Austin [3]IBM Research NY. Correspondence to: Zaiwei Chen <zchen458@gatech.edu>.

ICML Workshop on Reinforcement Learning Theory, 2021.

Although off-policy sampling is more practical than on-policy sampling, it is more challenging to analyze and is known to have high variance (Glynn & Iglehart, 1989), which is due to the presence of the product of the importance sampling ratios, and is a fundamental difficulty in off-policy learning. To overcome this difficulty, many variants of off-policy TD-learning algorithms have been proposed in the literature, such as the $Q^\pi(\lambda)$ (Harutyunyan et al., 2016) algorithm, the TB(λ) algorithm (Precup et al., 2000), the Retrace(λ) algorithm (Munos et al., 2016), and the Q-trace algorithm (Khodadadian et al., 2021), etc.

Main Contributions. In this work, we establish finite-sample bounds of a general n-step off-policy TD-learning algorithm that also subsumes several algorithms presented in the literature. The key step is to show that such algorithm can be modeled as a Markovian stochastic approximation (SA) algorithm for solving a generalized Bellman equation. We present sufficient conditions under which the generalized Bellman operator is contractive with respect to a weighted ℓ_p-norm for every $p \in [1, \infty)$, with a uniform contraction factor for all p. Our result shows that the sample complexity scales as $\tilde{\mathcal{O}}(\epsilon^{-2})$, where ϵ is the required accuracy. It also involves a factor that depends on the problem parameters, in particular, the generalized importance sampling ratios, and explicitly demonstrates the bias-variance trade-off.

Our result immediately gives finite-sample guarantees for variants of multi-step off-policy TD-learning algorithms including $Q^\pi(\lambda)$, TB(λ), Retrace(λ), and Q-trace. For $Q^\pi(\lambda)$, TB(λ), and Retrace(λ), we establish the first-known results in the literature, while for Q-trace, we improve the best known results in (Khodadadian et al., 2021). In this paper we only present the finite-sample bound for Q-trace.

1.1. Preliminaries

The RL problem is usually modeled as a Markov decision process (MDP). In this work, we consider an MDP with a finite set of states $\mathcal{S}$, a finite set of actions $\mathcal{A}$, a set of unknown action dependent transition probability matrices $\mathcal{P} = \{P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|} \mid a \in \mathcal{A}\}$, an unknown reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto [0, 1]$, and a discount factor $\gamma \in (0, 1)$. In order for an MDP to progress, we must specify the policy of selecting actions based on the state of the

environment. Specifically, a policy π is a mapping from the state-space to probability distributions supported on the action space, i.e., $\pi : \mathcal{S} \mapsto \Delta^{|\mathcal{A}|}$. The state-action value function Q^π associated with a policy π is defined by $Q^\pi(s,a) = \mathbb{E}_\pi[\sum_{k=0}^\infty \gamma^k \mathcal{R}(S_k, A_k) \mid S_0 = s, A_0 = a]$ for all (s,a). The goal in policy evaluation is to estimate the state-action value function Q^π for a given policy π.

Since the transition probabilities as well as the reward function are unknown, such state-action value function cannot be directly computed. The TD-learning algorithm is designed to estimate Q^π using the SA method. Specifically, in TD-learning, we first collect a sequence of samples $\{(S_k, A_k)\}$ from the model using some behavior policy π_b. Then the value function Q^π is iteratively estimated using the samples $\{(S_k, A_k)\}$. When $\pi_b = \pi$, the algorithm is called on-policy TD-learning, otherwise it is called off-policy TD-learning.

2. Finite-sample analysis of general off-policy TD-learning

In this section, we present finite-sample analysis of off-policy TD-learning using generalized importance sampling ratios and multi-step bootstrapping.

2.1. A generic model for n-step off-policy TD

Algorithm 1 presents our generic algorithm model. Due to off-policy sampling, the two functions $c, \rho : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}_+$ are introduced in Algorithm 1 to serve as generalized importance sampling ratios in order to account for the discrepancy between π and π_b. We denote $c_{\max} = \max_{s,a} c(s,a)$ and $\rho_{\max} = \max_{s,a} \rho(s,a)$. We next show how Algorithm 1 captures variants of off-policy TD-learning algorithms in the literature by using different $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$.

