A Lyapunov Theory for Finite-Sample Guarantees of Asynchronous Q-Learning and TD-Learning Variants

Zaiwei Chen
Georgia Institute of Technology
zchen458@gatech.edu

Siva Theja Maguluri
Georgia Institute of Technology
siva.theja@gatech.edu

Sanjay Shakkottai
The University of Texas at Austin
sanjay.shakkottai@utexas.edu

Karthikeyan Shanmugam
IBM Research NY
Karthikeyan.Shanmugam2@ibm.com

Abstract

This paper develops an unified framework to study finite-sample convergence guarantees of a large class of value-based asynchronous Reinforcement Learning (RL) algorithms. We do this by first reformulating the RL algorithms as *Markovian* Stochastic Approximation (SA) algorithms to solve fixed-point equations. We then develop a Lyapunov analysis and derive mean-square error bounds on the convergence of the *Markovian* SA. Based on this central result, we establish finite-sample mean-square convergence bounds for asynchronous RL algorithms such as Q-learning, n-step TD, TD(λ), and off-policy TD algorithms including V-trace. As a by-product, by analyzing the performance bounds of the TD(λ) (and n-step TD) algorithm for general λ (and n), we demonstrate a bias-variance trade-off, i.e., efficiency of bootstrapping in RL. This was first posed as an open problem in [37].

1 Introduction

Reinforcement Learning (RL) is a promising approach to solve sequential decision making problems in complex and stochastic systems [38]. Despite the empirical successes of RL [32, 33], the convergence properties of RL algorithms are not well understood. In particular, even in the basic tabular setting (i.e., without using function approximation), finite-sample convergence guarantees of many popular RL algorithms are in general not established.

Most of the value-based RL algorithms can be viewed as Stochastic Approximation (SA) algorithms for solving suitable Bellman's equations. Due to the nature of sampling in RL, many such algorithms inevitably perform the so-called asynchronous update. That is, in each iteration, only a subset of the components of the vector-valued iterate is updated. Moreover, the components being updated are usually selected in a stochastic manner along a *single trajectory* based on an underlying Markov chain. Handling such asynchronous updates is one of the main challenges in analyzing the behavior of RL algorithms. In this paper, we study such asynchronous RL algorithms through the lens of Markovian SA algorithms, and develop a unified Lyapunov approach to derive finite-sample bounds on the mean-square error.

1.1 Main Contributions

Finite-Sample Bounds for Markovian SA. We establish finite-sample convergence guarantees (under various choices of stepsizes) of a stochastic approximation algorithm, which involves a contraction mapping, and is driven by both Markovian and martingale difference noise . Specifically, when using constant stepsize (i.e., $\epsilon_k \equiv \epsilon$), the convergence rate is geometric, with asymptotic accuracy roughly $O(\epsilon \log(1/\epsilon))$. When using diminishing stepsizes of the form ϵ/k^ξ (where $\xi \in (0,1]$), the convergence rate is $O(\log(k)/k^\xi)$, provided that ϵ is appropriately chosen.

Finite-Sample Bounds for Q-Learning. We establish convergence bounds of the asynchronous Q-learning algorithm. In the constant stepsize regime, when compared to [5, 4], our result improves the (best case) dependence on the size of the state-action space from $(|\mathcal{S}||\mathcal{A}|)^4$ to $(|\mathcal{S}||\mathcal{A}|)^2 \log(|\mathcal{S}||\mathcal{A}|)$ (both in the best-case scenario).

Finite-Sample Bounds for V-Trace. We establish finite-sample convergence bounds for V-trace [15], which is an off-policy variant of n-step TD. We do this showing that the *asynchronous* V-trace algorithm can be viewed as a Markovian SA algorithm involving an operator that is contractive with respect to the ℓ_∞-norm $\|\cdot\|_\infty$. This enables us to apply our central SA results.

Efficiency of Bootstrapping in RL. By deriving explicit finite-sample performance bounds of the TD(λ) (and n-step TD) algorithm as a function of λ (and n), we provide theoretical insight into the bias-variance trade-off in choosing λ (and n)[1]. For instance, in the constant-stepsize TD(λ) algorithm, after the k-th iteration, the "bias" is of the size $(1-\Theta(\frac{1}{1-\beta\lambda}))^k$ while the "variance" is of the size $\Theta(\frac{1}{(1-\beta\lambda)\log(1/(\beta\lambda))})$. Such trade-offs are also demonstrated for the V-trace and n-step TD algorithms.

1.2 Motivation and Technical Approach

Illustration via Q-Learning. To illustrate our approach of dealing with RL algorithms having asynchronous update, we use the Q-learning algorithm as a motivating example.

The Q-learning algorithm is a recursive approach for finding the optimal policy corresponding to a Markov Decision Process (MDP) (see Section 3.1 for details). At time step k, the algorithm updates a vector (of dimension state-space size $\times$ action-space size) Q_k, which is an estimate of the Q-function, using noisy samples collected along a single trajectory (aka sample-path). After a sufficient number of iterations, the vector Q_k is a close approximation of the true Q-function, which (after some straightforward computations) delivers the optimal policy for the MDP. Concretely, let $\{(S_k, A_k)\}$ be a sample trajectory of state-action pairs collected by applying some behavior policy to the model (see Section 3.1). The Q-learning algorithm performs a scalar update to the (vector-valued) iterate Q_k based on:

$$Q_{k+1}(s,a) = Q_k(s,a) + \epsilon_k \Gamma_1(Q_k, S_k, A_k, S_{k+1}) \tag{1}$$

when $(s,a) = (S_k, A_k)$, and $Q_{k+1}(s,a) = Q_k(s,a)$ otherwise. Further, $\Gamma_1(Q_k, S_k, A_k, S_{k+1}) = \mathcal{R}(S_k, A_k) + \beta \max_{a' \in \mathcal{A}} Q_k(S_{k+1}, a') - Q_k(S_k, A_k)$ is a function of the reward *sample* $\mathcal{R}(S_k, A_k)$ and the *temporal difference* in the Q-function iterate.

[1] The question of bias-variance trade-off for TD(λ) has been a long-standing open problem, first posed by [37].

At a high-level, this recursion approximates the fixed-point of the Bellman's equation through samples along a single trajectory. There are, however, two sources of noise in this approximation: (1) *asynchronous update* where only one of the components in the vector Q_k is updated (component corresponding to the state-action pair (S_k, A_k) encountered at time k), and other components in the vector Q_k are left unchanged, and (2) *stochastic noise* due to the expectation in the Bellman's operator being replaced by a single-sample estimate $\Gamma_1(\cdot)$ at time step k.

Reformulation through Markovian SA. To overcome the challenge of asynchronism (aka scalar update of the vector Q_k), our first step is to reformulate asynchronous Q-learning as a *Markovian SA algorithm* [9] [Chapter 7] by introducing an operator that captures asynchronous updates along a trajectory. A Markovian SA algorithm is an iterative approach to solve fixed-point equations (see Section 2), and leads to recursions of the form:

$$x_{k+1} = x_k + \epsilon_k(F(x_k, Y_k) - x_k + w_k). \tag{2}$$

Here x_k is the main iterate, ϵ_k is the stepsize, $F(\cdot)$ is an operator that is (in an appropriate expected sense) contractive with respect to a suitable norm, Y_k is noise derived from the evolution of a Markov Chain, and w_k is additive noise (see Section 2 for details). To cast Q-learning as a Markovian SA, let $F : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be an operator defined by $[F(Q, s_0, a_0, s_1)](s, a) = \mathbb{1}_{\{(s_0,a_0)=(s,a)\}}\Gamma_1(Q, s_0, a_0, s_1) + Q(s, a)$ for all (s, a). Then the Q-learning algorithm (1) can be rewritten as:

$$Q_{k+1} = Q_k + \epsilon_k \left(F(Q_k, S_k, A_k, S_{k+1}) - Q_k\right), \tag{3}$$

which is of the form of (2) with x_k replaced by Q_k, $w_k = 0$, and $Y_k = (S_k, A_k, S_{k+1})$. *The key takeaway is that in (3), the various noise terms (both due to performing asynchronous update and due to samples replacing an expectation in the Bellman's equation) are encoded through introducing the operator $F(\cdot)$ and the associated evolution of the Markovian noise $\{Y_k\}$.*

Analyzing the Markovian SA. To study the Markovian SA (3), let $\bar{F}(\cdot)$ be the expectation of $F(\cdot, S_k, A_k, S_{k+1})$, where the expectation is taken with respect to the steady-state distribution of the Markov chain $\{(S_k, A_k, S_{k+1})\}$. Under mild conditions, we show that $\bar{F}(Q) = N\mathcal{H}(Q) + (I - N)Q$. Here $\mathcal{H}$ is the Bellman's optimality operator for the Q-function [7]. The matrix N is a diagonal matrix with $\{p(s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ sitting on its diagonal, where $p(s, a)$ is the steady-state visitation probability of the state-action pair (s, a).

An important insight about the operator $\bar{F}(\cdot)$ is that it can be viewed as an "asynchronous" variant of the Bellman's operator $\mathcal{H}$. To see this, consider a state-action pair (s, a). The value of $[\bar{F}(Q)](s, a)$ can be interpreted as the expectation of a random variable, which takes $[\mathcal{H}(Q)](s, a)$ w.p. $p(s, a)$, and takes $Q(s, a)$ w.p. $1 - p(s, a)$. This precisely captures the asynchronous update in the Q-learning algorithm (1) in that, at steady-state, $Q_k(s, a)$ is updated w.p. $p(s, a)$, and remains unchanged otherwise. Moreover, since it is well-known that $\mathcal{H}$ is a contraction mapping w.r.t. $\|\cdot\|_\infty$, we also show that $\bar{F}$ is a contraction mapping w.r.t. $\|\cdot\|_\infty$, with the optimal Q-function being its unique fixed-point.

Thus, by recentering the iteration in (3) about $\bar{F}(\cdot)$, we have:

$$Q_{k+1} = Q_k + \underbrace{\epsilon_k \left(\bar{F}(Q_k) - Q_k\right)}_{\text{expected update}} + \underbrace{\epsilon_k(F(Q_k, S_k, A_k, S_{k+1}) - \bar{F}(Q_k))}_{\text{Markovian noise}}. \tag{4}$$

In summary, we have recast asynchronous Q-learning as an iterative update that decomposes the Q_k update into an expected update (averaged over the stationary distribution of the noise Markov chain) and a "residual *update" due to the Markovian noise. As will see in Section 2, this update equation has the interpretation of solving the fixed-point equation $\bar{F}(Q) = Q$, with Markovian noise in the update.*

Finite-Sample Bounds for Markovian SA. We use a unified *Lyapunov approach* for deriving finite-sample bounds on the update in (4). Specifically, the Lyapunov approach handles both (1) non-smooth $\|\cdot\|_\infty$-contraction of the averaged operator $\bar{F}(\cdot)$, and (2) Markovian noise that depends on the state-action trajectory. To handle $\|\cdot\|_\infty$-contraction, or more generally arbitrary norm contraction, inspired by [12], we use the Generalized Moreau Envelope as the Lyapunov function. To handle the Markovian noise, we use the conditioning argument along with the geometric mixing of the underlying Markov chain [7, 35]. Finally, for recursions beyond Q-learning, we deal with additional extraneous martingale difference noise through the tower property of the conditional expectation.

As we later discuss, beyond Q-learning, TD-learning variants such as V-trace, n-step TD, and TD(λ) can all be modeled by Markovian SA algorithms involving a contraction mapping (possibly w.r.t. different norm), and Markovian noise. Therefore, our approach unifies the finite-sample analysis of value-based RL algorithms.

1.3 Related Literature

SA Algorithms. Stochastic Approximation method was first introduced in [30] for iteratively solving systems of equations. Since then, it has been thoroughly studied in the literature due to its wide applications in optimization and machine learning [6, 23, 9, 24, 11]. Specifically, asymptotic convergence of SA algorithms involving a contraction mapping was studied in [41, 19]. As for finite-sample analysis, SA algorithms with martingale difference noise were studied in [5, 4] for bounded noise, and in [12] for unbounded noise. For SA algorithms with Markovian noise, finite-sample bounds were established for linear SA in [35, 8], and for nonlinear SA in [13] under Lipschitz and strong monotone assumptions.

RL Algorithms. The Q-learning algorithm is perhaps one of the most well-known RL algorithms in the literature [44]. The asymptotic convergence of Q-learning was established in [41, 9, 10, 19, 25]. As for finite-sample bounds, [4, 5, 43, 12] study the mean-square bounds of synchronous Q-learning. For Q-learning with asynchronous update, [4, 5] study the mean-square convergence bounds for using constant stepsize, and [16, 21, 20, 29, 27] study the high-probability bounds. When Q-learning is used along with function approximation, the asymptotic convergence and finite-sample bounds are studied in [28, 13, 25, 46, 17]. Variants of Q-learning algorithms such as double Q-learning, speedy Q-learning, and fitted Q-iteration are also studied in the literature [18, 1, 45, 14].

The V-trace algorithm [15] is a multi-step off-policy TD-learning variant. The asymptotic convergence of the V-trace algorithm was established in [15], and finite-sample bounds were derived in [12] under the assumption that synchronous update is performed. The n-step TD and TD(λ) algorithms are popular variants in the TD-learning family [36]. The asymptotic convergence of TD-learning was established in [41]. As for TD-learning with function approximation, asymptotic convergence and finite-sample bounds are studied in [42, 35, 8]. Variants of TD-learning algorithms such as LSTD(λ) and true online TD(λ) are also studied in the literature [40, 39, 31].

A key idea in n-step TD or TD(λ) is to improve the performance of the algorithm by adjusting the degree of bootstrapping, i.e., tuning the parameters λ (or n). Numerical experiments indicate that one should choose an intermediate value of λ (or n) [34]. Theoretical justification of this observation is, to some extent, provided in [22], where they study a variant of the TD(λ) algorithm called *phased* TD.

2 Markovian Stochastic Approximation

2.1 Problem Setting

Suppose we want to solve for $x^* \in \mathbb{R}^d$ in the equation

$$\mathbb{E}_{Y\sim\mu}[F(x,Y)] = x, \tag{5}$$

where $Y \in \mathcal{Y}$ is a random variable with distribution μ, and $F : \mathbb{R}^d \times \mathcal{Y} \mapsto \mathbb{R}^d$ is a general nonlinear operator. We assume the set $\mathcal{Y}$ is finite, and denote $\bar{F}(x) = \mathbb{E}_{Y\sim\mu}[F(x,Y)]$ as the expected operator.

We next present the SA algorithm for estimating x^*. Let $\{Y_k\}$ be a Markov chain with stationary distribution μ. Then the SA algorithm iteratively updates the estimate x_k by:

$$x_{k+1} = x_k + \epsilon_k \left(F(x_k, Y_k) - x_k + w_k\right), \tag{6}$$

where $\{\epsilon_k\}$ is a sequence of stepsizes, and $\{w_k\}$ is a random process representing the additive extraneous noise. To establish finite-sample convergence bounds of Algorithm (6), we next formally state our assumptions. Let $\|\cdot\|_c$ be some arbitrary norm in $\mathbb{R}^d$.

Assumption 2.1. There exist $A_1, B_1 > 0$ such that $\|F(x_1,y)-F(x_2,y)\|_c \leq A_1\|x_1-x_2\|_c$ and $\|F(\mathbf{0},y)\|_c \leq B_1$ for any $x_1, x_2 \in \mathbb{R}^d$ and $y \in \mathcal{Y}$.

Assumption 2.2. There exists $\gamma \in (0,1)$ such that $\|\bar{F}(x_1) - \bar{F}(x_2)\|_c \leq \gamma\|x_1 - x_2\|_c$ for any $x_1, x_2 \in \mathbb{R}^d$.

By Banach fixed-point theorem [2], Assumption 2.2 guarantees that the target equation (5) has a unique solution, which we have denoted by x^*.

Assumption 2.3. The Markov chain $\mathcal{M} = \{Y_k\}$ has a unique stationary distribution μ, and there exist $C > 0$ and $\sigma \in (0,1)$ such that $\max_{y\in\mathcal{Y}} \|P^k(y,\cdot) - \mu(\cdot)\|_{\text{TV}} \leq C\sigma^k$ for all $k \geq 0$.

