Finite Sample Analysis of Average-Reward TD Learning and Q-Learning

Sheng Zhang [*1] **Zhe Zhang** [*1] **Siva Theja Maguluri** [1]

Abstract

The focus of this paper is on sample complexity guarantees of average-reward reinforcement learning algorithms, which are known to be more challenging to study than their discounted-reward counterparts. To the best of our knowledge, we provide the first known finite sample guarantees using both constant and diminishing step sizes of (i) average-reward TD(λ) with linear function approximation for policy evaluation and (ii) average-reward Q-learning in the tabular setting to find an optimal policy. A major challenge is that since the value functions are agnostic to an additive constant, the corresponding Bellman operators are no longer contraction mappings under any norm. We obtain the results for TD(λ) by working in an appropriately defined subspace that ensures uniqueness of the solution. For Q-learning, we exploit the span seminorm contractive property of the Bellman operator and construct a novel Lyapunov function obtained by infimal convolution of the generalized Moreau envelope and the indicator function of a set.

1. Introduction

The average-reward setting is a classical setting for formulating the goal in an infinite-horizon Markov decision process (MDP) (Sutton, 1988). The need to maximize the average reward has been demonstrated in many applications, including scheduling automatic guided vehicles (Tadepalli et al., 1994), inventory management in supply chains (Giannoccaro & Pontrandolfo, 2002), communication system control and routing (Marbach et al., 2000) and cooperative multi-robot learning (Tangamchit et al., 2002). In these problems, the discounted-reward criterion usually leads to poor long-time performance since the system operates over an extended period of time and the main objective is to perform consistently well over the long run.

Even though there is a well developed theory of average-reward MDPs (Howard, 1960; Blackwell, 1962; Puterman, 2014), the theoretical understanding of average-reward reinforcement learning (RL) methods is still quite limited. Most existing results are focused only on asymptotic convergence (Tsitsiklis & Van Roy, 1999; Abounadi et al., 2001; Wan et al., 2020; Zhang et al., 2021). The focus of this paper is to understand the sample efficiency. *How much data is required to guarantee a given level of accuracy?*

Recent literature obtains finite sample guarantees for discounted reward TD and Q learning algorithms (Bhandari et al., 2018; Srikant & Ying, 2019; Chen et al., 2019; Qu & Wierman, 2020; Chen et al., 2020) by developing novel analytical techniques. Such a study of average-reward RL algorithms is not undertaken. Analysis of average-reward RL algorithms is known to be more challenging to study than their discounted-reward counterparts. The key property that is exploited in the study of discounted-reward problems is the contraction property of the underlying Bellman operator. In the average-reward setting, such a contraction property does not hold under any norm, and the Bellman equation is known to have multiple fixed points.

In this work, we take the first step toward understanding finite sample guarantees of model-free average-reward RL algorithms. Specifically, we consider (i) average-reward TD(λ) with linear function approximation for policy evaluation, and (ii) average-reward tabular Q-learning in the synchronous setting for the control problem.

1.1. Contributions and Summary of Our Techniques

We establish the first finite sample convergence guarantees of average-reward TD(λ) with linear function approximation and average-reward Q-learning in the literature.

TD(λ) Results. We study the average-reward TD(λ) with linear function approximation under a general asynchronous update. We present finite-sample bounds under both constant and diminishing step sizes. With a constant step-size, the iterates converge at an exponential rate to a small cylinder around the set of TD fixed points. With properly chosen diminishing step sizes, the mean-square distance of the iter-

[*]Equal contribution [1]The H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Sheng Zhang <shengzhang@gatech.edu>.

Accepted at the ICML 2021 Workshop on Reinforcement Learning Theory. Copyright 2021 by the author(s).

ates to this set converges with an $\tilde{\mathcal{O}}\left(\frac{1}{T}\right)$ rate, and this leads to a sample complexity of $\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2}\right)$. Our sample complexity bound suggests a trade-off in choosing λ, i.e., the optimal λ should be neither too large nor too small. The dependence on the effective horizon plays a key role in the study of discounted-reward RL algorithms (Pananjady & Wainwright, 2020; Chen et al., 2021). There is no such effective horizon in average-reward problems, and the spectral gap of an appropriately defined matrix plays a key role instead.