Algorithm 1 General n-Step Off-Policy TD-Learning

Input: K, $\{\alpha_k\}$, Q_0, π, π_b, generalized importance sampling ratios $c, \rho : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}_+$, and sample trajectory $\{(S_k, A_k)\}_{0 \le k \le K+n}$ collected under the behavior policy π_b.
for $k = 0, 1, \cdots, K-1$ **do**
  $\alpha_k(s,a) = \alpha_k \mathbb{I}_{\{(s,a)=(S_k,A_k)\}}$ for all (s,a)
  $\Delta(S_i, A_i, S_{i+1}, A_{i+1}, Q_k) = \mathcal{R}(S_i, A_i) + \gamma\rho(S_{i+1}, A_{i+1})Q_k(S_{i+1}, A_{i+1}) - Q_k(S_i, A_i)$ for all $i \in \{k, k+1, ..., k+n-1\}$.
  $Q_{k+1}(s,a) = Q_k(s,a) + \alpha_k(s,a) \sum_{i=k}^{k+n-1} \gamma^{i-k} \prod_{j=k+1}^{i} c(S_j, A_j) \times \Delta(S_i, A_i, S_{i+1}, A_{i+1}, Q_k)$ for all (s,a)
end for
Output: Q_K

Vanilla IS. When $c(s,a) = \rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$ for all (s,a), Algorithm 1 is the standard off-policy TD-learning with importance sampling (Precup et al., 2000). We will refer to this algorithm as Vanilla IS.

$Q^\pi(\lambda)$. When $c(s,a) = \lambda$ and $\rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, Algorithm 1 is the $Q^\pi(\lambda)$ algorithm (Harutyunyan et al., 2016).

TB(λ). When $c(s,a) = \lambda\pi(a|s)$ and $\rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, we have the TB($\lambda$) algorithm (Precup et al., 2000).

Retrace(λ). When $c(s,a) = \lambda \min(1, \frac{\pi(a|s)}{\pi_b(a|s)})$ and $\rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, we have the Retrace($\lambda$) algorithm.

Q-trace. When $c(s,a) = \min(\bar{c}, \frac{\pi(a|s)}{\pi_b(a|s)})$ and $\rho(s,a) = \min(\bar{\rho}, \frac{\pi(a|s)}{\pi_b(a|s)})$, where $\bar{\rho} \ge \bar{c}$, Algorithm 1 is the Q-trace algorithm (Khodadadian et al., 2021). The Q-trace algorithm is an analog of the V-trace algorithm (Espeholt et al., 2018).

From now on, we focus on studying the generic algorithm 1. We make the following assumption about the behavior policy π_b, which is fairly standard in off-policy TD-learning.

Assumption 2.1. The behavior policy π_b satisfies $\pi_b(a|s) > 0$ for all (s,a), and the Markov chain $\{S_k\}$ induced by π_b is irreducible and aperiodic.

Irreducibility and aperiodicity together imply that the Markov chain $\{S_k\}$ has a unique stationary distribution, which we denote by $\kappa_S \in \Delta^{|\mathcal{S}|}$. Moreover, the Markov chain $\{S_k\}$ mixes geometrically fast in that there exist $C > 0$ and $\sigma \in (0,1)$ such that $\max_{s \in \mathcal{S}} \|P^k(s,\cdot) - \kappa_S(\cdot)\|_{\text{TV}} \le C\sigma^k$ for all $k \ge 0$, where $\|\cdot\|_{\text{TV}}$ is the total variation distance (Levin & Peres, 2017). Let $\kappa_{SA} \in \Delta^{|\mathcal{S}||\mathcal{A}|}$ be such that $\kappa_{SA}(s,a) = \kappa_S(s)\pi_b(a|s)$ for all (s,a). Note that κ_{SA} is the stationary distribution of the Markov chain $\{(S_k, A_k)\}$. Let $\mathcal{K}_S = \text{diag}(\kappa_S) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ and $\mathcal{K}_{SA} = \text{diag}(\kappa_{SA}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$, and denote the minimal diagonal entries of $\mathcal{K}_S$ and $\mathcal{K}_{SA}$ by $\mathcal{K}_{S,\min}$ and $\mathcal{K}_{SA,\min}$ respectively.