Under Assumption 2.3, we next introduce the notion of mixing time, which will be frequently used in our derivation.

Definition 2.1. For any $\delta > 0$, the mixing time $t_\delta(\mathcal{M})$ of the Markov chain $\mathcal{M} = \{Y_k\}$ with precision δ is defined by $t_\delta(\mathcal{M}) = \min\{k \geq 0 : \max_{y\in\mathcal{Y}} \|P^k(y,\cdot) - \mu(\cdot)\|_{\text{TV}} \leq \delta\}$.

For simplicity of notation, in this section, we will just write t_δ for $t_\delta(\mathcal{M})$, and further use t_k for t_{ϵ_k}, where ϵ_k is the stepsize used in the k-th iteration of Algorithm (6).

Assumption 2.4. The random process $\{w_k\}$ satisfies $\mathbb{E}[w_k|\mathcal{F}_k] = 0$ and $\|w_k\|_c \leq A_2\|x_k\|_c + B_2$ for all $k \geq 0$, where $\mathcal{F}_k$ is the Sigma algebra generated by $\{(x_i, Y_i, w_i)\}_{0\leq i\leq k}$, and $A_2, B_2 > 0$ are constants.

Lastly, we specify the requirements for choosing the stepsize sequence $\{\epsilon_k\}$. We will consider using stepsizes of the form $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$, where $\epsilon, h > 0$ and $\xi \in [0,1]$.

Condition 2.1. *(1) Constant Stepsize.* When $\xi = 0$, there exists a threshold $\bar{\epsilon} \in (0,1)$ (defined in Appendix A.2) such that the stepsize ϵ is chosen to be in $(0, \bar{\epsilon})$. *(2) Linear Stepsize.* When $\xi = 1$, for each $\epsilon > 0$, there exists a threshold $\bar{h} > 0$ (defined in Appendix A.2) such that h is chosen to be at least $\bar{h}$. *(3) Polynomial Stepsize.* For each $\xi \in (0,1)$ and $\epsilon > 0$, there exists a threshold $\bar{h} > 0$ (defined in Appendix A.2) such that h is chosen to be at least $\bar{h}$.

2.2 Finite-Sample Convergence Bounds

For simplicity of notation, let $A = A_1 + A_2 + 1$, which can be viewed as the combined effective "Lipschitz constant", and $B = B_1 + B_2$. Let $c_1 = (\|x_0 - x^*\|_c + \|x_0\|_c + \frac{B}{A})^2$, and $c_2 = (A\|x^*\|_c + B)^2$. The constants $\{\alpha_i\}_{1 \leq i \leq 3}$ we are going to use are defined explicitly in Appendix A, and depend only on the contraction norm $\|\cdot\|_c$ and the contraction factor γ.

Theorem 2.1. *Consider $\{x_k\}$ of Algorithm (6). Suppose that Assumptions 2.1 – 2.4 are satisfied. Let $K = \min\{k : k \geq t_k\}$. Then we have the following results.*

(1) When $k \in [0, K-1]$, we have $\|x_k - x^*\|_c^2 \leq c_1$ a.s.

(2) When $k \geq K$, we have the following finite-sample convergence bounds.

(a) ***Constant Stepsize.*** *Under Condition 2.1 (1), we have:*

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 (1 - \alpha_2 \epsilon)^{k - t_\epsilon} + \frac{\alpha_3 c_2}{\alpha_2} \epsilon t_\epsilon.$$

(b) ***Linear Stepsize.*** *Under Condition 2.1 (2), we have:*

(i) when $\epsilon < 1/\alpha_2$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2 \epsilon} + \frac{8\epsilon^2 \alpha_3 c_2}{1 - \alpha_2 \epsilon} \frac{t_k}{(k+h)^{\alpha_2 \epsilon}}.$$

(ii) when $\epsilon = 1/\alpha_2$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \frac{K+h}{k+h} + 8\epsilon^2 \alpha_3 c_2 \frac{t_k \log(k+h)}{k+h}.$$

(iii) when $\epsilon > 1/\alpha_2$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2 \epsilon} + \frac{8e\epsilon^2 \alpha_3 c_2}{\alpha_2 \epsilon - 1} \frac{t_k}{k+h}.$$

(c) ***Polynomial Stepsize.*** *Under Condition 2.1 (3), we have:*

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 e^{-\frac{\alpha_2 \epsilon}{1-\xi}\left((k+h)^{1-\xi} - (K+h)^{1-\xi}\right)} + \frac{4\alpha_3 c_2 \epsilon}{\alpha_2} \frac{t_k}{(k+h)^\xi}.$$

Remark. Note that under Assumption 2.3, we can further bound the mixing time t_k of the Markov chain $\{Y_k\}$ by $\frac{\xi \log(k+h) + \log(C/(\epsilon\sigma))}{\log(1/\sigma)}$, which introduces an additional logarithmic factor.

In all cases of Theorem 2.1, we state the results as a combination of two terms. The first term is usually viewed as the "bias", and it involves the error in the initial estimate x_0 (through the constant c_1), and the geometric decay term that depends on the contraction factor γ (for constant stepsize case). The second term is usually understood as the "variance", and hence involves the constant c_2, which represents the noise variance at x^*. This form of convergence bounds is common in related literature. See for example in [11] about the results for SGD.

From case (2) (a), we see that constant stepsize is very efficient in killing the bias, but cannot drive the variance to zero. When using linear stepsize, the convergence bounds crucially depend on the value of ϵ. In order to balance the bias and the variance terms to achieve the optimal convergence rate, we need to choose $\epsilon > 1/\alpha_2$ (case (2) (b) (iii)), and the resulting optimal convergence rate is roughly $O(\frac{\log(k)}{k})$. When using polynomial stepsize, although the convergence rate is the sub-optimal $O(\frac{\log(k)}{k^\xi})$, it is more robust in the sense that it does not depend on the value of ϵ.

Switching focus, we now consider the constants $\{\alpha_i\}_{1\leq i\leq 3}$ in Theorem 2.1, which depend on the contraction norm $\|\cdot\|_c$ and the contraction factor γ. In the following lemma, we consider two cases where (1) $\|\cdot\|_c=\|\cdot\|_2$ and (2) $\|\cdot\|_c=\|\cdot\|_\infty$. Both of them will be useful when we study convergence bounds of RL algorithms. The proof of the following result is presented in Appendix A.4.1.

Lemma 2.1. *(1) When $\|\cdot\|_c=\|\cdot\|_2$, we have $\alpha_1\leq 1$, $\alpha_2\geq 1-\gamma$, and $\alpha_3\leq 228$. (2) When $\|\cdot\|_c=\|\cdot\|_\infty$, we have $\alpha_1\leq 3$, $\alpha_2\geq\frac{1-\gamma}{2}$, and $\alpha_3\leq\frac{456e\log(d)}{1-\gamma}$.*

Note that under $\|\cdot\|_\infty$-contraction, the constant α_3 can be of the size $\frac{\log(d)}{1-\gamma}$. It was argued in [12] that such $\log(d)$ factor is inevitable.

2.3 Outline of the Proof

In this section, we present the key ideas in proving Theorem 2.1. The detailed proof is presented in Appendix A. At a high level, we use a *Lyapunov* function $M:\mathbb{R}^d\mapsto\mathbb{R}$ such that the following one-step contractive inequality holds:

$$\mathbb{E}[M(x_{k+1}-x^*)]\leq(1-O(\epsilon_k)+o(\epsilon_k))\mathbb{E}[M(x_k-x^*)]+o(\epsilon_k), \tag{7}$$

which then can be repeatedly used to derive finite-sample bounds of Algorithm (6).

2.3.1 Moreau Envelope as a Lyapunov Function

Inspired by [12], we will use $M(x)=\min_{u\in\mathbb{R}^d}\{\frac{1}{2}\|u\|_c^2+\frac{1}{2\theta}\|x-u\|_p^2\}$ as the Lyapunov function, where $\theta>0$ and $p\geq 2$ are tunable parameters. The function $M(\cdot)$ is called the Generalized Moreau Envelope, which is known to be a smooth approximation of the function $\frac{1}{2}\|x\|_c^2$, with smoothness parameter $\frac{p-1}{\theta}$. See [12] for more details about using the Generalized Moreau Envelope as a Lyapunov function.

Using the smoothness property of $M(\cdot)$ and the update equation (6), we have for all $k\geq 0$:

$$\begin{aligned}&\mathbb{E}[M(x_{k+1}-x^*)]\\ &\leq\mathbb{E}[M(x_k-x^*)]+\underbrace{\epsilon_k\mathbb{E}[\langle\nabla M(x_k-x^*),\bar{F}(x_k)-x_k\rangle]}_{T_1:\text{ Expected update}}+\underbrace{\epsilon_k\mathbb{E}[\langle\nabla M(x_k-x^*),F(x_k,Y_k)-\bar{F}(x_k)\rangle]}_{T_2:\text{ Error due to Markovian noise }Y_k}\\ &\quad+\underbrace{\epsilon_k\mathbb{E}[\langle\nabla M(x_k-x^*),w_k\rangle]}_{T_3:\text{ Error due to Martingale difference noise }w_k}+\underbrace{\frac{(p-1)\epsilon_k^2}{2\theta}\mathbb{E}[\|F(x_k,Y_k)-x_k+w_k\|_p^2]}_{T_4:\text{ Error due to discretization and noises}}.\end{aligned}\tag{8}$$

What remains to do is to bound the terms T_1 to T_4. The term T_1 represents the expected update. One can show that it is negative and is of the order $O(\epsilon_k)$, hence giving the negative drift term in the target one-step contractive inequality (7). The error terms T_3 and T_4 are relatively easy to control. In fact, we have $T_3=0$ and $T_4=O(\epsilon_k^2)$. The main challenge here is to control the error term T_2, which arises due to the Markovian noise $\{Y_k\}$.

2.3.2 Handling the Markovian Noise

To control the term T_2, we need to carefully use a conditioning argument along with the geometric mixing of $\{Y_k\}$. Specifically, we first show that the error is small when we replace x_k by x_{k-t_k} in the term T_2, where t_k is the mixing time of the Markov chain $\{Y_k\}$ with precision ϵ_k. Now, consider the resulting term

$$\tilde{T}_2 = \epsilon_k \mathbb{E}[\langle \nabla M(x_{k-t_k} - x^*), F(x_{k-t_k}, Y_k) - \bar{F}(x_{k-t_k}) \rangle].$$

Taking expectation conditioning on x_{k-t_k} and Y_{k-t_k}, then we can utilize the mixing time (See definition 2.1) of $\{Y_k\}$ to show that the difference between $\mathbb{E}[F(x_{k-t_k}, Y_k) \mid x_{k-t_k}, Y_{k-t_k}]$ and $\bar{F}(x_{k-t_k})$ (i.e., $\mathbb{E}_\mu[F(x, Y)]$ evaluated at $x = x_{k-t_k}$) is of the size $O(\epsilon_k)$, hence concluding that $\tilde{T}_2 = O(\epsilon_k^2) = o(\epsilon_k)$ by the tower property of conditional expectation.

This type of conditioning argument was first introduced in [7] to establish the asymptotic convergence of linear SA with Markovian noise. Later, it was used more explicitly in [35] to study finite-sample bounds of linear SA, and in [13] to study nonlinear SA under strong monotone assumptions.

Using the upper bounds we have for the terms T_1 to T_4 in Eq. (8), we obtain the desired one-step contractive inequality (7). The rest of the proof follows by repeatedly using this inequality and evaluating the final expression for using different stepsize sequence $\{\epsilon_k\}$.

In summary, we have stated finite-sample convergence bounds of a general stochastic approximation algorithm, and highlighted the key ideas in the proof. Next, we use Theorem 2.1 as a central result to study convergence bounds of Reinforcement Learning algorithms.

3 Finite-Sample Convergence Bounds of Reinforcement Learning Algorithms

The RL problem is usually modeled by an MDP where the transition dynamics are unknown. In this work we consider an MDP consisting of a finite set of states $\mathcal{S}$, a finite set of actions $\mathcal{A}$, a set of unknown transition probability matrices that are indexed by actions, a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, and a discount factor $\beta \in (0, 1)$. We assume without loss of generality that the range of the reward function is $[0, 1]$.

The goal in RL is to find an optimal policy π^* so that the cumulative reward received by using π^* is maximized. More formally, given a policy π, define its state-value function $V_\pi : \mathcal{S} \mapsto \mathbb{R}$ by $V_\pi(s) = \mathbb{E}_\pi[\sum_{k=0}^\infty \beta^k \mathcal{R}(S_k, A_k) \mid S_0 = s]$ for all s, where $\mathbb{E}_\pi[\,\cdot\,]$ means that the actions are selected according to the policy π. Then, a policy π^* is said to be optimal if $V_{\pi^*}(s) \geq V_\pi(s)$ for any state s and policy π.

In RL, the problem of finding an optimal policy is called the control problem, which is solved with popular algorithms such as Q-learning [44]. A sub-problem is to find the value function of a given policy, which is called the prediction problem. This is solved with TD-learning and its variants such as TD(λ), n-step TD [38], and the off-policy V-trace [15]. We next show that our SA results can be used to establish finite-sample convergence bounds of all the RL algorithms listed above.

3.1 Off-Policy Control: Q-Learning

We begin by defining the Q-function associated with a policy π by $Q_\pi(s,a) = \mathbb{E}_\pi[\sum_{k=0}^\infty \beta^k \mathcal{R}(S_k, A_k)|S_0 = s, A_0 = a]$ for all (s,a). Denote Q^* as the Q-function associated with an optimal policy π^* (all optimal policies share the same optimal Q-function). The motivation of the Q-learning algorithm is based on the following result [7, 38]: π^* is an optimal policy if and only if $\pi^*(a|s) \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)$ for any (s,a). The above result implies that knowing the optimal Q-function alone is enough to compute an optimal policy.

The Q-learning algorithm is an iterative method to estimate the optimal Q-function. First, a sample trajectory $\{(S_k, A_k)\}$ is collected using a suitable behavior policy π_b. Then, initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. For each $k \geq 0$ and state-action pair (s,a), the iterate $Q_k(s,a)$ is updated by

$$Q_{k+1}(s,a) = Q_k(s,a) + \epsilon_k \Gamma_1(Q_k, S_k, A_k, S_{k+1}) \tag{9}$$

when $(s,a) = (S_k, A_k)$, and $Q_{k+1}(s,a) = Q_k(s,a)$ otherwise. Here $\Gamma_1(Q_k, S_k, A_k, S_{k+1}) = \mathcal{R}(S_k, A_k) + \beta \max_{a'\in\mathcal{A}} Q_k(S_{k+1}, a') - Q_k(S_k, A_k)$ is called the temporal difference. To establish the convergence bounds of the Q-learning algorithm, we make the following assumption.

Assumption 3.1. The behavior policy π_b satisfies $\pi_b(a|s) > 0$ for all (s,a), and the Markov chain $\mathcal{M}_S = \{S_k\}$ induced by π_b is irreducible and aperiodic.

The requirement that $\pi_b(a|s) > 0$ for all (s,a) is necessary even for the asymptotic convergence of Q-learning [41]. Since we work with finite-state MDPs, Assumption 3.1 on $\mathcal{M}_S$ implies that $\mathcal{M}_S$ has a unique stationary distribution, denoted by $\kappa_b \in \Delta^{|\mathcal{S}|}$, and $\mathcal{M}_S$ mixes at a geometric rate [26].

3.1.1 Properties of the Q-Learning Algorithm

We will follow the road map described in Section 1.2. We begin by remodelling the Q-learning algorithm. Let $Y_k = (S_k, A_k, S_{k+1})$ for all $k \geq 0$. Note that the random process $\mathcal{M}_Y = \{Y_k\}$ is also a Markov chain, whose state-space is denoted by $\mathcal{Y}$. Define an operator $F : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \times \mathcal{Y} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by $[F(Q,y)](s,a) = [F(Q, s_0, a_0, s_1)](s,a) = \mathbb{1}_{\{(s_0,a_0)=(s,a)\}} \Gamma_1(Q, s_0, a_0, s_1) + Q(s,a)$ for all (s,a). Then the Q-learning algorithm (9) can be written by

$$Q_{k+1} = Q_k + \epsilon_k \left(F(Q_k, Y_k) - Q_k\right),$$

which is in the same form of the SA algorithm (6) with w_k being identically equal to zero. Next, we establish the properties of the operator $F(\cdot,\cdot)$ and the Markov chain $\{Y_k\}$ in the following proposition, which guarantees that Assumptions 2.1 – 2.3 are satisfied in the context of Q-learning.