TD(λ) Analysis. A major challenge in the analysis is that the projected Bellman operator is not a contraction under any norm. Moreover, even though the projected Bellman equation can be written as a linear set of equations, they are underdetermined. So existing techniques (Bhandari et al., 2018; Srikant & Ying, 2019) are not directly applicable. Since the value function is unique up to an additive constant, we have a unique solution of the projected Bellman equation when restricted to an appropriately defined subspace. We exploit this property and work in this subspace and use a quadratic Lyapunov function to obtain finite sample guarantees.

Q-learning Results. We consider a J-step synchronous Q-learning algorithm. We present finite sample error bounds under diminishing step-sizes. The span seminorm of a vector is defined to be the difference between the maximum and minimum element. Since the optimal value function, Q^*, is agnostic up to an additive constant, we show that the mean-square of the span seminorm of the error $Q_t - Q^*$ converges at $\mathcal{O}(\frac{1}{t})$ rate. This implies a sample complexity of $\mathcal{O}(\frac{1}{\epsilon^2})$ to find an ϵ-optimal differential Q-function.

Q-learning Analysis. While the average reward Bellman operator is not a contraction under any norm, it is known to be a contraction under the span seminorm. The span seminorm can be interpreted as the ℓ_∞ distance to the space spanned by the all-ones vector. Finite sample bounds for stochastic approximation of ℓ_∞-norm contractive operators were obtained in (Chen et al., 2020) by using generalized Moreau envelope as a smooth Lyapunov function. Here, we generalize this approach and introduce a new Lyapunov function to study span seminorm contractive operators. Our Lyapunov function is obtained by applying an infimal convolution with respect to an indicator function to the generalized Moreau envelop used in (Chen et al., 2020).

1.2. Related Literature

Average-Reward MDP. There is an extensive body of literature on average-reward MDPs. Several authors have made early contributions to the average-reward problem (Gillette, 2016; Howard, 1960; Blackwell, 1962; Brown, 1965; Veinott, 1966). There are well known classical dynamic programming algorithms for finding optimal policies such as policy iteration (Howard, 1960) and value iteration (White, 1963). However, these algorithms require knowledge of the state transition probabilities, and are also computationally intractable in large state spaces (Mahadevan, 1996).

Average-Reward Policy Evaluation. Tsitsiklis and Van Roy (Tsitsiklis & Van Roy, 1999) proved the convergence of an average-reward TD(λ) with linear function approximation, and provided approximation error bounds. The paper did not study finite-sample bounds. Yu and Bertsekas (Yu & Bertsekas, 2009) proved the convergence of an average-reward LSPE(λ) and its rate of convergence for constant step size. Both TD(λ) and LSPE(λ) aim to solve the same projected Bellman equation. However, LSTD(λ) uses simulation to construct directly the low-dimensional quantities defining the equation, instead of only the solution itself like TD(λ). Both papers assumed that the set of basis functions are independent of the all-ones vector, which apparently does not hold in the tabular setting. We do not require such an assumption in this paper. Note that both algorithms above are on-policy. Recently, in addition to a convergent off-policy tabular method presented in (Wan et al., 2020), Zhang et al. (Zhang et al., 2021) introduced a convergent off-policy linear function approximation algorithm. Both did not study finite-time convergence guarantees.

Average-Reward Control. The earliest control algorithms were those introduced by Schwartz (Schwartz, 1993) and Singh (Singh, 1994) without convergence proofs. The first provably convergent algorithms are RVI Q-learning and SSP Q-learning, introduced by Abounadi, Bertsekas, and Borkar (Abounadi et al., 2001). SSP Q-learning and the algorithm introduced later by Gosavi (Gosavi, 2004) are limited to MDPs with a special state that is recurrent under all stationary policies, whereas RVI Q-learning is convergent for more general MDPs. Recently, Wan et al.(Wan et al., 2020) introduced an algorithm without a reference function, which is needed in RVI Q-learning, and proved its convergence with the techniques that are a slight generalization of those in (Abounadi et al., 2001). To the best of our knowledge, our paper is the first work in the literature that studies the finite sample guarantees of a general average-reward Q-learning algorithm.