2.2. Identifying the generalized Bellman operator

In this section, we identify the generalized Bellman equation Algorithm 1 aims at solving, and also the corresponding generalized Bellman operator and its asynchronous variant. Let $\mathcal{T}_c, \mathcal{H}_\rho : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be two operators defined by

$$[\mathcal{T}_c(Q)](s,a) = \sum_{i=0}^{n-1} \gamma^i \mathbb{E}_{\pi_b}\left[\prod_{j=1}^{i} c(S_j, A_j) Q(S_i, A_i) | S_0 = s, A_0 = a\right],$$

$$[\mathcal{H}_\rho(Q)](s,a) = \mathcal{R}(s,a) + \gamma \mathbb{E}_{\pi_b}[\rho(S_{k+1}, A_{k+1}) Q(S_{k+1}, A_{k+1}) \mid S_k = s, A_k = a].$$

for all (s,a). Note that the operator $\mathcal{T}_c(\cdot)$ depends on the generalized importance sampling ratio $c(\cdot,\cdot)$, while the operator $\mathcal{H}_\rho(\cdot)$ depends on $\rho(\cdot,\cdot)$.

With $\mathcal{T}_c(\cdot)$ and $\mathcal{H}_\rho(\cdot)$ defined above, Algorithm 1 can be viewed as an asynchronous SA algorithm for solving the generalized Bellman equation $\mathcal{B}_{c,\rho}(Q) = Q$, where the generalized Bellman operator $\mathcal{B}_{c,\rho}(\cdot)$ is defined by $\mathcal{B}_{c,\rho}(Q) = \mathcal{T}_c(\mathcal{H}_\rho(Q) - Q) + Q$. Since Algorithm 1 per-

forms asynchronous update, using the terminology in (Chen et al., 2021a), we further define the asynchronous variant $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ of the generalized Bellman operator $\mathcal{B}_{c,\rho}(\cdot)$ by

$$\tilde{\mathcal{B}}_{c,\rho}(Q) := \mathcal{K}_{SA}\mathcal{B}_{c,\rho}(Q) + (I - \mathcal{K}_{SA})Q. \quad (1)$$

Each component of the asynchronous generalized Bellman operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ can be thought of as a convex combination with identity, where the weights are the stationary probabilities of visiting state-action pairs $(s,a) \in \mathcal{S} \times \mathcal{A}$. This captures the fact that when performing asynchronous update, the corresponding component is updated only when the state-action pair (s,a) is visited. It is clear from its definition that $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ has the same fixed-points as $\mathcal{B}_{c,\rho}(\cdot)$ (provided that they exist).

Under some mild conditions on the generalized importance sampling ratios $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$, we will show in the next section that both the asynchronous generalized Bellman operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ and the operator $\mathcal{H}_\rho(\cdot)$ are contraction mappings. Therefore, since $\mathcal{T}_c(\mathbf{0}) = \mathbf{0}$, the operators $\mathcal{H}_\rho(\cdot)$, $\mathcal{B}_{c,\rho}(\cdot)$, $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ all share the same unique fixed-point. Since the fixed-point of the operator $\mathcal{H}_\rho(\cdot)$ depends only on the generalized importance sampling ratio $\rho(\cdot,\cdot)$, but not on $c(\cdot,\cdot)$, we can flexibly choose $c(\cdot,\cdot)$ to control the variance while maintaining the fixed-point of the operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$.

2.3. Establishing the contraction property

In this section, we study the fixed-point and the contraction property of the asynchronous generalized Bellman operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$. We begin by introducing some notation. Let $D_c, D_\rho \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ be two diagonal matrices such that $D_c((s,a),(s,a)) = \sum_{a\in\mathcal{A}} \pi_b(a|s)c(s,a)$ and $D_\rho((s,a),(s,a)) = \sum_{a\in\mathcal{A}} \pi_b(a|s)\rho(s,a)$ for all (s,a). We denote $D_{c,\min}$ ($D_{c,\max}$) and $D_{\rho,\min}$ ($D_{\rho,\max}$) as the minimal (maximal) diagonal entries of the matrices D_c and D_ρ respectively. In view of the definition of $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ in Eq. (1), the fixed-point of $\mathcal{H}_\rho(\cdot)$ must also be a fixed-point of $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$. We first study the fixed point of $\mathcal{H}_\rho(\cdot)$ by establishing its contraction property.