Let $N \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ be the diagonal matrix with $\{\kappa_b(s)\pi_b(a|s)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ sitting on its diagonal. Let $N_{\min} = \min_{(s,a)} \kappa_b(s)\pi_b(a|s)$. Note that we have $N_{\min} > 0$ under Assumption 3.1. The proof of the following proposition is presented in Appendix B.1.

Proposition 3.1. *Suppose that Assumption 3.1 is satisfied, Then we have the following results.*

(1) The operator $F(\cdot,\cdot)$ satisfies $\|F(Q_1,y) - F(Q_2,y)\|_\infty \leq 2\|Q_1 - Q_2\|_\infty$ and $\|F(\mathbf{0},y)\|_\infty \leq 1$ for any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and $y \in \mathcal{Y}$.

(2) The Markov chain $\mathcal{M}_Y = \{Y_k\}$ has a unique stationary distribution μ, and there exist $C_1 > 0$ and $\sigma_1 \in (0,1)$ s.t. $\max_{y\in\mathcal{Y}} \|P^{k+1}(y,\cdot) - \mu(\cdot)\|_{TV} \leq C_1\sigma_1^k$ for any $k \geq 0$.

(3) Define the expected operator $\bar{F}:\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}\mapsto\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ of $F(\cdot,\cdot)$ by $\bar{F}(Q)=\mathbb{E}_{Y\sim\mu}[F(Q,Y)]$. Then

(a) $\bar{F}(\cdot)$ is explicitly given by $\bar{F}(Q)=N\mathcal{H}(Q)+(I-N)Q$, where $\mathcal{H}$ is the Bellman's operator for the Q-function.

(b) $\bar{F}(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_\infty$, with contraction factor $\gamma_1:=1-N_{\min}(1-\beta)$.

(c) $\bar{F}(\cdot)$ has a unique fixed-point Q^*.

3.1.2 Finite-Sample Bounds of Q-Learning

Proposition 3.1 enables us to apply Theorem 2.1 and Lemma 2.1 (2) to the Q-learning algorithm. For ease of exposition, we only present the result of using constant stepsize. The proof of the following result and the result for using diminishing stepsizes are presented in Appendix B.2.

Theorem 3.1. *Consider $\{Q_k\}$ of Algorithm (9). Suppose that Assumption 3.1 is satisfied, and $\epsilon_k=\epsilon$ for all $k\geq 0$, where ϵ is chosen properly. Then there exists $K_1>0$ such that the following inequality holds for all $k\geq K_1$:*

$$\mathbb{E}[\|Q_k-Q^*\|_\infty^2]\leq c_{Q,1}(1-\frac{(1-\gamma_1)\epsilon}{2})^{k-K_1}+c_{Q,2}\frac{\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma_1)^2}\epsilon t_\epsilon(\mathcal{M}_Y),$$

where $c_{Q,1}=3(\|Q_0-Q^*\|_\infty+\|Q_0\|_\infty+1)^2$ and $c_{Q,2}=912e(3\|Q^*\|_\infty+1)^2$.

Remark. Recall that $t_\epsilon(\mathcal{M}_Y)$ is the mixing time of the Markov chain $\{Y_k\}$ with precision ϵ. Using Proposition 3.1 (2), we see that $t_\epsilon(\mathcal{M}_Y)$ produces an additional $\log(1/\epsilon)$ factor in the bound.

Rate of Convergence. Similar to Theorem 2.1, we view the first term on the RHS of the convergence bound as the the "bias", and the second term as the "variance". Since we are using constant stepsize, the bias term goes to zero geometrically fast while the variance is of the size $O(\epsilon\log(1/\epsilon))$.

Dependence on the Contraction Factor. The contraction factor appears as $\frac{1}{(1-\gamma_1)^2}$ in the "variance". It is argued in [43] that for SA algorithms under $\|\cdot\|_\infty$-contraction, such $\frac{1}{(1-\gamma)^2}$ factor is in general unimprovable.

Dependence on the Size of the State-Action Space. Observe that we have $\log(|\mathcal{S}||\mathcal{A}|)$ in our bound, which is also present in synchronous Q-learning [43, 12]. Besides the logarithmic factor, since $(1-\gamma_1)^2$ appears in the denominator of the variance term, and $1-\gamma_1=N_{\min}(1-\beta)\leq(1-\beta)/(|\mathcal{S}||\mathcal{A}|)$ (Proposition 3.1), the actual dependence on the size of the state-action space is at least $(|\mathcal{S}||\mathcal{A}|)^2\log(|\mathcal{S}||\mathcal{A}|)$. This improves the results in [4, 5], where they have at least $(|\mathcal{S}||\mathcal{A}|)^4$ dependence.

3.2 Off-Policy Prediction: V-Trace

We next switch our focus to solving the prediction problem using TD-learning variants. Specifically, we first consider the V-trace algorithm for off-policy TD-learning, which subsumes the on-policy n-step TD as a special case.

The V-trace algorithm, proposed in [15], is a multi-step TD-learning algorithm which uses off-policy sampling with truncated Importance Sampling (IS) ratios.

Let π_b be a behavior policy used to collect samples, π be the target policy (i.e., we want to evaluate V_π), and n be a positive integer. Let $c(s,a)=\min(\bar{c},\frac{\pi(a|s)}{\pi_b(a|s)})$ and $\rho(s,a)=\min(\bar{\rho},\frac{\pi(a|s)}{\pi_b(a|s)})$ be the truncated IS ratios at (s,a), where $\bar{\rho}\geq\bar{c}\geq 1$ are the two truncation levels. Suppose a sequence of state-action pairs $\{(S_k,A_k)\}$

is collected under the behavior policy π_b. Then, with initialization $V_0 \in \mathbb{R}^{|\mathcal{S}|}$, for each $k \geq 0$ and $s \in \mathcal{S}$, the V-trace algorithm updates the estimate $V_k(s)$ by

$$V_{k+1}(s) = V_k(s) + \epsilon_k \sum_{i=k}^{k+n-1} \beta^{i-k} \eta_{k,i} \Gamma_2(V_k, S_i, A_i, S_{i+1}) \tag{10}$$

when $s = S_k$, and $V_{k+1}(s) = V_k(s)$ otherwise. Here $\eta_{k,i} = \prod_{j=k}^{i-1} c(S_j, A_j)\rho(S_i, A_i)$ is the product of the truncated IS ratios, and $\Gamma_2(V_k, S_i, A_i, S_{i+1}) = \mathcal{R}(S_i, A_i) + \beta V_k(S_{i+1}) - V_k(S_i)$ is the temporal difference. Note that when $\pi_b = \pi$, and $\bar{c} = \bar{\rho} = 1$, Eq. (10) reduces to the update equation for the on-policy n-step TD [38].

To establish finite-sample convergence bounds of Algorithm (10), we make the following assumption.

Assumption 3.2. The behavior policy π_b satisfies for all $s \in \mathcal{S}$: $\{a \in \mathcal{A} \mid \pi(a|s) > 0\} \subseteq \{a \in \mathcal{A} \mid \pi_b(a|s) > 0\}$, and the Markov chain $\mathcal{M}_{\mathcal{S}} = \{S_k\}$ induced by π_b is irreducible and aperiodic.

The first part of Assumption 3.2 is call the *coverage* assumption, which states that, for any state, if it is possible to explore a specific action under the target policy π, then it is also possible to explore such an action under the behavior policy π_b. This requirement is necessary for off-policy RL. The second part of Assumption 3.2 implies that $\{S_k\}$ has a unique stationary distribution, denoted by $\kappa_b \in \Delta^{|\mathcal{S}|}$. Moreover, the Markov chain $\{S_k\}$ mixes at a geometric rate [26].

3.2.1 Finite-Sample Bounds of V-Trace

To establish the convergence bounds of the V-trace algorithm, similarly as in the Q-learning section, we first rewrite the V-trace algorithm in the form of SA algorithm (6). Then we investigate the properties of the corresponding operator $F(\cdot, \cdot)$ and the Markov chain $\{Y_k\}$, and establish a proposition analogous to Proposition 3.1. In particular, we show that the "asynchronous" Bellman's operator $\bar{F}(\cdot)$ associated with the V-trace algorithm is a contraction with respect to $\|\cdot\|_\infty$. A similar property for synchronous V-trace was shown in [15]. See Appendix C for details.

We next present the finite-sample convergence guarantees of the V-trace algorithm. In order to do so, we need to introduce more notations. For any $s \in \mathcal{S}$, let $C(s) = \sum_{a \in \mathcal{A}} \min(\bar{c}\pi_b(a|s), \pi(a|s))$, and $D(s) = \sum_{a \in \mathcal{A}} \min(\bar{\rho}\pi_b(a|s), \pi(a|s))$. Let $C_{\min} = \min_{s \in \mathcal{S}} C(s)$, $D_{\min} = \min_{s \in \mathcal{S}} D(s)$, and $\kappa_{\min} = \min_{s \in \mathcal{S}} \kappa_b(s)$. It is clear that $0 < C_{\min} \leq D_{\min} \leq 1$, and $\kappa_{\min} > 0$ (Assumption 3.2). Moreover, let $\pi_{\bar{\rho}}$ be a policy defined by $\pi_{\bar{\rho}}(a|s) = \frac{\min(\bar{\rho}\pi_b(a|s), \pi(a|s))}{D(s)}$ for any (s, a). Let $\gamma_2 = 1 - \kappa_{\min} \frac{(1-\beta)(1-(\beta C_{\min})^n) D_{\min}}{1 - \beta C_{\min}}$, which is the contraction factor of the "asynchronous" Bellman's operator $\bar{F}(\cdot)$ associated with the V-trace algorithm.

We next present the convergence bounds of V-trace for using constant stepsize, whose proof and the result for using diminishing stepsize are presented in Appendix C.

Theorem 3.2. *Consider $\{V_k\}$ of Algorithm (10). Suppose that Assumption 3.2 is satisfied, and $\epsilon_k \equiv \epsilon$ with properly chosen ϵ. Then there exists $K_2 > 0$ such that the following inequality holds for all $k \geq K_2$:*

$$\mathbb{E}[\|V_k - V_{\pi_{\bar{\rho}}}\|_\infty^2] \leq c_{V,1}(1 - \frac{1-\gamma_2}{2}\epsilon)^{k-K_2} + c_{V,2} \frac{\log(|\mathcal{S}|)(\bar{\rho}+1)^2 z(\beta, \bar{c})^2}{(1-\gamma_2)^2} \epsilon (t_\epsilon(\mathcal{M}_S) + n)$$

where $c_{V,1} = 3(\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2$, $c_{V,2} = 3648e(\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2$, $z(\beta, \bar{c}) = n$ when $\beta\bar{c} = 1$ and $= \frac{1-(\beta\bar{c})^n}{1-\beta\bar{c}}$ otherwise.

Remark. Similarly as in Q-learning, under Assumption 3.2, we can further bound the mixing time $t_\epsilon(\mathcal{M}_S)$ by $L(\log(1/\epsilon)+1)$, where $L>0$ is a constant.

The rate of convergence (i.e., geometric convergence with asymptotic accuracy $O(\epsilon\log(1/\epsilon))$) and the dependence on the size of the state-space (i.e., $|\mathcal{S}|^2\log(|\mathcal{S}|)$) are similar to that of Q-learning. Here we focus on the dependence on the truncation levels $\bar{\rho}$ and $\bar{c}$, and the parameter n.

Dependence on the Truncation Levels. The truncation level $\bar{\rho}$ appears quadratically in the bound, and determines the limit point $V_{\pi_{\bar{\rho}}}$. Moreover, the bias of the limit point $V_{\pi_{\bar{\rho}}}$ can be controlled by $\|V_{\pi_{\bar{\rho}}}-V_\pi\|_\infty\leq\frac{1}{(1-\beta)^2}\|\pi-\pi_{\bar{\rho}}\|_\infty$ (See [12] for its proof). The truncation level $\bar{c}$ appears either linearly or exponentially in the bound through the term $z(\beta,\bar{c})$. These observations agree with results in [12], where synchronous V-trace is studied.

Dependence on the Parameter n. To analyze the impact of the parameter n, we begin by rewriting the convergence bounds in Theorem 3.2 focusing only on n-dependent terms. Using the explicit expression of the contraction factor γ_2, in the k-th iteration, the bias term is of the size $(1-\Theta(1-(\beta C_{\min})^n))^k$. Since the mixing time $t_\epsilon(\mathcal{M}_S)$ of the original Markov chain $\{S_k\}$ does not depend on n, the variance term is of the size $\Theta(\frac{n}{1-(\beta C_{\min})^n})$. Now we can clearly see that as n increases, the bias goes down while the variance goes up, thereby demonstrating a bias-variance trade-off in the V-trace algorithm. Since on-policy n-step TD is a special case of the V-trace algorithm, such trade-off is also present in n-step TD. See Appendix D for more details.

3.3 On-Policy Prediction: TD(λ)

We next consider the on-policy TD(λ) algorithm, which effectively uses a convex combination of all the multi-step temporal differences at each update.

We begin by describing the TD(λ) algorithm for estimating the value function V_π of a policy π. Suppose we have collected a sample trajectory $\{(S_k,A_k)\}$ using the policy π. Then, with initialization $V_0\in\mathbb{R}^{|\mathcal{S}|}$, for any $\lambda\in(0,1)$, the estimate V_k is iteratively updated according to

$$V_{k+1}(s)=V_k(s)+\epsilon_k z_k(s)\Gamma_4(V_k,S_k,A_k,S_{k+1}) \tag{11}$$

for all $s\in\mathcal{S}$, where $\Gamma_4(V_k,S_k,A_k,S_{k+1})=\mathcal{R}(S_k,A_k)+\beta V_k(S_{k+1})-V_k(S_k)$ is the temporal difference, and $z_k(s)=\sum_{i=0}^k(\beta\lambda)^{k-i}\mathbb{1}_{\{S_i=s\}}$ is the eligibility trace [7, 38].

A key idea in the TD(λ) algorithm is to use the parameter λ to adjust the bootstrapping effect. When $\lambda=0$, Algorithm (11) becomes the standard TD(0) update, which is pure bootstrapping. Another extreme case is when $\lambda=1$. This corresponds to using pure Monte Carlo method. Theoretical understanding of the efficiency of bootstrapping is a long-standing open problem in RL [37].

In the following section, we establish finite-sample convergence bounds of the TD(λ) algorithm. By evaluating the resulting bound as a function of λ, we provide theoretical insight into the bias-variance trade-off in choosing λ. To achieve that, we need the following assumption.

Assumption 3.3. The Markov chain $\mathcal{M}_\mathcal{S}=\{S_k\}$ induced by the target policy π is irreducible and aperiodic.

As a result of Assumption 3.3, the Markov chain $\{S_k\}$ has a unique stationary distribution, denoted by $\kappa\in\Delta^{|\mathcal{S}|}$, and the geometric mixing property [26].

3.3.1 Finite-Sample Bounds of TD(λ)

Unlike Q-learning and V-trace, TD(λ) cannot be viewed as a direct variant of the SA algorithm (6). This is because of the geometric averaging induced by the eligibility trace in TD(λ), which creates dependencies over the *entire* past trajectory. We overcome this difficulty by using an additional truncation argument. We show that the truncated "asynchronous" Bellman's operator $\bar{F}(\cdot)$ associated with the TD(λ) algorithm is a contraction with respect to the ℓ_p-norm $\|\cdot\|_p$ for any $p \geq 1$, with a common contraction factor $\gamma_4 = 1 - \kappa_{\min}\frac{1-\beta}{2(1-\beta\lambda)}$, where $\kappa_{\min} = \min_{s\in\mathcal{S}} \kappa(s)$. This enables us to use our SA approach. The residual error due to truncation can be separately handled.

Next we present the finite-sample convergence bound of the TD(λ) algorithm for using constant stepsize, where we exploit only the $\|\cdot\|_2$-contraction property. The proof is presented in Appendix E.2.