Stochastic Approximation. Many RL algorithms can be viewed through the lens of stochastic approximation (SA). There is a well developed asymptotic theory of SA (Kushner & Yin, 2003; Borkar, 2009; Benveniste et al., 2012). The ODE method is a dominant approach used in most asymptotic convergence proofs in RL (Borkar & Meyn, 2000). However, this is a coarse tool, since it is not able to generate insight into an algorithm's sensitivity to noise in the system and step-size choices. Driven by the interest in finite-sample guarantees of RL algorithms, recent years have witnessed a focus shifted from asymptotic analysis to non-asymptotic analysis of SA schemes. For example, a finite-time bound for linear SA is given in (Srikant & Ying, 2019), which leads to finite time error bounds for TD learning. (Qu & Wierman, 2020) provides a finite-time analysis of asynchronous non-linear SA, which yields finite-time bounds on asynchronous Q-learning.

Others. There are other related papers which are beyond the scope of the present paper. For instance, there is a line of work (Jaksch et al., 2010; Abbasi-Yadkori et al., 2019; Wei et al., 2020) on regret guarantees, which is a different focus compared to our work, for learning in average-reward MDPs. In addition, there are RL methods based on linear programming (Wang, 2017; Neu et al., 2017), or learning automata (Wheeler & Narendra, 1986; Chang, 2009).

1.3. The Average-Reward Problem Setting

We consider an infinite-horizon average-reward MDP described by $(\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$, where $\mathcal{S} = \{1, 2, \cdots, n\}$ is a finite state space, $\mathcal{A} = \{1, 2, \cdots, m\}$ is a finite action space, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function (i.e., $\mathcal{R}(s, a)$ is the immediate reward received upon executing action a while in state s), and $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the transition dynamics of the environment (i.e., $p(s'|s, a)$ is the probability of transitioning into state s' upon taking action a in state s). An agent interacts with the environment according to the following protocol: at each time step $t = 0, 1, 2, \cdots$, the agent is in a state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$, then receives from the environment an immediate reward $\mathcal{R}(s_t, a_t)$ and the next state s_{t+1} which is a sample drawn from $p(\cdot|s_t, a_t)$. The average reward per step of a deterministic stationary policy $\mu : \mathcal{S} \to \mathcal{A}$ starting from state $s \in \mathcal{S}$ is defined as

$$r^{\mu}(s) := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \mathcal{R}(s_t, \mu(s_t))|s_0 = s\right]. \quad (1)$$

Let $r^*(s) := \sup_{\mu\in\mathcal{M}} r^{\mu}(s)$, where $\mathcal{M}$ is the set of deterministic stationary policies. A policy $\mu^* \in \mathcal{M}$ is said to be optimal if it satisfies $r^{\mu^*}(s) = r^*(s)$ for all $s \in \mathcal{S}$.

2. Policy Evaluation Algorithm: TD Learning

2.1. Problem Formulation and Average-Reward TD(λ)

We consider the problem of evaluating a given policy $\mu \in \mathcal{M}$ in an average-reward MDP when the data is generated by applying the policy μ. Since the environment is an induced Markov reward process, we employ the notation $\mathcal{R}(s) := \mathcal{R}(s, \mu(s))$ for rewards, and $P(s, s') := p(s'|s, \mu(s))$ for transition probabilities. We make the following standard assumption to ensure there is a unique stationary distribution π with $\pi(i) > 0$ for all $i \in \mathcal{S}$. Let $\mathbb{E}_{\pi}[\cdot]$ stand for expectation with respect to π and $D = \text{diag}(\pi_1, \cdots, \pi_n)$ denote the diagonal matrix consisting of π. The D-weighted norm, $\|x\|_D = \sqrt{x^\top Dx}$, is helpful.

Assumption 1. *The Markov chain associated with P is irreducible and aperiodic.*

Under Assumption 1, it is known that the average reward satisfies $r^{\mu}(s) = r(\mu) := \mathbb{E}_{\pi}[\mathcal{R}(s_t)]$ for all $s \in \mathcal{S}$, and the set of differential value functions takes the form $\{v^{\mu} + ce|c \in \mathbb{R}\}$, where $v^{\mu} : \mathcal{S} \to \mathbb{R}$, known as the *basic* differential value function, is given by $v^{\mu} := \sum_{t=0}^{\infty} P^t \left(\mathcal{R} - r(\mu)e\right)$.