Proposition 2.1. *Suppose that $D_{\rho,\max} < 1/\gamma$. Then the operator $\mathcal{H}_\rho(\cdot)$ is a contraction mapping with respect to the ℓ_∞-norm, with contraction factor $\gamma D_{\rho,\max}$. In this case, the unique fixed-point $Q^{\pi,\rho}$ of $\mathcal{H}_\rho(\cdot)$ satisfies the following inequality: $\|Q^\pi - Q^{\pi,\rho}\|_\infty \leq \frac{\gamma \max_{s,a} |\pi(a|s) - \pi_b(a|s)\rho(s,a)|}{(1-\gamma)(1-\gamma D_{\rho,\max})}$.*

Observe from Proposition 2.1 that when $\rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$, which is the case for $Q^\pi(\lambda)$, TB(λ), and Retrace(λ), the unique fixed-point $Q^{\pi,\rho}$ is exactly the target value function Q^π. This agrees with the definition of the operator $\mathcal{H}_\rho(\cdot)$ in that it reduces to the regular Bellman operator $\mathcal{H}_\pi(\cdot)$ when $\rho(s,a) = \frac{\pi(a|s)}{\pi_b(a|s)}$ for all (s,a). If $\rho(s,a) \neq \frac{\pi(a|s)}{\pi_b(a|s)}$ for some (s,a), then in general the fixed-point of $\mathcal{H}_\rho(\cdot)$ is different from Q^π. In that case, Proposition 2.1 provides an error bound on the difference between the potentially biased limit $Q^{\pi,\rho}$ and Q^π. Such error bound will be useful for us to study the Q-trace algorithm.

To further guarantee the uniqueness of the fixed-point of $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$, we establish the contraction property of $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$. We begin with the following definition.

Definition 2.1. Let $\{\mu_i\}_{1\leq i\leq d}$ be positive weights. For any $x \in \mathbb{R}^d$, the weighted ℓ_p-norm ($p \in [1,\infty)$) of x with weights $\{\mu_i\}$ is defined by $\|x\|_{\mu,p} = (\sum_i \mu_i |x_i|^p)^{1/p}$.

We next establish the contraction property of the operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ in the following theorem. Let $\omega = \mathcal{K}_{SA,\min} f(\gamma D_{c,\min})(1 - \gamma D_{\rho,\max})$, where the function $f : \mathbb{R} \mapsto \mathbb{R}$ is defined by $f(x) = n$ when $x = 1$, and $f(x) = \frac{1-x^n}{1-x}$ when $x \neq 1$.

Theorem 2.1. *Suppose that $c(s,a) \leq \rho(s,a)$ for all (s,a) and $D_{\rho,\max} < 1/\gamma$. Then for any $\theta \in (0,1)$, there exists a weight vector $\mu \in \Delta^{|\mathcal{S}||\mathcal{A}|}$ satisfying $\mu(s,a) \geq \frac{\omega(1-\theta)}{(1-\theta\omega)|\mathcal{S}||\mathcal{A}|}$ for all (s,a) such that the operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_{\mu,p}$ for any $p \in [1,\infty)$, with a common contraction factor $1 - \theta\omega$.*

Theorem 2.1 is the key result for our finite-sample analysis, and we present its proof in the next section. The weighted ℓ_p-norm (especially the weighted ℓ_2-norm) contraction property we established for the operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ has a far-reaching impact even beyond the finite-sample analysis of tabular RL in this paper. Specifically, recall that the key property used for establishing the convergence and finite-sample bound of on-policy TD-learning with *linear function approximation* in the seminal work (Tsitsiklis & Van Roy, 1997) is that the corresponding Bellman operator is a contraction mapping not only with respect to the ℓ_∞-norm, but also with respect to a weighted ℓ_2-norm. We establish the same property in the off-policy setting, and hence lay down the foundation for extending our results to the function approximation setting. This is an immediate future research direction.

2.4. Proof of Theorem 2.1

We begin by explicitly computing the asynchronous generalized Bellman operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$. Let π_c and π_ρ be two policies defined by $\pi_c(a|s) = \frac{\pi_b(a|s)c(s,a)}{D_c((s,a),(s,a))}$ and $\pi_\rho(a|s) = \frac{\pi_b(a|s)\rho(s,a)}{D_\rho((s,a),(s,a))}$ for all (s,a).

Proposition 2.2. *The operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ is explicitly given by $\tilde{\mathcal{B}}_{c,\rho}(Q) = AQ + b$ for any Q, where $A = I - \mathcal{K}_{SA}\sum_{i=0}^{n-1}(\gamma P_{\pi_c} D_c)^i (I - \gamma P_{\pi_\rho} D_\rho)$ and $b = \mathcal{K}_{SA}\sum_{i=0}^{n-1}(\gamma P_{\pi_c} D_c)^i R$.*

In light of Proposition 2.2, to prove Theorem 2.1, it is enough to work with the matrix A. To proceed, we require the following definition.

Definition 2.2. Given $\beta \in [0,1]$, a matrix $M \in \mathbb{R}^{d\times d}$ is

called a substochastic matrix with modulus β if and only if $M_{ij} \geq 0$ for all i, j and $\sum_j M_{ij} \leq 1-\beta$ for all i.