Theorem 3.3. *Consider $\{V_k\}$ of Algorithm (11). Suppose that Assumption 3.3 is satisfied and $\epsilon_k \equiv \epsilon$ for properly chosen ϵ. Then there exists $K_4 > 0$ such that the following inequality holds for all $k \geq K_4$:*

$$\mathbb{E}[\|V_k - V_\pi\|_2^2] \leq \tilde{c}_1 \left(1 - (1-\gamma_4)\epsilon\right)^{k-K_4} + \tilde{c}_2 \frac{\epsilon \left(t_\epsilon(\mathcal{M}_S) + \frac{\log(1/\epsilon)}{\log(1/(\beta\lambda))}\right)}{(1-\beta\lambda)^2(1-\gamma_4)},$$

where $\tilde{c}_1 = (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 1)^2$ and $\tilde{c}_2 = 112(4\|V_\pi\|_2 + 1)^2$.

Remark. Under Assumption 3.3, the mixing time $t_\epsilon(\mathcal{M}_S)$ is at most an affine function of $\log(1/\epsilon)$. More importantly, it does not depend on the choice of λ.

The convergence rate is similar to that of Q-learning and V-trace. We here focus on the dependence on the size of the state-space, and the parameter λ.

Dependence on the Size of the State-Space. There is no explicit dependence on $|\mathcal{S}|$ as in V-trace. Implicitly, since $1-\gamma_4 \propto \kappa_{\min} \leq 1/|\mathcal{S}|$, we have at least an $|\mathcal{S}|$ factor in the bound (more specifically in the "variance" term). This is different from the V-trace algorithm. The reason is that we are able to exploit the $\|\cdot\|_2$-contraction property of the "asynchronous" Bellman's operator $\bar{F}(\cdot)$ in the TD(λ) algorithm, while in the V-trace algorithm we only have $\|\cdot\|_\infty$-contraction. The $\|\cdot\|_\infty$-contraction property produces an additional $\log(d)$ factor as stated in Lemma 2.1.

Dependence on the Parameter λ. We begin by rewriting both the "bias" term and the "variance" term in the resulting convergence bound of Theorem 3.3 focusing only on λ-dependent terms. Then, the bias term is of the size $(1-\Theta(\frac{1}{1-\beta\lambda}))^k$ while the variance term is between $\Theta(\frac{1}{(1-\beta\lambda)\log(1/(\beta\lambda))})$ and $\Theta(\frac{1}{1-\beta\lambda})$. Now observe that the bias term is in favor of large λ (i.e., less bootstrapping, more Monte Carlo) while the variance term is in favor of small λ (i.e., more bootstrapping, less Monte Carlo). This observation agrees with empirical results in the literature [38, 22]. Therefore, we demonstrate a bias-variance trade-off in choosing λ, which addresses one of the open problems in [37] on the efficiency of bootstrapping in RL.

References

[1] Azar, M. G., Munos, R., Ghavamzadeh, M., and Kappen, H. J. (2011). Speedy Q-learning. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, pages 2411–2419.

[2] Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. math*, 3(1):133–181.

[3] Beck, A. (2017). *First-order methods in optimization*, volume 25. SIAM.

[4] Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208.

[5] Beck, C. L. and Srikant, R. (2013). Improved upper bounds on the expected error in constant step-size Q-learning. In *2013 American Control Conference*, pages 1926–1931. IEEE.

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

[7] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

[8] Bhandari, J., Russo, D., and Singal, R. (2018). A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. In *Conference On Learning Theory*, pages 1691–1692.

[9] Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

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

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

[12] Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2020). Finite-Sample Analysis of Contractive Stochastic Approximation Using Smooth Convex Envelopes. *Advances in Neural Information Processing Systems*, 33.

[13] Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. (2019). Finite-Sample Analysis of Nonlinear Stochastic Approximation with Applications in Reinforcement Learning. *Preprint arXiv:1905.11425*.

[14] Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556.

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

[16] Even-Dar, E. and Mansour, Y. (2003). Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25.

[17] Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). A theoretical analysis of deep Q-learning. In *Learning for Dynamics and Control*, pages 486–489. PMLR.

[18] Hasselt, H. (2010). Double Q-learning. *Advances in neural information processing systems*, 23:2613–2621.

[19] Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710.

[20] Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878.

[21] Kearns, M. and Singh, S. (1999). Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, pages 996–1002.

[22] Kearns, M. J. and Singh, S. P. (2000). Bias-Variance Error Bounds for Temporal Difference Updates. In *COLT*, pages 142–147. Citeseer.

[23] Kushner, H. (2010). Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):87–96.

[24] Lan, G. (2020). *First-order and Stochastic Optimization Methods for Machine Learning*. Springer.

[25] Lee, D. and He, N. (2019). A Unified Switching System Perspective and ODE Analysis of Q-Learning Algorithms. *Preprint arXiv:1912.02270*.

[26] Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

[27] Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Preprint arXiv:2006.03041*.

[28] Melo, F. S., Meyn, S. P., and Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671.

[29] Qu, G. and Wierman, A. (2020). Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR.

[30] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407.

[31] Seijen, H. and Sutton, R. (2014). True online TD(λ). In *International Conference on Machine Learning*, pages 692–700. PMLR.

[32] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

[33] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, 550(7676):354.

[34] Singh, S. P. and Sutton, R. S. (1996). Reinforcement learning with replacing eligibility traces. *Machine learning*, 22(1):123–158.

[35] Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830.

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

[37] Sutton, R. S. (1999). Open theoretical questions in reinforcement learning. In *European Conference on Computational Learning Theory*, pages 11–17. Springer.

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

[39] Tagorti, M. and Scherrer, B. (2015). On the Rate of Convergence and Error Bounds for LSTD(λ). In *International Conference on Machine Learning*, pages 1521–1529. PMLR.

[40] Thiery, C. and Scherrer, B. (2010). Least-squares λ policy iteration: Bias-variance trade-off in control problems. In *International Conference on Machine Learning*.

[41] Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202.

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

[43] Wainwright, M. J. (2019). Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *Preprint arXiv:1905.06265.*

[44] Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

[45] Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Finite-time analysis for double Q-learning. *Advances in Neural Information Processing Systems*, 33.

[46] Xu, P. and Gu, Q. (2020). A finite-time analysis of Q-learning with neural network function approximation. In *International Conference on Machine Learning*, pages 10555–10565. PMLR.

Appendices

A Proof of Theorem 2.1

We will state and prove a more general version of Theorem 2.1. To do that, we need to introduce more notations and explicitly specify the requirement for choosing the stepsize sequence $\{\epsilon_k\}$.

More Notations. Let $g(x)=\frac{1}{2}\|x\|_s^2$, where the norm $\|\cdot\|_s$ is properly chosen so that the function $g(\cdot)$ is a strongly smooth function with respect to the norm $\|\cdot\|_s$. That is, the function $g(\cdot)$ is convex, differentiable, and there exists $L>0$ such that $g(x_2)\leq g(x_1)+\langle\nabla g(x_1),x_2-x_1\rangle+\frac{L}{2}\|x_1-x_2\|_s^2$ for any $x_1,x_2\in\mathbb{R}^d$. For example, ℓ_p-norm with $p\in[2,\infty)$ works [3]. Since we work with finite-dimensional space $\mathbb{R}^d$, there exist $\ell_{cs},u_{cs}>0$ such that $\ell_{cs}\|\cdot\|_c\leq\|\cdot\|_c\leq u_{cs}\|\cdot\|_s$. Let $\theta>0$ be chosen such that $\gamma^2<\frac{1+\theta\ell_{cs}^2}{1+\theta u_{cs}^2}$, which is always possible since $\gamma\in(0,1)$. Denote $\alpha_1=\frac{1+\theta u_{cs}^2}{1+\theta\ell_{cs}^2}$, $\alpha_2=1-\gamma\alpha_1^{1/2}$, and $\alpha_3=\frac{114L(1+\theta u_{cs}^2)}{\theta\ell_{cs}^2}$. Note that $\alpha_2\in(0,1)$ under our choice of θ.

Now we state the requirement in choosing the stepsizes. For simplicity, we use $\epsilon_{i,j}$ for $\sum_{k=i}^j\epsilon_k$.

Condition A.1. The stepsize sequence $\{\epsilon_k\}$ is non-increasing and satisfies $\epsilon_{k-t_k,k-1}\leq\min(\frac{\alpha_2}{\alpha_3A^2},\frac{1}{4A})$ for all $k\geq t_k$.

Now we are ready to state a more general version of Theorem 2.1.

Theorem A.1. *Consider $\{x_k\}$ generated by Algorithm (6). Suppose that Assumptions 2.1 – 2.4 are satisfied, and the stepsize sequence $\{\epsilon_k\}$ satisfies Condition A.1. Let $K = \min\{k \geq 0 \ : \ t_k \leq k\}$, then we have the following results.*

(1) For any $k \in [0, K-1]$, we have: $\|x_k - x^*\|_c^2 \leq c_1$ a.s.

(2) For any $k \geq K$, we have

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \prod_{j=K}^{k-1}(1-\alpha_2\epsilon_j) + \alpha_3 c_2 \sum_{i=K}^{k-1} \epsilon_i \epsilon_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\alpha_2\epsilon_j),$$

where $c_1 = (\|x_0 - x^*\|_c + \|x_0\|_c + \frac{B}{A})^2$ and $c_2 = (A\|x^*\|_c + B)^2$.

Once we have Theorem A.1, we can evaluate the bound when $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$ to get Theorem 2.1. This is presented in Appendix A.3. In Appendix A.3, we also show how Condition 2.1 is obtained from Condition A.1 and the explicit requirements on the thresholds $\bar{c}$ and $\bar{h}$. We next present the complete proof of Theorem A.1.

A.1 Proof of Theorem A.1

A.1.1 Step One: Constructing a Valid Lyapunov Function

We will use the Generalized Moreau Envelope of $f(\cdot)$ with respect to $g(\cdot)$, i.e., $M_f^{\theta,g}(x) = \min_{u\in\mathbb{R}^d}\{f(u) + \frac{1}{\theta}g(x-u)\}$ as the Lyapunov function to study stochastic iterative algorithm (6). We first summarize the properties of $M_f^{\theta,g}(\cdot)$ in the following proposition, which was established in [12]. For simplicity, we will just write $M(\cdot)$ for $M_f^{\theta,g}(\cdot)$ in the following unless we want to emphasize the dependence on the choices of θ and $g(\cdot)$.

Proposition A.1. *The function $M(x)$ has the following properties.*

(1) $M(x)$ is convex, and $\frac{L}{\theta}$-smooth w.r.t. $\|\cdot\|_s$, i.e., $M(y) \leq M(x) + \langle \nabla M(x), y - x\rangle + \frac{L}{2\theta}\|x-y\|_s^2$ for all $x, y \in \mathbb{R}^d$.

(2) There exists a norm, denoted by $\|\cdot\|_m$, such that $M(x) = \frac{1}{2}\|x\|_m^2$.

(3) Let $\ell_{cm} = (1+\theta\ell_{cs}^2)^{1/2}$ and $u_{cm} = (1+\theta u_{cs}^2)^{1/2}$. Then it holds that $\ell_{cm}\|\cdot\|_m \leq \|\cdot\|_c \leq u_{cm}\|\cdot\|_m$.

Using Proposition A.1 (1) and the update equation (6), we have for any $k \geq 0$:

$$\begin{aligned}
M(x_{k+1} - x^*) &\leq M(x_k - x^*) + \langle \nabla M(x_k - x^*), x_{k+1} - x_k\rangle + \frac{L}{2\theta}\|x_{k+1} - x_k\|_s^2 \\
&= M(x_k - x^*) + \epsilon_k\langle \nabla M(x_k - x^*), F(x_k, Y_k) - x_k + w_k\rangle + \frac{L\epsilon_k^2}{2\theta}\|F(x_k, Y_k) - x_k + w_k\|_s^2 \\
&= M(x_k - x^*) + \underbrace{\epsilon_k\langle \nabla M(x_k - x^*), \bar{F}(x_k) - x_k\rangle}_{T_1\text{: Expected update}} + \underbrace{\epsilon_k\langle \nabla M(x_k - x^*), w_k\rangle}_{T_2\text{: Error due to Martingale difference noise } w_k} \\
&\quad + \underbrace{\epsilon_k\langle \nabla M(x_k - x^*), F(x_k, Y_k) - \bar{F}(x_k)\rangle}_{T_3\text{: Error due to Markovian noise } Y_k} + \underbrace{\frac{L\epsilon_k^2}{2\theta}\|F(x_k, Y_k) - x_k + w_k\|_s^2}_{T_4\text{: Error due to discretization and noises}}.
\end{aligned} \tag{12}$$

The term T_1 represents the expected update of the stochastic iterative algorithm (6), and is bounded in the following lemma, whose proof can be found in [12].

Lemma A.1. *The following inequality holds for all $k \geq 0$: $T_1 \leq -2\left(1-\gamma\frac{u_{cm}}{\ell_{cm}}\right)\epsilon_k M(x_k - x^*)$.*

As we have seen in Lemma A.1, the term T_1 provides us the desired negative drift, i.e., the $-O(\epsilon_k)$ term in the target one-step contractive inequality (7). What remains to do is to control all the error terms T_2 to T_4 in Eq. (12).

A.1.2 Step Two: Bounding the Error Terms

We begin with the term T_2. Since $\{w_k\}$ is a martingale difference sequence with respect to the filtration $\mathcal{F}_k$ (Assumption 2.4), while x_k is measurable with respect to $\mathcal{F}_k$, we have by the tower property of conditional expectation that

$$\mathbb{E}[T_2] = \mathbb{E}[\mathbb{E}[T_2 \mid \mathcal{F}_k]] = \epsilon_k \mathbb{E}[\langle \nabla M_f^{\theta,g}(x_k - x^*), \mathbb{E}[w_k \mid \mathcal{F}_k]]\rangle = 0.$$

Next we analyze the error term T_3, which is due to the Markovian noise $\{Y_k\}$. We first decompose T_3 in the following way:

$$\begin{aligned}
T_3 &= \epsilon_k \langle \nabla M(x_k - x^*), F(x_k, Y_k) - \bar{F}(x_k)\rangle \\
&= \epsilon_k \underbrace{\langle \nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*), F(x_k, Y_k) - \bar{F}(x_k)\rangle}_{T_{31}} \\
&\quad + \epsilon_k \underbrace{\langle \nabla M(x_{k-t_k} - x^*), F(x_k, Y_k) - F(x_{k-t_k}, Y_k) + \bar{F}(x_{k-t_k}) - \bar{F}(x_k)\rangle}_{T_{32}} \\
&\quad + \epsilon_k \underbrace{\langle \nabla M(x_{k-t_k} - x^*), F(x_{k-t_k}, Y_k) - \bar{F}(x_{k-t_k})\rangle}_{T_{33}}. \qquad (13)
\end{aligned}$$

To proceed, we need the following lemma, which allows us to control the difference between x_{k_1} and x_{k_2} when $|k_1 - k_2|$ is relatively small. The proof can be found in Appendix A.4.2.

Lemma A.2. *Given non-negative integers $k_1 \leq k_2$ satisfying $\epsilon_{k_1,k_2-1} \leq \frac{1}{4A}$, we have for all $k \in [k_1, k_2]$:*

$$\|x_k - x_{k_1}\|_c \leq 2\epsilon_{k_1,k_2-1}(A\|x_{k_1}\|_c + B), \text{ and } \|x_k - x_{k_1}\|_c \leq 4\epsilon_{k_1,k_2-1}(A\|x_{k_2}\|_c + B).$$

Using the assumption that $\epsilon_{k_1,k_2-1} \leq \frac{1}{4A}$ in the resulting inequality of Lemma A.2, we have the following corollary, which will also be frequently used during the derivation.

Corollary A.1.1. *Given non-negative integers $k_1 \leq k_2$ satisfying $\epsilon_{k_1,k_2-1} \leq \frac{1}{4A}$, we have for all $k \in [k_1, k_2]$:*

$$\|x_k - x_{k_1}\|_c \leq \max(\|x_{k_1}\|_c, \|x_{k_2}\|_c) + \frac{B}{A}.$$

Recall that we require $\epsilon_{k-t_k,k-1} \leq \frac{1}{4A}$ for all $k \geq t_k$ in Condition A.1. Therefore, Lemma A.2 is applicable when $k_1 = k - t_k$ and $k_2 = k - 1$ for any $k \geq t_k$.