We consider approximation to differential value functions using a function of the form $V_{\theta}(i) = \phi(i)^\top\theta$ for all $i \in \mathcal{S}$, where $\phi(i) := [\phi_1(i), \cdots, \phi_d(i)]^\top \in \mathbb{R}^d$ is the feature vector for state $i \in \mathcal{S}$ and $\theta \in \mathbb{R}^d$ is a tunable parameter vector. Here, $\{\phi_k : \mathcal{S} \to \mathbb{R}|k = 1, 2, \cdots, d\}$ is a given set of d basis functions with $d \le n$. With this notation, V_{θ} can be expressed compactly in the form $V_{\theta} = \Phi\theta$, where Φ is an $n \times d$ matrix whose k-th column is ϕ_k. We assume that Φ has full column rank and $\|\phi(i)\|_2 \le 1$ for all $i \in \mathcal{S}$.

Algorithm 1: TD(λ) with linear function approximation

```
Input  : initial guess $\bar{r}_0$ and $\theta_0$, basis functions
         $\{\phi_k\}_{k=1}^{d}$, step-size sequence $\{\beta_t\}_{t\in\mathbb{N}}$ and
         positive constant $c_\alpha$.
Initialize: $z_{-1} = 0$, $\lambda \in [0, 1)$.
for $t = 0, 1, \dots$ do
    Observe tuple: $O_t = (s_t, \mathcal{R}(s_t), s_{t+1})$
    Get TD error:
      $\delta_t(\theta_t) = \mathcal{R}(s_t) - \bar{r}_t + \phi(s_{t+1})^\top\theta_t - \phi(s_t)^\top\theta_t$
    Update eligibility trace: $z_t = \lambda z_{t-1} + \phi(s_t)$
    Update average-reward estimate:
      $\bar{r}_{t+1} = \bar{r}_t + c_\alpha\beta_t(\mathcal{R}(s_t) - \bar{r}_t)$
    Update parameter vector:
      $\theta_{t+1} = \theta_t + \beta_t\delta_t(\theta_t)z_t$
end
```

We present in Algorithm 1 the average-reward TD(λ) with linear function approximation introduced by (Tsitsiklis & Van Roy, 1999). Tsitsiklis and Van Roy proved the asymptotic convergence of Algorithm 1 and provided approximation error bounds. However, they did not study finite-sample bounds, and to ensure the TD(λ) limit point is unique, they assumed that the set of basis functions are independent of the all-ones vector, which apparently does not hold in the tabular setting. We do not require such an assumption in this paper.

2.2. Finite-Time Bounds for Average-Reward TD(λ)

We study the drift of an appropriately chosen Lyapunov function to obtain an upper bound on the mean-square error. Denote by $S(\Phi, e) := \text{span}\left(\{\theta|\Phi\theta = e\}\right)$ the linear span of solutions to $\Phi\theta = e$. Since Φ has full rank,

$S(\Phi, e) = \{0\}$ if $e \notin W := \{\Phi\theta | \theta \in \mathbb{R}^d\}$; otherwise, $S(\Phi, e) = \{c\theta_e | \Phi\theta_e = e, c \in \mathbb{R}\}$. Let E be the orthogonal complement of $S(\Phi, e)$. We can think of E as the set of equivalent classes with the equivalence relation $\sim$ on $\mathbb{R}^d$ defined by $\theta_1 \sim \theta_2$ if and only if $\theta_1 - \theta_2$ is in $S(\Phi, e)$. The following lemma characterizes the set of TD(λ) limit points.

Lemma 1. *Under Assumption 1, the set of TD(λ) limit points is $\mathcal{L} := \{\theta^* + c\theta_e \cdot \mathbb{1}\{e \in W\} | c \in \mathbb{R}\}$, where $\theta^* \in E$ is the fixed point of the projected Bellman equation $\Phi\theta = \Pi_{D,W^*} T^{(\lambda)} \Phi\theta$. Here, $\Pi_{D,W^*}(\cdot)$ is the projection operator onto the subspace $W^* := \{\Phi\theta | \theta \in E\}$ with respect to the norm $\|\cdot\|_D$, and $T^{(\lambda)}$ is the TD(λ) operator defined by $T^{(\lambda)} v = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m \left(\sum_{t=0}^{m} P^t (\mathcal{R} - r(\mu) e) + P^{m+1} J \right)$.*

We consider the Lyapunov function $\Phi(\bar{r}, \theta) := (\bar{r} - r(\mu))^2 + \|\Pi_{2,E} (\theta - \theta^*)\|_2^2$. Here, $\Pi_{2,E}$ is the projection matrix onto the subspace E with respect to the 2-norm. Note that $\|\Pi_{2,E} (\theta - \theta^*)\|_2^2$ measures the distance between θ and the set of TD(λ) limit points. With the following lemma, we can show that the Lyapunov function Φ has a one-time-step negative drift.