We next show in the following two propositions that (1) the matrix A given in Proposition 2.2 is a sub-stochastic matrix with modulus ω, and (2) for any sub-stochastic matrix M with a positive modulus, there exist weights $\{\mu_i\}$ such that the induced matrix norm $\|M\|_{\mu,p}$ is strictly less than 1. These two results together imply Theorem 2.1.

Proposition 2.3. *Suppose that $c(s,a) \leq \rho(s,a)$ for all (s,a) and $D_{\rho,\max} < 1/\gamma$. Then the matrix A given in Proposition 2.2 is a sub-stochastic matrix with modulus ω, where $\omega = \mathcal{K}_{SA,\min} f(\gamma D_{c,\min})(1-\gamma D_{\rho,\max})$.*

The condition $c(s,a) \leq \rho(s,a)$ ensures that the matrix A is non-negative, and the condition $D_{\rho,\max} < 1/\gamma$ ensures that the each row of the matrix A sums up to at most $1-\omega$. Together they imply the substochasticity of A. The modulus ω is an important parameter for our finite-sample analysis. In view of Theorem 2.1, we see that large modulus gives smaller (or better) contraction factor of $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$.

Proposition 2.4. *For any sub-stochastic matrix $M \in \mathbb{R}^{d\times d}$ with a positive modulus $\beta \in (0,1)$, for any $\theta \in (0,1)$, there exists $\mu \in \Delta^d$ satisfying $\mu_i \geq \frac{\beta(1-\theta)}{(1-\theta\beta)d}$ for all i such that $\|M\|_{\mu,p} \leq 1-\theta\beta$ for any $p \in [1,\infty)$. Furthermore, if M is irreducible [1], then we can choose $\theta = 1$.*

Note that Proposition 2.4 introduces the tunable parameter θ. It is clear that large θ gives better contraction factor of $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ but worse lower bound on the entries of the weight vector μ. In general, when M is not irreducible, we cannot hope to choose a weight vector $\mu \in \Delta^d$ with positive components and obtain $\|M\|_{\mu,p} \leq 1-\omega$. To see this, consider the example where $M = (1-\omega)[\mathbf{0}, \mathbf{0}, \cdots, \mathbf{1}]$, which is clearly a substochastic matrix with modulus ω, but is not an irreducible matrix. For any weight vector $\mu \in \Delta^d$, we have $\|M\|_{\mu,p} = (1-\omega)\max_{x\in\mathbb{R}^d:\|x\|_{\mu,p}=1}|x_d| = (1-\omega)/\mu_d^{1/p} > 1-\omega$. However, by choosing μ_d close to unity, we can get $\|M\|_{\mu,p}$ arbitrarily close to $1-\omega$. This is analogous to choosing θ close to one in Proposition 2.4. Since Proposition 2.4 is the major result for proving Theorem 2.1, we provide its proof sketch in Section 3.

2.5. Finite-sample convergence guarantees

In light of Theorem 2.1, Algorithm 1 is a Markovian SA algorithm for solving a fixed-point equation $\tilde{\mathcal{B}}_{c,\rho}(Q) = Q$, where the fixed-point operator $\tilde{\mathcal{B}}_{c,\rho}(\cdot)$ is a contraction mapping. Therefore, to establish the finite-sample bounds, we use a Lyapunov drift argument where we choose $W(Q) = \|Q - Q^{\pi,\rho}\|_{\mu,p}^2$ as the Lyapunov function. This leads to a finite-sample bound on $\mathbb{E}[\|Q_k - Q^{\pi,\rho}\|_{\mu,p}^2]$. However, since μ is unknown, to make the finite-sample bound independent of μ, we use the lower bound on $\mu(s,a)$ provided in Theorem 2.1 and also tune the parameters p and θ to obtain a finite-sample bound on $\mathbb{E}[\|Q_k - Q^{\pi,\rho}\|_\infty^2]$. The fact that the contraction factor $1-\theta\omega$ (cf. Theorem 2.1) is independent of p plays an important role in such tuning process.