Now we are ready to control the terms T_{31}, T_{32}, and T_{33} in the following lemma. The terms T_{31} and T_{32} are controlled mainly by constantly applying Lemma A.2 and the Lipschitz property of the operator F (Assumptions 2.1). Bounding the term T_{33} requires using the geometric mixing of the Markov chain $\{Y_k\}$ (Assumption 2.3). The proof is presented in Appendix A.4.3.

Lemma A.3. *The following inequalities hold for all $k \geq t_k$:*

(1) $T_{31} \leq \frac{16LA^2u_{cm}^2\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}M(x_k - x^*) + \frac{8L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2$,

(2) $T_{32} \leq \frac{64LA^2u_{cm}^2\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}M(x_k - x^*) + \frac{32L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2$,

(3) $\mathbb{E}[T_{33}] \leq \frac{32LA^2u_{cm}^2\epsilon_k}{\theta\ell_{cs}^2}\mathbb{E}[M(x_k - x^*)] + \frac{16L\epsilon_k}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2$.

Now that Lemma A.3 provides upper bounds on the terms T_{31}, T_{32}, and T_{33}, using them in Eq. (13) and we have the following result.

Lemma A.4. *The following inequality holds for all $k \geq t_k$:*

$$\mathbb{E}[T_3] \leq \frac{112LA^2u_{cm}^2\epsilon_k\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}\mathbb{E}[M(x_k - x^*)] + \frac{56L\epsilon_k\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2.$$

Lastly, we bound the error term T_4 (i.e., the discretization error) in the following lemma, whose proof is provided in Appendix A.4.4.

Lemma A.5. *It holds for any $k \geq 0$ that* $T_4 \leq \frac{2LA^2u_{cm}^2\epsilon_k^2}{\theta\ell_{cs}^2}M(x_k - x^*) + \frac{L\epsilon_k^2}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2$.

Now we have control on all the error terms T_1 to T_4. Using them in Eq. (12), we obtain the following result. The proof is presented in Appendix A.4.5

Lemma A.6. *The following inequality holds for all $k \geq t_k$:*

$$\begin{aligned}\mathbb{E}[M(x_{k+1} - x^*)] \leq\; & \left(1 - 2\left(1 - \gamma\frac{u_{cm}}{\ell_{cm}}\right)\epsilon_k + \frac{114LA^2u_{cm}^2\epsilon_k\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}\right)\mathbb{E}[M(x_k - x^*)] \\ & + \frac{57L\epsilon_k\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2.\end{aligned}$$

Note that Lemma A.6 provides the desired one-step contractive inequality. We next repeatedly use Lemma A.6 to derive finite-sample convergence bounds of Algorithm (6). Recall our notations $\alpha_1 = \frac{u_{cm}^2}{\ell_{cm}^2} = \frac{1+\theta u_{cs}^2}{1+\theta\ell_{cs}^2}$, $\alpha_2 = 1 - \gamma\alpha_1^{1/2}$, and $\alpha_3 = \frac{114Lu_{cm}^2}{\theta\ell_{cs}^2}$. Then Lemma A.6 reads:

$$\mathbb{E}[M(x_{k+1} - x^*)] \leq \left(1 - 2\alpha_2\epsilon_k + \alpha_3A^2\epsilon_k\epsilon_{k-t_k,k-1}\right)\mathbb{E}[M(x_k - x^*)] + \frac{\alpha_3\epsilon_k\epsilon_{k-t_k,k-1}}{2u_{cm}^2}(A\|x^*\|_c + B)^2.$$

Since $\epsilon_{k-t_k,k-1} \leq \frac{\alpha_2}{\alpha_3A^2}$ for all $k \geq K$ (see Condition A.1 and the definition of K in Theorem A.1), we have by Lemma A.6 that

$$\mathbb{E}[M(x_{k+1} - x^*)] \leq (1 - \alpha_2\epsilon_k)\,\mathbb{E}[M(x_k - x^*)] + \frac{\alpha_3\epsilon_k\epsilon_{k-t_k,k-1}}{2u_{cm}^2}(A\|x^*\|_c + B)^2$$

for all $k \geq K$. Recursively using the previous inequality starting from K, we have for any $k \geq K$:

$$\begin{aligned}\mathbb{E}[\|x_k - x^*\|_c^2] &\leq 2u_{cm}^2\mathbb{E}[M(x_k - x^*)] && \text{(Proposition A.1)} \\ &\leq 2u_{cm}^2\mathbb{E}[M(x_K - x^*)]\prod_{j=K}^{k-1}(1 - \alpha_2\epsilon_j) + \alpha_3(A\|x^*\|_c + B)^2\sum_{i=K}^{k-1}\epsilon_i\epsilon_{i-t_i,i-1}\prod_{j=i+1}^{k-1}(1 - \alpha_2\epsilon_j) \\ &\leq \frac{u_{cm}^2}{\ell_{cm}^2}\mathbb{E}[\|x_K - x^*\|_c^2]\prod_{j=K}^{k-1}(1 - \alpha_2\epsilon_j) + \alpha_3(A\|x^*\|_c + B)^2\sum_{i=K}^{k-1}\epsilon_i\epsilon_{i-t_i,i-1}\prod_{j=i+1}^{k-1}(1 - \alpha_2\epsilon_j) \\ & && \text{(Proposition A.1)}\end{aligned}$$

$$\leq \alpha_1 \mathbb{E}[\|x_K - x^*\|_c^2] \prod_{j=K}^{k-1}(1-\alpha_2\epsilon_j) + \alpha_3 c_2 \sum_{i=K}^{k-1} \epsilon_i \epsilon_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\alpha_2\epsilon_j).$$

According to Condition A.1, we have $\epsilon_{0,K-1} \leq \frac{1}{4A}$. Using Corollary A.1.1 one more time and we have

$$\mathbb{E}[\|x_K - x^*\|_c^2] \leq \mathbb{E}[(\|x_K - x_0\|_c + \|x_0 - x^*\|_c)^2] \leq \left(\|x_0 - x^*\|_c + \|x_0\|_c + \frac{B}{A}\right)^2 = c_1.$$

Therefore, we finally obtain for all $k \geq K$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \prod_{j=K}^{k-1}(1-\alpha_2\epsilon_j) + \alpha_3 c_2 \sum_{i=K}^{k-1} \epsilon_i \epsilon_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\alpha_2\epsilon_j). \tag{14}$$

This proves Theorem A.1.

A.2 Finite-Sample Convergence Bounds for Using Various Stepsizes

We next proceed to prove Theorem 2.1 by evaluating the convergence bounds in Theorem A.1 when the stepsize sequence is chosen by $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$, where $\epsilon, h > 0$ and $\xi \in (0,1)$. We begin by restating Theorem 2.1 in full details.

Theorem A.2. *Consider $\{x_k\}$ of Algorithm (6). Suppose that Assumptions 2.1 – 2.4 are satisfied. Let $K = \min\{k : k \geq t_k\}$. Then we have the following results.*

(1) When $k \in [0, K-1]$, we have $\|x_k - x^*\|_c^2 \leq c_1$ almost surely.

(2) When $k \geq K$, we have the following finite-sample convergence bounds.

*(a) **Constant Stepsize.** Let $\bar{\epsilon} \in (0,1)$ be chosen such that $\epsilon t_\epsilon \leq \min(\frac{\alpha_2}{\alpha_3 A^2}, \frac{1}{4A})$ for all $\epsilon \in (0, \bar{\epsilon})$. Then when $\epsilon_k \equiv \epsilon$, we have for all $k \geq t_\epsilon$:*

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 (1-\alpha_2\epsilon)^{k-t_\epsilon} + \frac{\alpha_3 c_2}{\alpha_2}\epsilon t_\epsilon.$$

*(b) **Linear Stepsize.** When $\epsilon_k = \frac{\epsilon}{k+h}$, for any $\epsilon > 0$, let $\bar{h}$ be chosen such that $\epsilon_{0,K-1} \leq \min(\frac{\alpha_2}{\alpha_3 A^2}, \frac{1}{4A})$ for all $h \geq \bar{h}$. Then*

(i) When $\epsilon < 1/\alpha_2$, we have for all $k \geq K$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2\epsilon} + \frac{8\epsilon^2\alpha_3 c_2}{1-\alpha_2\epsilon}\frac{t_k}{(k+h)^{\alpha_2\epsilon}}.$$

(ii) When $\epsilon = 1/\alpha_2$, we have for all $k \geq K$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \frac{K+h}{k+h} + 8\epsilon^2\alpha_3 c_2 \frac{t_k \log(k+h)}{k+h}.$$

(iii) When $\epsilon > 1/\alpha_2$, we have for all $k \geq K$:

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2\epsilon} + \frac{8e\epsilon^2\alpha_3 c_2}{\alpha_2\epsilon - 1}\frac{t_k}{k+h}.$$

*(c) **Polynomial Stepsize.** When $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$, for any $\xi \in (0,1)$ and $\epsilon > 0$, let $\bar{h}$ be chosen such that $\bar{h} \geq [2\xi/(\alpha_2\epsilon)]^{1/(1-\xi)}$ and $\epsilon_{0,K-1} \leq \min(\frac{\alpha_2}{\alpha_3 A^2}, \frac{1}{4A})$ for any $h \geq \bar{h}$. Then we have for all $k \geq K$:*

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 e^{-\frac{\alpha_2\epsilon}{1-\xi}\left((k+h)^{1-\xi} - (K+h)^{1-\xi}\right)} + \frac{4\alpha_3 c_2 \epsilon}{\alpha_2}\frac{t_k}{(k+h)^\xi}.$$

A.3 Proof of Theorem A.2

Theorem A.2 (1) directly follows from Theorem A.1 (1). We next prove Theorem A.2 (2) (a).

A.3.1 Constant Stepsize

It is clear that Condition A.1 is satisfied when $\epsilon t_\epsilon \leq \min(\frac{\alpha_2}{\alpha_3 A^2}, \frac{1}{4A})$. We next verify the existence of such threshold $\bar{\epsilon}$. Note that we have by definition of t_ϵ and Assumption 2.3 that

$$t_\epsilon \leq \min\left\{k \geq 0 \;:\; C\sigma^k \leq \epsilon\right\} = \min\left\{k \geq 0 \;:\; k \geq \frac{\log(1/\epsilon) + \log(C)}{\log(1/\sigma)}\right\} \leq \frac{\log(1/\epsilon) + \log(C/\sigma)}{\log(1/\sigma)}.$$

It follows that $\lim_{\epsilon \to 0} \epsilon t_\epsilon = 0$. Hence there exists $\bar{\epsilon} \in (0,1)$ such that Condition A.1 is satisfied for all $\epsilon \in (0, \bar{\epsilon})$, which is stated in Condition 2.1 (1). We next evaluate Eq. (14). When $\epsilon_k \equiv \epsilon$, we have for all $k \geq t_\epsilon$:

$$\begin{aligned}
\mathbb{E}[\|x_k - x^*\|_c^2] &\leq \alpha_1 c_1 \prod_{j=K}^{k-1}(1 - \alpha_2 \epsilon_j) + \alpha_3 c_2 \sum_{i=K}^{k-1} \epsilon_i \epsilon_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1 - \alpha_2 \epsilon_j) \\
&= \alpha_1 c_1 (1 - \alpha_2 \epsilon)^{k-K} + \alpha_3 c_2 \sum_{i=K}^{k-1} \epsilon^2 t_\epsilon (1 - \alpha_2 \epsilon)^{k-i-1} \\
&\leq \alpha_1 c_1 (1 - \alpha_2 \epsilon)^{k-K} + \frac{\alpha_3 c_2}{\alpha_2} \epsilon t_\epsilon.
\end{aligned}$$

This proves Part (2) (a) of Theorem 2.1.

A.3.2 Linear Stepsize

We first verify the existence of the threshold $\bar{h}$. We begin by comparing ϵ_{k-t_k} with ϵ_k. Using Assumption 2.3 and we have

$$t_k \leq \frac{\log(k+h) + \log(C/(\sigma\epsilon))}{\log(1/\sigma)}.$$

It follows that

$$\frac{\epsilon_k}{\epsilon_{k-t_k}} = 1 - \frac{t_k}{k+h} \longrightarrow 1 \text{ as } (k+h) \to \infty.$$

Therefore, there exists $\bar{h}_1 > 0$ such that $\epsilon_{k-t_k} \leq 2\epsilon_k$ holds for any $k \geq t_k$ when $h \geq \bar{h}_1$. Now consider the requirement stated in Condition A.1. Using the fact that $\{\epsilon_k\}$ is non-increasing, we have

$$\epsilon_{k-t_k,k-1} \leq t_k \epsilon_{k-t_k} \leq 2\epsilon_k t_k \to 0 \text{ as } (k+h) \to \infty.$$

Hence there exists $\bar{h}_2 > 0$ such that $\epsilon_{k-t_k,k-1} \leq \min(\frac{\alpha_2}{\alpha_3 A^2}, \frac{1}{4A})$ holds for any $k \geq t_k$ when $h \geq \bar{h}_2$. Now choosing $\bar{h} = \max(\bar{h}_1, \bar{h}_2)$, Condition A.1 is satisfied. This is stated in Condition 2.1 (2). Furthermore, by construction we have $\epsilon_{k-t_k} \leq 2\epsilon_k$ for any $k \geq t_k$, which will be useful later when we evaluate Eq. (14).

We next evaluate the RHS of Eq. (14) in the following lemma, whose proof is presented in Appendix A.4.6.

Lemma A.7. *The following inequality hold for all $k \geq K$:*

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \begin{cases} \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2 \epsilon} + \frac{8\alpha_3 c_2 \epsilon^2}{1-\alpha_2\epsilon} \frac{t_k}{(k+h)^{\alpha_2\epsilon}}, & \epsilon < \frac{1}{\alpha_2}, \\ \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2 \epsilon} + 8\alpha_3 c_2 \epsilon^2 \frac{t_k \log(k+h)}{k+h}, & \epsilon = \frac{1}{\alpha_2}, \\ \alpha_1 c_1 \left(\frac{K+h}{k+h}\right)^{\alpha_2 \epsilon} + \frac{8e\alpha_3 c_2 \epsilon^2}{\alpha_2\epsilon - 1} \frac{t_k}{k+h}, & \epsilon > \frac{1}{\alpha_2}. \end{cases}$$

This proves Part (2) (b) of Theorem 2.1.

A.3.3 Polynomially Diminishing Stepsizes

Now we consider using $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$, where $\xi \in (0,1)$ and $\epsilon, h > 0$. Similarly as in the previous section, we can show that for any $\xi \in (0,1)$ and $\epsilon > 0$, there exists $\bar{h} > 0$ such that Condition A.1 is satisfied for all $h \geq \bar{h}$. Furthermore, we can assume without loss of generality that $\epsilon_{k-t_k} \leq 2\epsilon_k$ for all $k \geq t_k$ and $\bar{h} \geq [2\xi/(\alpha_2\epsilon)]^{1/(1-\xi)}$.

We next evaluate the RHS of Eq. (14) in the following lemma, whose proof is presented in Appendix A.4.7.

Lemma A.8. *The following inequality hold for all $k \geq K$:*

$$\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \exp\left[-\frac{\alpha_2\epsilon}{1-\xi}\left((k+h)^{1-\xi} - (K+h)^{1-\xi}\right)\right] + \frac{4\alpha_3 c_2 \epsilon}{\alpha_2} \frac{t_k}{(k+h)^\xi}.$$

This proves Part (2) (c) of Theorem 2.1.

A.4 Proof of Technical Lemmas

A.4.1 Proof of Lemma 2.1

(1) When $\|\cdot\|_c = \|\cdot\|_2$, we choose $\theta = 1$ and $g(x) = \frac{1}{2}\|x\|_2^2$. It follows that $L = 1$ and $u_{cs} = \ell_{cs} = 1$. Therefore, we have $\alpha_1 = 1$, $\alpha_2 = 1-\gamma$, and $\alpha_3 = 228$.