Lemma 2. *Under Assumption 1, we have $\Delta := \min_{\|\theta\|_2 = 1, \theta \in E} \theta^\top \Phi^\top D \left(I - P^{(\lambda)} \right) \Phi\theta > 0$, where $P^{(\lambda)} = (1 - \lambda) \sum_{m=0}^{\infty} \lambda^m P^{m+1}$.*

To handle the Markovian noise, we use the conditioning argument along with the geometric mixing of the underlying Markov chain $\{X_t := (s_t, s_{t+1}, z_t)\}$ (see Lemma 6.7 in (Bertsekas & Tsitsiklis, 1996)).

Lemma 3. *Under Assumption 1, the Markov chain $\{X_t\}$ has a geometric mixing time, i.e., there exists a constant $K \geq 1$ such that given a small positive constant ϵ, $\tau(\epsilon) \leq K \ln\left(\frac{1}{\epsilon}\right)$.*

We now state two finite-time bounds on the performance of TD(λ). Part (a) studies TD(λ) applied with sufficiently small constant step-size, which is common in practice. In this case, the iterate θ_t will never converge to TD(λ) limit points due to the noise variance, but our result shows that the expected distance to the set of TD(λ) fixed points decreases at an exponential rate below some level that depends on the choice of step-size. Part (b) attains an $\tilde{\mathcal{O}}\left(\frac{1}{T}\right)$ convergence rate to TD(λ) limit points with a carefully chosen decaying step-size sequence.

Theorem 1. *Consider iterates $\{(\bar{r}_t, \theta_t)\}$ generated by Algorithm 1 with Assumption 1 and $c_\alpha \geq \Delta + \sqrt{\frac{1}{\Delta^2 (1-\lambda)^4} - \frac{1}{(1-\lambda)^2}}$. Let $\xi_1 = 8 \left(\sqrt{\bar{r}_0^2 + \|\theta_0\|_2^2} + \sqrt{r(\mu)^2 + \|\theta^*\|_2^2} + 1 \right)^2$ and $\xi_2 = 228\eta^2 \left(\sqrt{r(\mu)^2 + \|\theta^*\|_2^2} + 1 \right)^2$, where $\eta := \sqrt{c_\alpha^2 + \frac{5}{(1-\lambda)^2}}$.*

(a) If $\beta_t = \beta$ for all t, where positive constant β is properly chosen such that $\Delta\beta < 2$ and $\beta\tau(\beta) \leq \min\{\frac{1}{4\eta}, \frac{\Delta}{228\eta^2}\}$. Then, for all $T \geq \tau(\beta)$, we have $\mathbb{E}\left[(\bar{r}_T - r(\mu))^2 + \|\Pi_{2,E} (\theta_T - \theta^*)\|_2^2 \right] \leq \xi_1 \left(1 - \frac{\Delta}{2}\beta\right)^{T - \tau(\beta)} + \xi_2 \frac{\beta\tau(\beta)}{\Delta}$.

(b) If $\beta_t = \frac{c_1}{t + c_2}$ where positive constants c_1 and c_2 are properly chosen such that $2 < \Delta c_1 < 2c_2$ and there exists a smallest positive integer t^* such that $\sum_{k=0}^{t^*-1} \beta_k \leq \frac{1}{2\eta}$, and for all $t \geq t^*$, $\sum_{k=t-\tau(\beta_t)}^{t-1} \beta_k \leq \min\{\frac{1}{4\eta}, \frac{\Delta}{228\eta^2}\}$. Then, for all $T \geq t^*$, we have $\mathbb{E}\left[(\bar{r}_T - r(\mu))^2 + \|\Pi_{2,E} (\theta_T - \theta^*)\|_2^2 \right] \leq \xi_1 \left(\frac{t^* + c_2}{T + c_2} \right)^{\frac{\Delta c_1}{2}} + \xi_2 \frac{8eKc_1^2}{\Delta c_1 - 2} \frac{\ln(T + c_2) - \ln(c_1)}{T + c_2 + 1}$.

With an appropriate step-size, the following sample complexity of TD(λ) can be obtained.