To present the results, we need to introduce more notation. For any $\delta > 0$, define $t_\delta(\mathcal{MC}_S)$ as the mixing time of the Markov chain $\{S_k\}$ (induced by π_b) with precision δ, i.e., $t_\delta(\mathcal{MC}_S) = \min\{k \geq 0 : \max_{s\in\mathcal{S}} \|P^k(s,\cdot) - \kappa_S(\cdot)\|_{\text{TV}} \leq \delta\}$. Under Assumption 2.1, we can easily verify that $t_\delta(\mathcal{MC}_S) \leq L(\log(1/\delta)+1)$ for some constant $L > 0$. Let $\tau_{\delta,n} = t_\delta(\mathcal{MC}_S) + n + 1$. The parameters $\{c_i\}_{1\leq i\leq 3}$ used in the following theorem are numerical constants.

Theorem 2.2. *Consider $\{Q_k\}$ of Algorithm 1. Suppose that (1) Assumptions 2.1 is satisfied, (2) $c(s,a) \leq \rho(s,a)$ for all (s,a) and $D_{\rho,\max} < 1/\gamma$, and (3) α is chosen such that $\alpha\tau_{\alpha,n} \leq \frac{c_1\omega}{\log(2|\mathcal{S}||\mathcal{A}|/\omega) f(\gamma c_{\max})^2(\gamma\rho_{\max}+1)^2}$. Then we have the following finite-sample convergence bound:*

$$\mathbb{E}[\|Q_k - Q^{\pi,\rho}\|_\infty^2] \leq \zeta_1\left(1-\frac{\omega\alpha}{2}\right)^{k-\tau_{\alpha,n}} + \zeta_2 \frac{f(\gamma c_{\max})^2(\gamma\rho_{\max}+1)^2\log(2|\mathcal{S}||\mathcal{A}|/\omega)}{\omega}\alpha\tau_{\alpha,n}, \quad (2)$$

where $\zeta_1 = c_2(\|Q_0 - Q^{\pi,\rho}\|_\infty + \|Q_0\|_\infty + 1)^2$, and $\zeta_2 = c_3(3\|Q^{\pi,\rho}\|_\infty + 1)^2$.

Theorem 2.2 enables one to design a wide class of off-policy TD variants with provable finite-sample guarantees by choosing appropriate generalized importance sampling ratios $c(\cdot,\cdot)$ and $\rho(\cdot,\cdot)$, which, as we will see soon, are closely related to the bias-variance trade-off in Algorithm 1. The first term on the RHS of Eq. (2) is usually called the convergence bias in SA literature (Bottou et al., 2018), and it goes to zero at a geometric rate. The second term on the RHS of Eq. (2) stands for the variance in the iterates, and it is a constant proportional to $\alpha\tau_{\alpha,n}$. To see more explicitly the bias-variance trade-off, we derive the sample complexity of Algorithm 1 in the following.

Corollary 2.2.1. *For an accuracy $\epsilon > 0$, to obtain $\mathbb{E}[\|Q_k - Q^{\pi,\rho}\|_\infty] \leq \epsilon$, the sample complexity is*

$$\underbrace{\tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2(1-\gamma)^2}\right)}_{T_1}\underbrace{\tilde{\mathcal{O}}\left(\frac{f(\gamma c_{\max})^2(\gamma\rho_{\max}+1)^2}{\mathcal{K}_{SA}^2 f(\gamma D_{c,\min})^2(1-\gamma D_{\rho,\max})^2}\right)}_{T_2}.$$

From Corollary 2.2.1, we see that the dependency on the accuracy is $\tilde{\mathcal{O}}(\epsilon^{-2})$, and the dependency on the parameter n and the effective horizon $1/(1-\gamma)$ is $\tilde{\mathcal{O}}(n/(1-\gamma)^2)$, both of which are the same as TD-learning in the on-policy setting (Chen et al., 2021a). The impact of performing off-policy sampling is captured by the term T_3, which depends on the choice of the importance sampling ratios.

In the numerator of T_3, we have $f(\gamma c_{\max})^2(\gamma\rho_{\max}+1)^2$, which is from the second term on the RHS of Eq. (2),

[1] A non-negative matrix is irreducible if and only if its associated graph is strongly connected (Berman & Plemmons, 1994).

and represents the impact of the variance on the sample complexity. It is clear that smaller $c_{\max}$ and $\rho_{\max}$ lead to smaller variance. As we will see later, this is the reason for the variance reduction of various off-policy TD-learning algorithms in the literature. In the denominator of T_3, we have $\mathcal{K}_{SA,\min}^2 f(\gamma D_{c,\min})^2(1-\gamma D_{\rho,\max})^2 = \omega^2$, which represents the effect of the contraction factor. To see this, recall from Theorem 2.1 that the contraction factor is $1-\theta\omega$. In light of the previous analysis, the bias-variance trade-off in general off-policy multi-step TD-learning algorithm 1 is intuitively of the form $\tilde{\mathcal{O}}\left(\frac{\text{Variance}}{(1-\text{ Contraction factor})^2}\right)$.