(2) Recall the definition of $\{\alpha_i\}_{1\leq i\leq 3}$ in the beginning of Appendix A: $\alpha_1 = \frac{1+\theta u_{cs}^2}{1+\theta \ell_{cs}^2}$, $\alpha_2 = 1 - \gamma\alpha_1^{1/2}$, and $\alpha_3 = \frac{114L(1+\theta u_{cs}^2)}{\theta \ell_{cs}^2}$. When $\|\cdot\|_c = \|\cdot\|_\infty$, we choose $\theta = \left(\frac{1+\gamma}{2\gamma}\right)^2 - 1$ and $g(x) = \frac{1}{2}\|x\|_p^2$ with $p = 2\log(d)$, where d is the dimension of the iterates x_k. It follows that $L = p - 1 \leq 2\log(d)$ [3], $u_{cs} = 1$, and $\ell_{cs} = 1/d^{1/p} = 1/\sqrt{e}$. Therefore, we have

$$\begin{aligned} \alpha_1 &= \frac{1+\theta u_{cs}^2}{1+\theta\ell_{cs}^2} = \frac{1+\theta}{1+\theta/\sqrt{e}} \leq \sqrt{e} \leq 3, \\ \alpha_2 &= 1 - \gamma\alpha_1^{1/2} \geq 1 - \gamma\frac{1+\gamma}{2\gamma} = \frac{1-\gamma}{2}, \\ \alpha_3 &= \frac{114L(1+\theta u_{cs}^2)}{\theta\ell_{cs}^2} \leq \frac{228e\log(d)(1+\theta)}{\theta} \leq \frac{456e\log(d)}{1-\gamma}. \end{aligned}$$

A.4.2 Proof of Lemma A.2

We first show that under Assumption 2.1, the size of $\|F(x,y)\|_c$ and $\|\bar{F}(x)\|_c$ can grow at most affinely in terms of $\|x\|_c$. Using Triangle inequality, we have for any $x \in \mathbb{R}^d$ and $y \in \mathcal{Y}$

$$\|F(x,y)\|_c - \|F(0,y)\|_c \leq \|F(x,y) - F(0,y)\|_c \leq A_1\|x\|_c,$$

where the last inequality follows from Assumption 2.1. It follows that

$$\|F(x,y)\|_c \leq A_1\|x\|_c + \|F(0,y)\|_c \leq A_1\|x\|_c + B_1.$$

Furthermore, we have by Jensen's inequality and the convexity of norms that

$$\|\bar{F}(x)\|_c = \|\mathbb{E}_{Y\sim\mu}[F(x,Y)]\|_c \leq \mathbb{E}_{Y\sim\mu}[\|F(x,Y)\|_c] \leq A_1\|x\|_c + B_1.$$

The previous two inequalities will be frequently used in the derivation here after. Now we proceed to prove Lemma A.2.

For any $k \in [k_1, k_2 - 1]$, using Triangle inequality, we have

$$\begin{aligned}
\|x_{k+1}\|_c - \|x_k\|_c &\leq \|x_{k+1} - x_k\|_c \\
&= \epsilon_k\|F(x_k, Y_k) - x_k + w_k\|_c \\
&\leq \epsilon_k(\|F(x_k, Y_k)\|_c + \|x_k\|_c + \|w_k\|_c) \\
&\leq \epsilon_k(A_1\|x_k\|_c + B_1 + \|x_k\|_c + A_2\|x_k\|_c + B_2). && \text{(Assumptions 2.1, 2.4)} \\
&\leq \epsilon_k(A\|x_k\|_c + B). && (15)
\end{aligned}$$

Note that the previous inequality is equivalent to

$$\|x_{k+1}\|_c + \frac{B}{A} \leq (1 + A\epsilon_k)\left(\|x_k\|_c + \frac{B}{A}\right),$$

which implies for all $k \in [k_1, k_2]$:

$$\|x_k\|_c \leq \prod_{j=k_1}^{k-1}(1 + A\epsilon_j)\left(\|x_{k_1}\|_c + \frac{B}{A}\right) - \frac{B}{A}.$$

Using the fact that $1 + x \leq e^x \leq 1 + 2x$ for all $x \in [0, 1/2]$, we have when $\epsilon_{k_1,k_2-1} \leq \frac{1}{4A}$:

$$\prod_{j=k_1}^{k-1}(1 + A\epsilon_j) \leq \exp\left(A\epsilon_{k_1,k-1}\right) \leq 1 + 2A\epsilon_{k_1,k-1}.$$

It follows that for all $k \in [k_1, k_2]$:

$$\|x_k\|_c \leq (1 + 2A\epsilon_{k_1,k-1})\|x_{k_1}\|_c + 2B\epsilon_{k_1,k-1}.$$

Using the previous inequality in Eq. (15) and we have for any $k \in [k_1, k_2]$:

$$\begin{aligned}
\|x_k - x_{k_1}\|_c &\leq \sum_{j=k_1}^{k-1}\|x_{j+1} - x_j\|_c \\
&\leq \sum_{j=k_1}^{k-1}\epsilon_j(A\|x_j\|_c + B)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{j=k_1}^{k-1} \epsilon_j [A((1+2A\epsilon_{k_1,j-1})\|x_{k_1}\|_c + 2B\epsilon_{k_1,j-1}) + B] \\
&\leq 2\sum_{j=k_1}^{k-1} \epsilon_j (A\|x_{k_1}\|_c + B) && (\epsilon_{k_1,j-1} \leq \tfrac{1}{4A}) \\
&= 2\epsilon_{k_1,k-1}(A\|x_{k_1}\|_c + B).
\end{aligned}
$$

Since $\epsilon_{k_1,k-1} \leq \epsilon_{k_1,k_2-1}$ when $k \in [k_1, k_2]$, we obtain the first claimed inequality:

$$
\|x_k - x_{k_1}\|_c \leq 2\epsilon_{k_1,k_2-1}(A\|x_{k_1}\|_c + B), \quad \forall\, k \in [k_1, k_2].
$$

Now for the second claimed inequality, since

$$
\begin{aligned}
\|x_{k_2} - x_{k_1}\|_c &\leq 2\epsilon_{k_1,k_2-1}(A\|x_{k_1}\|_c + B) \\
&\leq 2\epsilon_{k_1,k_2-1}(A\|x_{k_1} - x_{k_2}\|_c + A\|x_{k_2}\|_c + B) \\
&\leq \frac{1}{2}\|x_{k_2} - x_{k_1}\|_c + 2\epsilon_{k_1,k_2-1}(A\|x_{k_2}\|_c + B),
\end{aligned}
$$

we have $\|x_{k_2} - x_{k_1}\|_c \leq 4\epsilon_{k_1,k_2-1}(A\|x_{k_2}\|_c + B)$. Therefore, we have for any $k \in [k_1, k_2]$:

$$
\begin{aligned}
\|x_k - x_{k_1}\|_c &\leq 2\epsilon_{k_1,k_2-1}(A\|x_{k_1}\|_c + B) \\
&\leq 2\epsilon_{k_1,k_2-1}(A\|x_{k_1} - x_{k_2}\|_c + A\|x_{k_2}\|_c + B) \\
&\leq 2\epsilon_{k_1,k_2-1}(4A\epsilon_{k_1,k_2-1}(A\|x_{k_2}\|_c + B) + A\|x_{k_2}\|_c + B) \\
&\leq 4\epsilon_{k_1,k_2-1}(A\|x_{k_2}\|_c + B), && (\epsilon_{k_1,k_2-1} \leq \tfrac{1}{4A})
\end{aligned}
$$

which is the second claimed inequality.

A.4.3 Proof of Lemma A.3

(1) For the term T_{31}, using Hölder's inequality and we have

$$
\begin{aligned}
T_{31} &= \langle \nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*), F(x_k, Y_k) - \bar{F}(x_k) \rangle \\
&\leq \|\nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*)\|_s^* \|F(x_k, Y_k) - \bar{F}(x_k)\|_s \\
&\leq \frac{1}{\ell_{cs}} \|\nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*)\|_s^* \|F(x_k, Y_k) - \bar{F}(x_k)\|_c, && (\ell_{cs}\|\cdot\|_s \leq \|\cdot\|_c)
\end{aligned}
$$

where $\|\cdot\|_s^*$ denotes the dual norm of $\|\cdot\|_s$. We first control the term $\|\nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*)\|_s^*$. Recall that an equivalent definition of a convex function $h(x)$ been L – smooth with respect to norm $\|\cdot\|$ is that

$$
\|\nabla h(x_1) - \nabla h(x_2)\|_* \leq L\|x_1 - x_2\|, \quad \forall\, x_1, x_2,
$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ [3]. Therefore, since $M(x)$ is $\frac{L}{\theta}$-smooth w.r.t. $\|\cdot\|_s$, we have

$$
\begin{aligned}
\|\nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*)\|_s^* &\leq \frac{L}{\theta}\|x_k - x_{k-t_k}\|_s \\
&\leq \frac{L}{\theta\ell_{cs}}\|x_k - x_{k-t_k}\|_c \\
&\leq \frac{4L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}}(A\|x_k - x^*\|_c + A\|x^*\|_c + B), \qquad (16)
\end{aligned}
$$

where the last line follows from Lemma A.2.

We next control the term $\|F(x_k, Y_k) - \bar{F}(x_k)\|_c$. Using Assumption 2.1, Assumption 2.2, and the fact that $\bar{F}(x^*) = x^*$, we have

$$\begin{aligned}
\|F(x_k, Y_k) - \bar{F}(x_k)\|_c &= \|F(x_k, Y_k) - \bar{F}(x_k) + \bar{F}(x^*) - x^*\|_c \\
&\leq \|F(x_k, Y_k)\|_c + \|\bar{F}(x_k) - \bar{F}(x^*)\|_c + \|x^*\|_c \\
&\leq A_1\|x_k\|_c + B_1 + \|x_k - x^*\|_c + \|x^*\|_c \\
&\leq (A_1+1)\|x_k - x^*\|_c + (A_1+1)\|x^*\|_c + B_1 \\
&\leq A\|x_k - x^*\|_c + A\|x^*\|_c + B.
\end{aligned}$$

It follows that

$$\begin{aligned}
T_{31} &\leq \frac{1}{\ell_{cs}}\|\nabla M(x_k - x^*) - \nabla M(x_{k-t_k} - x^*)\|_s^*\|F(x_k, Y_k) - \bar{F}(x_k)\|_c \\
&\leq \frac{4L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x_k - x^*\|_c + A\|x^*\|_c + B)^2 \\
&\leq \frac{8L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}A^2\|x_k - x^*\|_c^2 + \frac{8L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2 \\
&\leq \frac{16LA^2u_{cm}^2\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}M(x_k - x^*) + \frac{8L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c + B)^2.
\end{aligned}$$

(2) Consider the term T_{32}. Using Hölder's inequality and we have

$$\begin{aligned}
T_{32} &= \langle \nabla M(x_{k-t_k} - x^*), F(x_k, Y_k) - F(x_{k-t_k}, Y_k) + \bar{F}(x_{k-t_k}) - \bar{F}(x_k)\rangle \\
&\leq \|\nabla M(x_{k-t_k} - x^*)\|_s^*\|F(x_k, Y_k) - F(x_{k-t_k}, Y_k) + \bar{F}(x_{k-t_k}) - \bar{F}(x_k)\|_s \qquad \text{(Hölder's inequality)} \\
&\leq \frac{1}{\ell_{cs}}\|\nabla M(x_{k-t_k} - x^*)\|_s^*\|F(x_k, Y_k) - F(x_{k-t_k}, Y_k) + \bar{F}(x_{k-t_k}) - \bar{F}(x_k)\|_c.
\end{aligned}$$

For the term $\|\nabla M(x_{k-t_k} - x^*)\|_s^*$, we have

$$\begin{aligned}
\|\nabla M(x_{k-t_k} - x^*)\|_s^* &= \|\nabla M(x_{k-t_k} - x^*) - \nabla M(x^* - x^*)\|_s^* \\
&\leq \frac{L}{\theta}\|x_{k-t_k} - x^*\|_s \qquad \text{(equivalent definition of smoothness)} \\
&\leq \frac{L}{\theta\ell_{cs}}\|x_{k-t_k} - x^*\|_c \\
&\leq \frac{L}{\theta\ell_{cs}}(\|x_{k-t_k} - x_k\|_c + \|x_k - x^*\|_c) \\
&\leq \frac{2L}{\theta\ell_{cs}}\left(\|x_k - x^*\|_c + \|x^*\|_c + \frac{B}{A}\right),
\end{aligned} \tag{17}$$

where the last line follow from Corollary A.1.1. For the term $\|F(x_k, Y_k) - F(x_{k-t_k}, Y_k) + \bar{F}(x_{k-t_k}) - \bar{F}(x_k)\|_c$, using Assumptions 2.1 and 2.2 and we obtain

$$\begin{aligned}
&\|F(x_k, Y_k) - F(x_{k-t_k}, Y_k) + \bar{F}(x_{k-t_k}) - \bar{F}(x_k)\|_c \\
\leq\ &\|F(x_k, Y_k) - F(x_{k-t_k}, Y_k)\|_c + \|\bar{F}(x_{k-t_k}) - \bar{F}(x_k)\|_c \\
\leq\ &2A_1\|x_k - x_{k-t_k}\|_c \\
\leq\ &2A\|x_k - x_{k-t_k}\|_c \\
\leq\ &8A\epsilon_{k-t_k,k-1}(A\|x_k - x^*\|_c + A\|x^*\|_c + B),
\end{aligned}$$

where in the last line we used Lemma A.2. It follows that

$$\begin{aligned}
T_{32} &\leq \frac{1}{\ell_{cs}}\|\nabla M(x_{k-t_k}-x^*)\|_s^*\|F(x_k,Y_k)-F(x_{k-t_k},Y_k)+\bar{F}(x_{k-t_k})-\bar{F}(x_k)\|_c \\
&\leq \frac{16L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x_k-x^*\|_c+A\|x^*\|_c+B)^2 \\
&\leq \frac{32LA^2\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}\|x_k-x^*\|_c^2+\frac{32L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2 \\
&\leq \frac{64LA^2u_{cm}^2\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}M(x_k-x^*)+\frac{32L\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2.
\end{aligned}$$

(3) Consider the term T_{33}. We first take expectation conditioning on x_{k-t_k} and Y_{k-t_k} to obtain

$$\begin{aligned}
&\mathbb{E}[T_{33}\mid x_{k-t_k},Y_{k-t_k}] \\
={}&\langle\nabla M(x_{k-t_k}-x^*),\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\bar{F}(x_{k-t_k})\rangle \\
\leq{}&\|\nabla M(x_{k-t_k}-x^*)\|_s^*\|\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\bar{F}(x_{k-t_k})\|_s \\
\leq{}&\frac{1}{\ell_{cs}}\|\nabla M(x_{k-t_k}-x^*)\|_s^*\|\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\bar{F}(x_{k-t_k})\|_c.
\end{aligned}$$

For the term $\|\nabla M(x_{k-t_k}-x^*)\|_s^*$, we have from Eq. (17) that

$$\|\nabla M(x_{k-t_k}-x^*)\|_s^*\leq\frac{2L}{\theta\ell_{cs}}\left(\|x_k-x^*\|_c+\|x^*\|_c+\frac{B}{A}\right).$$

For the term $\|\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\bar{F}(x_{k-t_k})\|_c$, using the geometric mixing of the Markov chain $\{Y_k\}$ (Assumption 2.3), we have

$$\begin{aligned}
&\|\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\bar{F}(x_{k-t_k})\|_c \\
={}&\|\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\mathbb{E}_{Y\sim\mu}[F(x_{k-t_k},Y)]\|_c \\
={}&\left\|\sum_{y\in\mathcal{Y}}\left(P^{t_k}(Y_{k-t_k},y)-\mu(y)\right)F(x_{k-t_k},y)\right\|_c \\
\leq{}&\sum_{y\in\mathcal{Y}}\left|P^{t_k}(Y_{k-t_k},y)-\mu(y)\right|\|F(x_{k-t_k},y)\|_c \\
\leq{}&2\max_{y_0\in\mathcal{Y}}\|P^{t_k}(y_0,\cdot)-\mu(\cdot)\|_{\mathrm{TV}}(A_1\|x_{k-t_k}\|_c+B_1) \\
\leq{}&2C\sigma^{t_k}(A_1\|x_k-x_{k-t_k}\|_c+A_1\|x_k\|_c+B_1) && \text{(Assumption 2.3)} \\
\leq{}&2\epsilon_k(A_1(\|x_k\|_c+B/A)+A_1\|x_k\|_c+B_1) && \text{(Definition of } t_k \text{ and Corollary A.1.1)} \\
\leq{}&4\epsilon_k(A\|x_k-x^*\|_c+A\|x^*\|_c+B).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}[T_{33}\mid x_{k-t_k},Y_{k-t_k}] &\leq\frac{1}{\ell_{cs}}\|\nabla M(x_{k-t_k}-x^*)\|_s^*\|\mathbb{E}[F(x_{k-t_k},Y_k)\mid x_{k-t_k},Y_{k-t_k}]-\bar{F}(x_{k-t_k})\|_c \\
&\leq\frac{8L\epsilon_k}{\theta\ell_{cs}^2}(A\|x_k-x^*\|_c+A\|x^*\|_c+B)^2 \\
&\leq\frac{16L\epsilon_k}{\theta\ell_{cs}^2}A^2\|x_k-x^*\|_c^2+\frac{16L\epsilon_k}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2 \\
&\leq\frac{32LA^2u_{cm}^2\epsilon_k}{\theta\ell_{cs}^2}M(x_k-x^*)+\frac{16L\epsilon_k}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2.
\end{aligned}$$

Taking the total expectation on both sides of the previous inequality yields the desired result.