Corollary 1. *For any $\epsilon > 0$, it takes at most $\tilde{\mathcal{O}}\left(\frac{K \log\left(\frac{1}{\Delta}\right) \|\theta^*\|_2^2}{\Delta^4 (1-\lambda)^4 \epsilon^2} \right)$ number of samples to find a pair $(\bar{r}_t, \theta_t)$ with $\mathbb{E}\left[(\bar{r}_t - r(\mu))^2 + \|\Pi_{2,E} (\theta_t - \theta^*)\|_2^2 \right] \leq \epsilon$.*

3. Control Algorithm: Q-learning

3.1. Problem Formulation and Synchronous Q-learning

We consider the problem of finding an optimal policy $\mu^* \in \mathcal{M}$ under the following unichain assumption (see Section 8.4 in (Puterman, 2014)).

Assumption 2. *An MDP is called unichain if the induced Markov chain consists of a single recurrent class plus a possibly empty set of transient states for any stationary deterministic policy.*

Under Assumption 2, it is known that the optimal average reward has equal components, i.e., $r^*(s) = r^*$ for all $s \in \mathcal{S}$, and that there exists $Q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (unique up to an additive constant) such that the Bellman optimality equation $Q^*(s, a) = H(Q^*)(s, a) - r^*$ holds for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, where $H(Q)(s, a) := R(s, a) + \sum_{s' \in \mathcal{S}} p(s'|s, a) \max_{a' \in \mathcal{A}} Q(s', a')$. Note that the optimal policy μ^* can be recovered by $\mu^*(s) = \text{argmax}_{a \in \mathcal{A}} Q^*(s, a)$ for all $s \in \mathcal{S}$.

Next we will use J-step synchronous Q-learning presented in Algorithm 2 to solve for Q^* approximately with a provable convergence guarantee. Notice that we include an offset function $f(\cdot)$ to ensure numerical stability (see Section 2.2 of (Abounadi et al., 2001)). If $J = 1$ and the offset function $f : \mathbb{R}^{n \times m} \to \mathbb{R}$ satisfies Assumption 2.2 in (Abounadi et al., 2001), Algorithm 2 recovers the RVI Q-learning algorithm.

Let $\mu_Q(s) := \arg\max_a Q(s, a)$ denote the Q-improving policy, and the J-step Bellman op-

erator be defined as $H^J(Q)(s,a) = \mathcal{R}(s,a) + \mathbb{E}_{s^1\sim p(\cdot|s,a),\dots,s^J\sim p(\cdot|s^{J-1},\mu_Q(s^{J-1}))}\big[\,R(s^1,\mu_Q(s^1)) + \dots + R(s^{J-1},\mu_Q(s^{J-1})) + Q(\mu^J,\mu_Q(s^J))\big]$.
Clearly, the associated Bellman optimality equation $Q(s,a) = H^J(Q)(s,a) - r^*$ has the same solutions as the above Bellman optimality equation.

3.2. Finite-Time Analysis

The existing approaches for showing convergence to a fixed point, e.g. (Chen et al., 2020), is no longer applicable because the Bellman operator H^J is indifferent to any constant shifting: $H^J(Q+ce) = H^J(Q)+ce$. And the induced Q-improving policy and its suboptimality gap are the same for any constant shifted Q (Puterman, 2014). Thus it is sensible to view all constant shifts of a Q function, $Q_{\bar{E}} := \{Q+ce : c\in\mathbb{R}\}$, as an equivalent class and directly analyze the convergence of those equivalent classes. We make a span contraction assumption on the operator H^J. Such an assumption is not restrictive because the aperiodic transformation suggested in Section 8.5.4 of (Puterman, 2014) can always be applied in the synchronous setting to ensure its satisfaction (with $J \le n$).

Assumption 3. *The Bellman operator H^J is a span contraction for some $J \ge 1$, i.e., there exists a $\gamma \in [0,1)$ such that for any Q_1 and Q_2 defined on $\mathcal{S}\times\mathcal{A}$, $\left\|H^JQ_1 - H^JQ_2\right\|_{sp} \le \gamma\left\|Q_1-Q_2\right\|_{sp}$, where $\|Q\|_{sp} := \max_{s,a} Q(s,a) - \min_{s,a} Q(s,a)$.*