2.6. Finite-sample analysis of Q-Trace

In this section, we apply Theorem 2.2 to the Q-trace algorithm and obtain an improved sample complexity compared to (Khodadadian et al., 2021). Similarly, our results can be used to obtain the first-known finite-sample bounds of Vanilla IS, $Q^\pi(\lambda)$, TB(λ), and Retrace(λ).

Consider the Q-trace algorithm, where $c(s,a) = \min(\bar{c}, \frac{\pi(a|s)}{\pi_b(a|s)})$ and $\rho(s,a) = \min(\bar{\rho}, \frac{\pi(a|s)}{\pi_b(a|s)})$ for all (s,a). This implies that $c_{\max} = \bar{c}$ and $\rho_{\max} = \bar{\rho}$. Moreover, we have $D_c(s,a) = \sum_a \min(\bar{c}\pi_b(a|s), \pi(a|s))$ and $D_\rho(s,a) = \sum_a \min(\bar{\rho}\pi_b(a|s), \pi(a|s))$ for all (s,a).

Theorem 2.3. *Consider Algorithm 1 with Q-trace update. Under Assumption 2.1, suppose $\bar{c} \le \bar{\rho}$ and α is chosen such that $\alpha\tau_{\alpha,n} \le \frac{c_1\omega}{\log(2|\mathcal{S}||\mathcal{A}|/\omega) f(\gamma\bar{c})^2(\gamma\bar{\rho}+1)^2}$. Then for all $k \ge \tau_{\alpha,n}$ we have $\mathbb{E}[\|Q_k - Q^{\pi,\rho}\|_\infty^2] \le \zeta_1\left(1-\frac{\omega\alpha}{2}\right)^{k-\tau_{\alpha,n}} + \zeta_2 \frac{f(\gamma\bar{c})^2(\gamma\bar{\rho}+1)^2\log(2|\mathcal{S}||\mathcal{A}|/\omega)}{\omega}\alpha\tau_{\alpha,n}$, where $\omega = \mathcal{K}_{SA,\min} f(\gamma D_{c,\min})(1-\gamma D_{\rho,\max})$. This implies a sample complexity of $\tilde{\mathcal{O}}\left(\frac{\log^2(1/\epsilon) n f(\gamma\bar{c})^2(\gamma\bar{\rho}+1)^2}{\epsilon^2\mathcal{K}_{SA}^2 f(\gamma D_{c,\min})^2(1-\gamma D_{\rho,\max})^2(1-\gamma)^2}\right)$.*

To avoid an exponential large variance, in view of the term $f(\gamma\bar{c})$ in our bound, we need to choose $\bar{c} \le 1/\gamma$. Due to the truncation level $\bar{\rho}$, the algorithm converges to a biased limit $Q^{\pi,\rho}$ instead of Q^π. Such truncation bias can be controlled using Proposition 2.1. These observations agree with the results (Khodadadian et al., 2021), where the finite-sample bounds of Q-trace were first established.

Compared to (Khodadadian et al., 2021), we have an improved sample complexity. Specifically, we have a sample complexity of $\tilde{\mathcal{O}}(\frac{\log^2(1/\epsilon) n f(\gamma\bar{c})^2(\gamma\bar{\rho}+1)^2}{\epsilon^2\mathcal{K}_{SA}^2 f(\gamma D_{c,\min})^2(1-\gamma D_{\rho,\max})^2})$, while the result in (Khodadadian et al., 2021) implies a sample complexity of $\tilde{\mathcal{O}}(\frac{\log^2(1/\epsilon) n f(\gamma\bar{c})^2(\gamma\bar{\rho}+1)^2}{\epsilon^2\mathcal{K}_{SA}^3 f(\gamma D_{c,\min})^3(1-\gamma D_{\rho,\max})^3})$, which has an additional factor of $(\mathcal{K}_{SA} f(\gamma D_{c,\min})(1-\gamma D_{\rho,\max}))^{-1}$. Since $\mathcal{K}_{SA,\min}^{-1} \ge |\mathcal{S}||\mathcal{A}|$, our result improves the dependency on the size of the state-action space by a factor of at least $|\mathcal{S}||\mathcal{A}|$ compared to (Khodadadian et al., 2021). Similarly, since V-trace is an analog of Q-trace, we can improve the sample complexity for V-trace in (Chen et al., 2021a).