A.4.4 Proof of Lemma A.5

Using Proposition A.1 (2), Assumption 2.1, and Assumption 2.4 (2), we have

$$
\begin{aligned}
T_4 &= \frac{L\epsilon_k^2}{2\theta}\|F(x_k,Y_k)-x_k+w_k\|_s^2 \\
&\leq \frac{L\epsilon_k^2}{2\theta\ell_{cs}^2}\|F(x_k,Y_k)-x_k+w_k\|_c^2 && \text{(Proposition A.1 (3))} \\
&\leq \frac{L\epsilon_k^2}{2\theta\ell_{cs}^2}(\|F(x_k,Y_k)\|_c+\|x_k\|_c+\|w_k\|_c)^2 \\
&\leq \frac{L\epsilon_k^2}{2\theta\ell_{cs}^2}(A\|x_k\|_c+B)^2 && \text{(Assumptions 2.1 and 2.4)} \\
&\leq \frac{L\epsilon_k^2}{2\theta\ell_{cs}^2}(A\|x_k-x^*\|_c+A\|x^*\|_c+B)^2 \\
&\leq \frac{L\epsilon_k^2}{\theta\ell_{cs}^2}A^2\|x_k-x^*\|_c^2+\frac{L\epsilon_k^2}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2 \\
&\leq \frac{2LA^2u_{cm}^2\epsilon_k^2}{\theta\ell_{cs}^2}M(x_k-x^*)+\frac{L\epsilon_k^2}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2.
\end{aligned}
$$

A.4.5 Proof of Lemma A.6

Using Lemmas A.1, A.4, and A.5 in Eq. (12) and we have for all $k\geq t_k$

$$
\begin{aligned}
\mathbb{E}[M(x_{k+1}-x^*)] \leq{}& \left(1-2\alpha_2\epsilon_k+\frac{114LA^2u_{cm}^2\epsilon_k\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}\right)\mathbb{E}[M(x_k-x^*)] \\
&+\frac{57L\epsilon_k\epsilon_{k-t_k,k-1}}{\theta\ell_{cs}^2}(A\|x^*\|_c+B)^2 \\
={}& \left(1-2\alpha_2\epsilon_k+\alpha_3A^2\epsilon_k\epsilon_{k-t_k,k-1}\right)\mathbb{E}[M(x_k-x^*)]+\frac{\alpha_3c_2\epsilon_k\epsilon_{k-t_k,k-1}}{2u_{cm}^2},
\end{aligned}
$$

where $\alpha_3=\frac{114Lu_{cm}^2}{\theta\ell_{cs}^2}$.

A.4.6 Proof of Lemma A.7

We first simplify the RHS of Eq. (14) using $\epsilon_k=\frac{\epsilon}{k+h}$. Since we have chosen h such that $\epsilon_{k-t_k,k-1}\leq 2\epsilon_k$ for any $k\geq t_k$, Eq. (14) implies

$$
\begin{aligned}
\mathbb{E}[\|x_k-x^*\|_c^2] &\leq \alpha_1c_1\prod_{j=K}^{k-1}(1-\alpha_2\epsilon_j)+\alpha_3c_2\sum_{i=K}^{k-1}\epsilon_i\epsilon_{i-t_i,i-1}\prod_{j=i+1}^{k-1}(1-\alpha_2\epsilon_j) \\
&\leq \alpha_1c_1\prod_{j=K}^{k-1}(1-\alpha_2\epsilon_j)+2\alpha_3c_2\sum_{i=K}^{k-1}\epsilon_i^2t_i\prod_{j=i+1}^{k-1}(1-\alpha_2\epsilon_j) \\
&= \alpha_1c_1\underbrace{\prod_{j=K}^{k-1}\left(1-\frac{\alpha_2\epsilon}{j+h}\right)}_{E_1}+2\alpha_3c_2t_k\underbrace{\sum_{i=K}^{k-1}\frac{\epsilon^2}{(i+h)^2}\prod_{j=i+1}^{k-1}\left(1-\frac{\alpha_2\epsilon}{j+h}\right)}_{E_2} \qquad (18)
\end{aligned}
$$

For the term E_1, we have

$$
E_1\leq\exp\left(-\alpha_2\epsilon\sum_{j=K}^{k-1}\frac{1}{j+h}\right)\leq\exp\left(-\alpha_2\epsilon\int_K^k\frac{1}{x+h}dx\right)=\left(\frac{K+h}{k+h}\right)^{\alpha_2\epsilon}.
$$

Now consider the term E_2. Similarly we have

$$
\begin{aligned}
E_2 &= \sum_{i=K}^{k-1} \frac{\epsilon^2}{(i+h)^2} \prod_{j=i+1}^{k-1} \left(1 - \frac{\alpha_2 \epsilon}{j+h}\right) \\
&\leq \sum_{i=K}^{k-1} \frac{\epsilon^2}{(i+h)^2} \left(\frac{i+1+h}{k+h}\right)^{\alpha_2\epsilon} \\
&\leq \frac{4\epsilon^2}{(k+h)^{\alpha_2\epsilon}} \sum_{i=K}^{k-1} \frac{1}{(i+1+h)^{2-\alpha_2\epsilon}} \\
&\leq \begin{cases} \dfrac{4\epsilon^2}{1-\alpha_2\epsilon} \dfrac{1}{(k+h)^{\alpha_2\epsilon}}, & \alpha_2\epsilon \in (0,1), \\ \dfrac{4\epsilon^2 \log(k+h)}{k+h}, & \alpha_2\epsilon = 1, \\ \dfrac{4e\epsilon^2}{\alpha_2\epsilon - 1} \dfrac{1}{k+h}, & \alpha_2\epsilon \in (1,\infty). \end{cases}
\end{aligned}
$$

The result then follows from using the upper bounds we obtained for the terms E_1 and E_2 in inequality (18).

A.4.7 Proof of Lemma A.8

When $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$, similarly we have from Eq. (14) that

$$
\mathbb{E}[\|x_k - x^*\|_c^2] \leq \alpha_1 c_1 \underbrace{\prod_{j=K}^{k-1} \left(1 - \frac{\alpha_2\epsilon}{(j+h)^\xi}\right)}_{E_1} + 2\alpha_3 c_2 t_k \underbrace{\sum_{i=K}^{k-1} \frac{\epsilon^2}{(i+h)^{2\xi}} \prod_{j=i+1}^{k-1} \left(1 - \frac{\alpha_2\epsilon}{(j+h)^\xi}\right)}_{E_2} \tag{19}
$$

The term E_1 can be controlled in the following way:

$$
\begin{aligned}
E_1 &= \prod_{j=K}^{k-1} \left(1 - \frac{\alpha_2\epsilon}{(j+h)^\xi}\right) \\
&\leq \exp\left(-\alpha_2\epsilon \sum_{j=K}^{k-1} \frac{1}{(j+h)^\xi}\right) \\
&\leq \exp\left(-\alpha_2\epsilon \int_K^k \frac{1}{(x+h)^\xi} dx\right) \\
&= \exp\left[-\frac{\alpha_2\epsilon}{1-\xi}\left((k+h)^{1-\xi} - (K+h)^{1-\xi}\right)\right].
\end{aligned}
$$

As for the term E_2, we will show by induction that $E_2 \leq \frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^\xi}$ for all $k \geq 0$. Consider a sequence $\{u_k\}_{k\geq 0}$ (with $u_0 = 0$) defined by

$$
u_{k+1} = \left(1 - \alpha_2 \frac{\epsilon}{(k+h)^\xi}\right) u_k + \frac{\epsilon^2}{(k+h)^{2\xi}}, \quad \forall\, k \geq 0.
$$

It can be easily verified that $u_k = E_2$. Since $u_0 = 0 \leq \frac{2\epsilon}{\alpha_2}\frac{1}{h^\xi}$, we have the base case. Now suppose $u_k \leq \frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^\xi}$ for some $k > 0$. Consider u_{k+1}, and we have

$$
\frac{2\epsilon}{\alpha_2}\frac{1}{(k+1+h)^\xi} - u_{k+1} = \frac{2\epsilon}{\alpha_2}\frac{1}{(k+1+h)^\xi} - \left(1 - \alpha_2\frac{\epsilon}{(k+h)^\xi}\right) u_k + \frac{\epsilon^2}{(k+h)^{2\xi}}
$$

$$
\begin{aligned}
&\geq \frac{2\epsilon}{\alpha_2}\frac{1}{(k+1+h)^\xi} - \left(1-\frac{\alpha_2\epsilon}{(k+h)^\xi}\right)\frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^\xi} - \frac{\epsilon^2}{(k+h)^{2\xi}} \\
&= \frac{2\epsilon}{\alpha_2}\left[\frac{1}{(k+1+h)^\xi} - \frac{1}{(k+h)^\xi} + \frac{\alpha_2\epsilon}{2}\frac{1}{(k+h)^{2\xi}}\right] \\
&= \frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^{2\xi}}\left[\frac{\alpha_2\epsilon}{2} - (k+h)^\xi\left(1-\left(\frac{k+h}{k+1+h}\right)^\xi\right)\right].
\end{aligned}
$$

Note that

$$
\left(\frac{k+h}{k+1+h}\right)^\xi = \left[\left(1+\frac{1}{k+h}\right)^{k+h}\right]^{-\frac{\xi}{k+h}} \geq \exp\left(-\frac{\xi}{k+h}\right) \geq 1-\frac{\xi}{k+h},
$$

where we used $(1+\frac{1}{x})^x < e$ for all $x > 0$ and $e^x \geq 1+x$ for all $x \in \mathbb{R}$. Therefore, we obtain

$$
\begin{aligned}
\frac{2\epsilon}{\alpha_2}\frac{1}{(k+1+h)^\xi} - u_{k+1} &\geq \frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^{2\xi}}\left[\frac{\alpha_2\epsilon}{2} - (k+h)^\xi\left(1-\left(\frac{k+h}{k+1+h}\right)^\xi\right)\right] \\
&\geq \frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^{2\xi}}\left[\frac{\alpha_2\epsilon}{2} - \frac{\xi}{(k+h)^{1-\xi}}\right] \\
&\geq 0,
\end{aligned}
$$

where the last line follows from $h \geq \bar{h} \geq [2\xi/(\alpha_2\epsilon)]^{1/(1-\xi)}$. The induction is now complete, and we have $E_2 \leq \frac{2\epsilon}{\alpha_2}\frac{1}{(k+h)^\xi}$ for all $k \geq 0$.

Using the upper bounds we obtained for the terms E_1 and E_2 in inequality (19) and we have the desired result.

B Q-Learning

B.1 Proof of Proposition 3.1

(1) For any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $y \in \mathcal{Y}$, we have

$$
\begin{aligned}
&\|F(Q_1,y) - F(Q_2,y)\|_\infty \\
&= \max_{(s,a)}\left|\beta\mathbb{1}_{\{(s_0,a_0)=(s,a)\}}(\max_{a_1\in\mathcal{A}} Q_1(s_1,a_1) - \max_{a_2\in\mathcal{A}} Q_2(s_1,a_2)) + \mathbb{1}_{\{(s_0,a_0)\neq(s,a)\}}(Q_1(s_0,a_0) - Q_2(s_0,a_0))\right| \\
&\leq 2\|Q_1 - Q_2\|_\infty.
\end{aligned}
$$

Similarly, for any $y \in \mathcal{Y}$, we have

$$
\|F(0,y)\|_\infty = \max_{(s,a)}\left|\mathbb{1}_{\{(s_0,a_0)=(s,a)\}}\mathcal{R}(s_0,a_0)\right| \leq 1.
$$

(2) It is clear from Assumption 3.1 that $\{Y_k\}$ has a unique stationary distribution, denoted by μ. Moreover, we have $\mu(s,a,s') = \kappa_b(s)\pi_b(a|s)P_a(s,s')$ for any $(s,a,s') \in \mathcal{Y}$.

Now consider the second claim. Using the definition of total variation distance, we have for all $k \geq 0$:

$$
\max_{y\in\mathcal{Y}}\|P^{k+1}(y,\cdot) - \mu(\cdot)\|_{\text{TV}} = \frac{1}{2}\max_{(s_0,a_0,s_1)\in\mathcal{Y}}\sum_{s,a,s'}|P_{\pi_b}^{k+1}((s_0,a_0,s_1),(s,a,s')) - \kappa_b(s)\pi_b(a|s)P_a(s,s')|
$$

$$
\begin{aligned}
&= \frac{1}{2}\max_{s_1\in\mathcal{S}}\sum_{s,a,s'}|P_{\pi_b}^k(s_1,s)\pi_b(a|s)P_a(s,s')-\kappa_b(s)\pi_b(a|s)P_a(s,s')|\\
&= \frac{1}{2}\max_{s_1\in\mathcal{S}}\sum_{s}|P_{\pi_b}^k(s_1,s)-\kappa_b(s)|\\
&= \max_{s\in\mathcal{S}}\|P_{\pi_b}^k(s,\cdot)-\kappa_b(\cdot)\|_{\text{TV}}\\
&\le C_1\sigma_1^k,
\end{aligned}
$$

where $C_1 > 0$ and $\sigma_1 \in (0,1)$ are constants. Note that the last line of the previous inequality follows from Assumption 3.1 and Theorem 4.9 in [26].

(3) (a) Using the Markov property, we have for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and (s,a):

$$
\begin{aligned}
&\mathbb{E}_{S_k\sim\kappa_b}\left[[F(Q,S_k,A_k,S_{k+1})](s,a)\right]\\
&= \mathbb{E}_{S_k\sim\kappa_b}\left[\mathbb{1}_{\{(S_k,A_k)=(s,a)\}}\left(\mathcal{R}(S_k,A_k)+\beta\max_{a'\in\mathcal{A}}Q(S_{k+1},a')-Q(S_k,A_k)\right)+Q(s,a)\right]\\
&= \mathbb{E}_{S_k\sim\kappa_b}\left[\mathbb{1}_{\{(S_k,A_k)=(s,a)\}}\left(\mathcal{R}(S_k,A_k)+\beta\max_{a'\in\mathcal{A}}Q(S_{k+1},a')\right)+(1-\mathbb{1}_{\{(S_k,A_k)=(s,a)\}})Q(S_k,A_k)\right]\\
&= \kappa_b(s)\pi_b(a|s)[\mathcal{H}(Q)](s,a)+(1-\kappa_b(s)\pi_b(a|s))Q(s,a),
\end{aligned}
$$

where $\mathcal{H}:\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the Bellman's optimality operator defined by

$$
[\mathcal{H}(Q)](s,a)=\mathbb{E}[\mathcal{R}(S_k,A_k)+\beta\max_{a'\in\mathcal{A}}Q(S_{k+1},a')\mid S_k=s,A_k=a]
$$

for any (s,a). Now use the definition of the matrix N and we have $\bar{F}(Q)=N\mathcal{H}(Q)+(I-N)Q$.