Now we construct a Lyaponov function to analyze the convergence of $\{Q_t\}$. The key insight is that the span seminorm $\|\cdot\|_{sp}$ can be interpreted as the infimal convolution of the ℓ_∞-norm and the indicator function of the set $\bar{E} := \{ce : c\in\mathbb{R}\}$, i.e., $\|x\|_{sp} = (\|\cdot\|_\infty \,\square\, \delta_{\bar{E}})(x) := \inf_y \|x-y\|_\infty + \delta_{\bar{E}}(y)$, where $\delta_{\bar{E}}(x) := \begin{cases} 0, & x\in\bar{E}, \\ \infty, & \text{otherwise.}\end{cases}$ The inifimal convolution operator has many desirable properties. For example, it is commutative, associative, convexity-preserving and smoothness-preserving. These nice properties allow us to design the Lyaponov function $M_{\bar{E}}(x) := M\square\delta_{\bar{E}}(x)$ for the equivalent classes by exploiting the smoothed Lyaponuv function proposed in (Chen et al., 2020), $M(Q) := \frac{1}{2}(\|\cdot\|_\infty^2 \,\square\, \frac{\|\cdot\|_p^2}{\mu})(Q)$ with $p := 4\log(mn)$ and $\mu := (\frac{1}{2}+\frac{1}{2\gamma})^2 - 1$.

We now state the finite-time error bound for Q-learning under diminishing step-sizes. Our result shows that the mean-square of the span seminorm of the error $Q_t - Q^*$ converges at $\mathcal{O}(\frac{1}{t})$ rate.

Theorem 2. *Let $\{Q_t\}$ be generated by Algorithm 2 with decreasing step sizes $\eta_t = \frac{2}{1-\gamma}\frac{1}{t+K}$ where $K = \frac{288}{(1-\gamma)^2}\log(mn)$, then we have $\mathbb{E}\left[\|Q_t - Q^*\|_{sp}^2\right] \le C\left(\frac{\|Q_0-Q^*\|_{sp}^2\log(mn)^2}{(1-\gamma)^4t^2} + \frac{\left(J^2+\|Q^*\|_{sp}^2\right)\log(mn)}{(1-\gamma)^3t}\right)$, for some universal constant C.*

Algorithm 2: J-step Synchronous Q-learning

```
Input   : initial guess $Q_0$, step-size sequence
          $\{\eta_t\}_{t\in\mathbb{N}}$, offset function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$
for $t = 0, 1, \dots$ do
    Compute the $Q$-improving policy
      $\mu_t(s) = \arg\max_{a\in\mathcal{A}} Q_t(s,a)$.
    for $(s,a) \in \mathcal{S}\times\mathcal{A}$ do
        Sample $s^1 \sim p(\cdot|s,a), s^2 \sim$
          $p(\cdot|s^1,\mu_t(s^1)), \dots, s^J \sim$
          $p(\cdot|s^{J-1},\mu_t(s^{J-1}))$.
        Compute $Q_{t+1}(s,a) \leftarrow Q_t(s,a) +$
          $\eta_t\big[\mathcal{R}(s,a) + \sum_{j=1}^{J-1}\mathcal{R}(s^j,\mu_t(s^j)) +$
          $Q_t(s^J,\mu_t(s^J)) - Q_t(s,a) - f(Q_t)\big]$.
    end
end
```

With the fact $\|Q^*\|_{sp} \le \frac{J}{1-\gamma}$, the following sample complexity of Q-learning can be derived, which is similar to the sample complexity of discounted Q-learning algorithm (Wainwright, 2019; Chen et al., 2020).

Corollary 2. *For any $\epsilon > 0$, it takes at most $\tilde{\mathcal{O}}\left(\frac{mnJ^3}{(1-\gamma)^5\epsilon^2}\right)$ number of samples to find a Q_t such that $\mathbb{E}\left[\|Q_t - Q^*\|_{sp}\right] \le \epsilon$.*

4. Conclusion

We establish the first finite-sample convergence bounds of (i) average-reward TD(λ) with linear function approximation under Markovian observation noise, and (ii) average-reward tabular Q-learning in the synchronous setting. These RL algorithms can be viewed as stochastic approximation schemes to solve average-reward Bellman equations. However, the Bellman operators are not contractive under any norm. To resolve this difficulty, we construct Lyapunov functions using projection and infimal convolution to analyze the convergence of equivalent classes generated by these algorithms. Our approach is simple and general, so we expect it to have broader applications in other problems.

When analyzing the average-reward Q-learning algorithm, we made a J-step span contraction assumption, which is not needed for the asymptotic convergence (Abounadi et al., 2001). However, it is unclear if such an assumption is necessary for establishing any finite time convergence bound. Since our results are the first finite sample bounds, a future research direction is on relaxing this assumption.