3. Proof sketch of Proposition 2.4

The idea is to construct a stochastic matrix M'' such that (1) M'' dominates M in the sense that $M''_{ij} \ge M_{ij}$ for all i,j, and (2) the Markov chain associated with M'' is irreducible, hence admits a unique stationary distribution μ satisfying $\mu_i > 0$ for all i. Using μ as weights, we have the desired result. The detailed analysis is presented in our online report (Chen et al., 2021b). We here only present how to construct such a stochastic matrix M''.

First of all, consider the special case where M itself is irreducible. Then we first scale up M by a factor of $1/(1-\omega)$ to obtain $M' = \frac{M}{1-\omega}$, which is clearly a substochastic matrix, with modulus zero. Hence there exists a stochastic matrix M'' that dominates M' (and also dominates M). Moreover, since M'' is also irreducible, its associated Markov chain admits a unique stationary distribution μ. This is equivalent to choosing $\theta = 1$ in Proposition 2.4. In fact, the matrix M being irreducible is only a sufficient condition for us to choose $\theta = 1$. What we need is the existence of a strictly positive stationary distribution of the stochastic matrix M'', which is guaranteed when M'' does not have transient states.

Now consider the general case where M is not necessarily irreducible. We construct the intermediate matrix M' by performing a convex combination of the matrix $\frac{M}{1-\omega}$ and the uniform stochastic matrix $\frac{E}{d}$, where E is the all one matrix, with weight $\frac{1-\omega}{1-\theta\omega}$. Specifically, for any $\theta \in (0,1)$, we define $M' = \left(\frac{1-\omega}{1-\theta\omega}\right)\frac{M}{1-\omega} + \left(1-\frac{1-\omega}{1-\theta\omega}\right)\frac{E}{d}$. Note that M' is a non-negative matrix. In addition, since $M'\mathbf{1} \le \frac{1-\omega}{1-\theta\omega}\mathbf{1} + \left(1-\frac{1-\omega}{1-\theta\omega}\right)\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the all one vector, the matrix M' is a substochatic matrix with modulus zero, and is also irreducible because all its entries are strictly positive. Therefore, there exists a stochastic matrix M'' such that $M'' \ge M'$. In addition, since M'' also has strictly positive entries, the Markov chain associated with M'' is irreducible, hence admits a unique stationary distribution $\mu \in \Delta^d$. By our construction, we can show a lower bound on the components of the stationary distribution μ.

4. Conclusion

In this work, we establish finite-sample guarantees of general n-step off-policy TD-learning algorithms. The key in our approach is to identify a generalized Bellman operator and establishes its contraction property with respect to a weighted ℓ_p-norm for each $p \in [1,\infty)$, with a uniform contraction factor. Our results are used to derive finite-sample guarantees of variants of n-step off-policy TD-learning algorithms in the literature. In particular, for Q-trace, we improve the result in (Khodadadian et al., 2021). The finite-sample bounds we establish also provide insights about the trade-off between the convergence rate and the variance.

References

Berman, A. and Plemmons, R. J. *Nonnegative matrices in the mathematical sciences*. SIAM, 1994.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *Preprint arXiv:2102.01567*, 2021a.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-Sample Analysis of Off-Policy TD-Learning via Generalized Bellman Operators. *arXiv preprint arXiv:2106.12729*, 2021b.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., et al. IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures. In *International Conference on Machine Learning*, pp. 1407–1416, 2018.

Glynn, P. W. and Iglehart, D. L. Importance sampling for stochastic simulations. *Management science*, 35(11): 1367–1392, 1989.

Harutyunyan, A., Bellemare, M. G., Stepleton, T., and Munos, R. Q(λ) with Off-Policy Corrections. In *International Conference on Algorithmic Learning Theory*, pp. 305–320. Springer, 2016.

Khodadadian, S., Chen, Z., and Maguluri, S. T. Finite-Sample Analysis of Off-Policy Natural Actor-Critic Algorithm. *The 38th International Conference on Machine Learning*, 2021.

Levin, D. A. and Peres, Y. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. Safe and efficient off-policy reinforcement learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1054–1062, 2016.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 759–766, 2000.

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

Tsitsiklis, J. N. and Van Roy, B. Analysis of temporal-difference learning with function approximation. In *Advances in neural information processing systems*, pp. 1075–1081, 1997.

Zhao, Y., Zeng, D., Socinski, M. A., and Kosorok, M. R. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.