References

Abbasi-Yadkori, Y., Bartlett, P., Bhatia, K., Lazic, N., Szepesvari, C., and Weisz, G. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pp. 3692–3702. PMLR, 2019.

Abounadi, J., Bertsekas, D., and Borkar, V. S. Learning algorithms for markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3): 681–698, 2001.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

Bertsekas, D. P. and Tsitsiklis, J. N. *Neuro-dynamic programming*. Athena Scientific, 1996.

Bhandari, J., Russo, D., and Singal, R. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

Blackwell, D. Discrete dynamic programming. *The Annals of Mathematical Statistics*, pp. 719–726, 1962.

Borkar, V. S. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

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

Brown, B. W. On the iterative method of dynamic programming on a finite space discrete time markov process. *The annals of mathematical statistics*, pp. 1279–1285, 1965.

Chang, H. S. Decentralized learning in finite markov chains: revisited. *IEEE Transactions on Automatic Control*, 54 (7):1648–1653, 2009.

Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, 2019.

Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. Finite-sample analysis of stochastic approximation using smooth convex envelopes. *arXiv preprint arXiv:2002.00874*, 2020.

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. *arXiv preprint arXiv:2102.01567*, 2021.

Giannoccaro, I. and Pontrandolfo, P. Inventory management in supply chains: a reinforcement learning approach. *International Journal of Production Economics*, 78(2): 153–161, 2002.

Gillette, D. 9. stochastic games with zero stop probabilities. In *Contributions to the Theory of Games (AM-39), Volume III*, pp. 179–188. Princeton University Press, 2016.

Gosavi, A. Reinforcement learning for long-run average cost. *European Journal of Operational Research*, 155(3): 654–674, 2004.

Howard, R. A. Dynamic programming and markov processes. 1960.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010.

Kushner, H. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media, 2003.

Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195, 1996.

Marbach, P., Mihatsch, O., and Tsitsiklis, J. N. Call admission control and routing in integrated services networks using neuro-dynamic programming. *IEEE Journal on selected areas in communications*, 18(2):197–208, 2000.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Pananjady, A. and Wainwright, M. J. Instance-dependent ℓ_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE Transactions on Information Theory*, 67 (1):566–585, 2020.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Qu, G. and Wierman, A. Finite-time analysis of asynchronous stochastic approximation and q-learning. In *Conference on Learning Theory*, pp. 3185–3205. PMLR, 2020.

Schwartz, A. A reinforcement learning method for maximizing undiscounted rewards. In *Proceedings of the tenth international conference on machine learning*, volume 298, pp. 298–305, 1993.

Singh, S. P. Reinforcement learning algorithms for average-payoff markovian decision processes. In *AAAI*, volume 94, pp. 700–705, 1994.

Srikant, R. and Ying, L. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pp. 2803–2830. PMLR, 2019.

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

Tadepalli, P., Ok, D., et al. H-learning: A reinforcement learning method to optimize undiscounted average reward. 1994.

Tangamchit, P., Dolan, J. M., and Khosla, P. K. The necessity of average rewards in cooperative multirobot learning. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, volume 2, pp. 1296–1301. IEEE, 2002.

Tsitsiklis, J. N. and Van Roy, B. Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808, 1999.

Veinott, A. F. On finding optimal policies in discrete dynamic programming with no discounting. *The Annals of Mathematical Statistics*, 37(5):1284–1294, 1966.

Wainwright, M. J. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

Wan, Y., Naik, A., and Sutton, R. S. Learning and planning in average-reward markov decision processes. *arXiv preprint arXiv:2006.16318*, 2020.

Wang, M. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017.

Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *International Conference on Machine Learning*, pp. 10170–10180. PMLR, 2020.

Wheeler, R. and Narendra, K. Decentralized learning in finite markov chains. *IEEE Transactions on Automatic Control*, 31(6):519–526, 1986.

White, D. J. Dynamic programming, markov chains, and the method of successive approximations. *Journal of Mathematical Analysis and Applications*, 6(3):373–376, 1963.

Yu, H. and Bertsekas, D. P. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7):1515–1531, 2009.

Zhang, S., Wan, Y., Sutton, R. S., and Whiteson, S. Average-reward off-policy policy evaluation with function approximation. *arXiv preprint arXiv:2101.02808*, 2021.