(b) Since it is well-known that the Bellman's optimality operator $\mathcal{H}$ is a β-contraction with respect to $\|\cdot\|_\infty$, we have for any $Q_1,Q_2\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$
\begin{aligned}
&\|\bar{F}(Q_1)-\bar{F}(Q_2)\|_\infty\\
&= \|N(\mathcal{H}(Q_1)-\mathcal{H}(Q_2))+(I-N)(Q_1-Q_2)\|_\infty\\
&= \max_{(s,a)}|\kappa_b(s)\pi_b(a|s)([\mathcal{H}(Q_1)](s,a)-[\mathcal{H}(Q_2)](s,a))+(1-\kappa_b(s)\pi_b(a|s))(Q_1(s,a)-Q_2(s,a))|\\
&\le \max_{(s,a)}[\kappa_b(s)\pi_b(a|s)|[\mathcal{H}(Q_1)](s,a)-[\mathcal{H}(Q_2)](s,a)|+(1-\kappa_b(s)\pi_b(a|s))|Q_1(s,a)-Q_2(s,a)|]\\
&\le \max_{(s,a)}[\kappa_b(s)\pi_b(a|s)\|\mathcal{H}(Q_1)-\mathcal{H}(Q_2)\|_\infty+(1-\kappa_b(s)\pi_b(a|s))\|Q_1-Q_2\|_\infty]\\
&\le \max_{(s,a)}[\kappa_b(s)\pi_b(a|s)\beta\|Q_1-Q_2\|_\infty+(1-\kappa_b(s)\pi_b(a|s))\|Q_1-Q_2\|_\infty]\\
&= \|Q_1-Q_2\|_\infty\max_{(s,a)}(1-(1-\beta)\kappa_b(s)\pi_b(a|s))\\
&= (1-N_{\min}(1-\beta))\|Q_1-Q_2\|_\infty.
\end{aligned}
$$

Therefore, the operator $\bar{F}(\cdot)$ is a contraction w.r.t. $\|\cdot\|_\infty$, with contraction factor $\gamma_1 = 1 - N_{\min}(1-\beta)$.

(c) It is enough to show that Q^* is a fixed-point of $\bar{F}(\cdot)$, the uniqueness part follows from $\bar{F}(\cdot)$ being a contraction [2]. Using the fact that $\mathcal{H}(Q^*)=Q^*$, we have

$$
\bar{F}(Q^*)=N\mathcal{H}(Q^*)+(I-N)Q^*=NQ^*+(I-N)Q^*=Q^*.
$$

B.2 Proof of Theorem 3.1

Since the contraction norm is $\|\cdot\|_\infty$, Lemma 2.1 (2) is applicable. To apply Theorem 2.1, we first identify the corresponding constants in the following, where we used Proposition 3.1:

$$A = A_1 + A_2 + 1 = 3,\ B = B_1 + B_2 = 1,\ \alpha_1 \leq 3,\ \alpha_2 \geq \frac{1-\gamma_1}{2},\ \alpha_3 \leq \frac{456e\log(|\mathcal{S}||\mathcal{A}|)}{1-\gamma_1},$$
$$c_1 \leq (\|Q_0 - Q^*\|_\infty + \|Q_0\|_\infty + 1)^2,\ c_2 = (3\|Q^*\|_\infty + 1)^2.$$

Now we apply Theorem 2.1 (1). When $\epsilon_k \equiv \epsilon$ with properly chosen ϵ, there exists $K_1 > 0$ such that we have for all $k \geq K_1$:

$$\begin{aligned}
&\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \\
&\leq \alpha_1 c_1 \left(1 - \frac{1-\gamma_1}{2}\epsilon\right)^{k-K_1} + \frac{\alpha_3 c_2}{\alpha_2}\epsilon t_\epsilon(\mathcal{M}_Y) \\
&\leq 3(\|Q_0 - Q^*\|_\infty + \|Q_0\|_\infty + 1)^2 \left(1 - \frac{1-\gamma_1}{2}\epsilon\right)^{k-K_1} + \frac{912e\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma_1)^2}(3\|Q^*\|_\infty + 1)^2 \epsilon t_\epsilon(\mathcal{M}_Y) \\
&= c_{Q,1}\left(1 - \frac{1-\gamma_1}{2}\epsilon\right)^{k-K_1} + c_{Q,2}\frac{\log(|\mathcal{S}||\mathcal{A}|)}{(1-\gamma_1)^2}\epsilon t_\epsilon(\mathcal{M}_Y),
\end{aligned}$$

where $c_{Q,1} = 3(\|Q_0 - Q^*\|_\infty + \|Q_0\|_\infty + 1)^2$ and $c_{Q,2} = 912e(3\|Q^*\|_\infty + 1)^2$.

B.3 Q-Learning with Diminishing Stepsizes

We next present the finite-sample bounds for Q-learning with diminishing stepsizes, whose proof follows by directly applying Theorem 2.1 (2) (b) and (2) (c).

Theorem B.1. *Consider $\{Q_k\}$ of Algorithm (9). Suppose Assumption 3.1 is satisfied, then we have the following results.*

(1) (a) When $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{1}{1-\gamma_1}$ and properly chosen h, there exists $K_1' > 0$ such that the following inequality holds for all $k \geq K_1'$:

$$\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq c_{Q,1}'\left(\frac{K_1'+h}{k+h}\right)^{1/2} + 2c_{Q,2}'\frac{\log(|\mathcal{S}||\mathcal{A})}{(1-\gamma_1)^3}\frac{t_k(\mathcal{M}_Y)}{k+h},$$

where $c_{Q,1}' = 3(\|Q_0 - Q^*\|_\infty + \|Q_0\|_\infty + 1)^2$ and $c_{Q,2}' = 3648e(3\|Q^*\|_\infty + 1)^2$

(b) When $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{2}{1-\gamma_1}$ and properly chosen h, there exists $K_1' > 0$ such that the following inequality holds for all $k \geq K_1'$:

$$\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq c_{Q,1}'\frac{K_1'+h}{k+h} + 4c_{Q,2}'\frac{\log(|\mathcal{S}||\mathcal{A})}{(1-\gamma_1)^3}\frac{t_k(\mathcal{M}_Y)\log(k+h)}{k+h}.$$

(c) When $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{4}{1-\gamma_1}$ and properly chosen h, there exists $K_1' > 0$ such that the following inequality holds for all $k \geq K_1'$:

$$\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq c_{Q,1}'\left(\frac{K_1'+h}{k+h}\right)^2 + 16c_{Q,2}'\frac{\log(|\mathcal{S}||\mathcal{A})}{(1-\gamma_1)^3}\frac{t_k(\mathcal{M}_Y)}{k+h}.$$

(2) When $\epsilon_k = \frac{\epsilon}{(k+h)^\xi}$ with $\xi \in (0,1)$, $\epsilon > 0$, and properly chosen h, there exists $K_1' > 0$ such that the following inequality holds for all $k \geq K_1'$:

$$\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq c_{Q,1}'\exp\left(-\frac{(1-\gamma_1)\epsilon}{2(1-\xi)}((k+h)^{1-\xi} - (K_1'+h)^{1-\xi})\right) + c_{Q,2}'\frac{\log(|\mathcal{S}||\mathcal{A})}{(1-\gamma_1)^2}\frac{t_k(\mathcal{M}_Y)}{k+h}.$$

C V-Trace

C.1 Properties of the V-trace Algorithm

To apply our main results in Section 2, we begin by rewriting the V-trace algorithm (10) in the form of SA algorithm (6). For any $k \geq 0$, let $Y_k = (S_k, A_k, ..., S_{k+n-1}, A_{k+n-1}, S_{k+n})$. It is clear that $\{Y_k\}$ is also a Markov chain, whose state space is denoted by $\mathcal{Y}$. Define an operator $F : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{Y} \mapsto \mathbb{R}^{|\mathcal{S}|}$ by

$$[F(V, y)](s) = \mathbb{1}_{\{s_0 = s\}} \sum_{i=0}^{n-1} \beta^i \left(\prod_{j=0}^{i-1} c(s_j, a_j) \right) \rho(s_i, a_i) \left(\mathcal{R}(s_i, a_i) + \beta V(s_{i+1}) - V(s_i) \right) + V(s)$$

for all $s \in \mathcal{S}$. Then the V-trace update equation (10) can be equivalently written by

$$V_{k+1} = V_k + \epsilon_k (F(V_k, Y_k) - V_k). \tag{20}$$

Under Assumptions 3.2, we next establish the properties of the operator F and the Markov chain $\{Y_k\}$, which allow us to call for our main results in Section 2.

Before that, recall our notations in the following:

- $C, D \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ are diagonal matrices such that $C(s) = \sum_{a \in \mathcal{A}} \min(\bar{c}\pi_b(a|s), \pi(a|s))$ and $D(s) = \sum_{a \in \mathcal{A}} \min(\bar{\rho}\pi_b(a|s), \pi(a|s))$ for all $s \in \mathcal{S}$.
- $C_{\min} = \min_{s \in \mathcal{S}} C(s)$ and $D_{\min} = \min_{s \in \mathcal{S}} D(s)$. Note that we have $0 < C_{\min} \leq D_{\min} \leq 1$ under Assumption 3.2.
- Let $\mathcal{K} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ be a diagonal matrix with diagonal entries $\{\kappa_b(s)\}_{s \in \mathcal{S}}$, and let $\kappa_{\min} = \min_{s \in \mathcal{S}} \kappa_b(s)$.
- Define policies $\pi_{\bar{c}}$ and $\pi_{\bar{\rho}}$ by

$$\pi_{\bar{c}}(a|s) = \frac{\min(\bar{c}\pi_b(a|s), \pi(a|s))}{C(s)}, \quad \text{and} \quad \pi_{\bar{\rho}}(a|s) = \frac{\min(\bar{\rho}\pi_b(a|s), \pi(a|s))}{D(s)}, \quad \forall\, (s, a).$$

Proposition C.1. *Under Assumptions 3.2, the V-trace algorithm (20) has the following properties:*

(1) The operator F satisfies for all $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$ and $y \in \mathcal{Y}$:

(a) $\|F(V_1, y) - F(V_2, y)\|_\infty \leq (2\bar{\rho} + 1) z(\beta, \bar{c}) \|V_1 - V_2\|_\infty$, where $z(\beta, \bar{c}) = \frac{1 - (\beta\bar{c})^n}{1 - \beta\bar{c}}$ when $\beta\bar{c} \neq 1$, and $z(\beta, \bar{c}) = n$ when $\beta\bar{c} = 1$.

(b) $\|F(\mathbf{0}, y)\|_\infty \leq \bar{\rho} z(\beta, \bar{c})$.

(2) The Markov chain $\{Y_k\}$ has a unique stationary distribution, denoted by μ. Moreover, there exists $C_2 > 0$ and $\sigma_2 \in (0, 1)$ such that $\max_{y \in \mathcal{Y}} \|P^{k+n}(y, \cdot) - \mu(\cdot)\|_{TV} \leq C_2 \sigma_2^k$ for all $k \geq 0$.

(3) Define the expected operator $\bar{F} : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ of F by $\bar{F}(V) = \mathbb{E}_{Y \sim \mu}[F(V, Y)]$ for all $V \in \mathbb{R}^{|\mathcal{S}|}$. Then

(a) $\bar{F}$ is explicitly given by $\bar{F}(V) = \left[I - \mathcal{K} \sum_{i=0}^{n-1} (\beta C P_{\pi_{\bar{c}}})^i D (I - \beta P_{\pi_{\bar{\rho}}}) \right] V + \mathcal{K} \sum_{i=0}^{n-1} (\beta C P_{\pi_{\bar{c}}})^i D R_{\pi_{\bar{\rho}}}$.

(b) $\bar{F}$ is a contraction mapping w.r.t. $\|\cdot\|_\infty$, with contraction factor $\gamma_2 := 1 - \kappa_{\min} \frac{(1-\beta)(1-(\beta C_{\min})^n) D_{\min}}{1 - \beta C_{\min}}$.

(c) $\bar{F}$ has a unique fixed-point $V_{\pi_{\bar{\rho}}}$.

Proof of Proposition C.1. (1) Using the definition of $F(V, y)$, we have for any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$, $y \in \mathcal{Y}$, and $s \in \mathcal{S}$:

$$|[F(V_1, y)](s) - [F(V_2, y)](s)|$$

$$= \left| \mathbb{1}_{\{s_0=s\}} \sum_{i=0}^{n-1} \beta^i \left(\prod_{j=0}^{i-1} c(s_j, a_j) \right) \rho(s_i, a_i) \left(\beta(V_1(s_{i+1}) - V_2(s_{i+1})) - (V_1(s_i) - V_2(s_i)) \right) + V_1(s) - V_2(s) \right|$$

$$\leq 2\|V_1 - V_2\|_\infty \sum_{i=0}^{n-1} \beta^i \bar{c}^i \bar{\rho} + \|V_1 - V_2\|_\infty$$

$$\leq \begin{cases} \frac{(2\bar{\rho}+1)(1-(\beta\bar{c})^n)}{1-\beta\bar{c}} \|V_1 - V_2\|_\infty, & \beta\bar{c} \neq 1, \\ (2\bar{\rho}+1)n\|V_1 - V_2\|_\infty, & \beta\bar{c} = 1. \end{cases}$$

It follows that $\|F(V_1, y) - F(V_2, y)\|_\infty \leq (2\bar{\rho}+1)z(\beta, \bar{c})\|V_1 - V_2\|_\infty$.

For any $y \in \mathcal{Y}$ and $s \in \mathcal{S}$, we have

$$\begin{aligned} |[F(\mathbf{0}, y)](s)| &= \left| \mathbb{1}_{\{s_0=s\}} \sum_{i=0}^{n-1} \beta^i \left(\prod_{j=0}^{i-1} c(s_j, a_j) \right) \rho(s_i, a_i) \mathcal{R}(s_i, a_i) \right| \\ &\leq \sum_{i=0}^{n-1} \beta^i \bar{c}^i \bar{\rho} \\ &= \begin{cases} \frac{\bar{\rho}(1-(\beta\bar{c})^n)}{1-\beta\bar{c}}, & \beta\bar{c} \neq 1, \\ n\bar{\rho}, & \beta\bar{c} = 1. \end{cases} \end{aligned}$$

It follows that $\|F(0, y)\|_\infty \leq z(\beta, \bar{c})$.

(2) The proof is identical to that of Proposition 3.1 (2).

(3) (a) Using the definition of $\bar{F}(\cdot)$, we have for any $V \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$:

$$\begin{aligned} &\mathbb{E}_{Y\sim\mu}[[F(V, Y)](s)] \\ &= \mathbb{E}_{S_0\sim\kappa} \left[\mathbb{1}_{\{S_0=s\}} \sum_{i=0}^{n-1} \beta^i \left(\prod_{j=0}^{i-1} c(S_j, A_j) \right) \rho(S_i, A_i) \left(\mathcal{R}(S_i, A_i) + \beta V(S_{i+1}) - V(S_i) \right) \right] + V(s). \end{aligned}$$

For any $0 \leq i \leq n-1$, we have by the Markov property and the tower property of conditional expectation that

$$\begin{aligned} &\mathbb{E}_{S_0\sim\kappa} \left[\mathbb{1}_{\{S_0=s\}} \beta^i \left(\prod_{j=0}^{i-1} c(S_j, A_j) \right) \rho(S_i, A_i) \left(\mathcal{R}(S_i, A_i) + \beta V(S_{i+1}) - V(S_i) \right) \right] \\ &= \mathbb{E}_{S_0\sim\kappa} \left[\mathbb{1}_{\{S_0=s\}} \beta^i \left(\prod_{j=0}^{i-1} c(S_j, A_j) \right) \mathbb{E}[\rho(S_i, A_i) \left(\mathcal{R}(S_i, A_i) + \beta V(S_{i+1}) - V(S_i) \right) \mid S_i] \right] \\ &= \mathbb{E}_{S_0\sim\kappa} \left[\mathbb{1}_{\{S_0=s\}} \beta^i \left(\prod_{j=0}^{i-1} c(S_j, A_j) \right) [D(R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}} V - V)](S_i) \right] \\ &= \mathbb{E}_{S_0\sim\kappa} \left[\mathbb{1}_{\{S_0=s\}} \beta^i \left(\prod_{j=0}^{i-2} c(S_j, A_j) \right) \mathbb{E}[c(S_{i-1}, A_{i-1})[D(R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}} V - V)](S_i) \mid S_{i-1}] \right] \\ &= \mathbb{E}_{S_0\sim\kappa} \left[\mathbb{1}_{\{S_0=s\}} \beta^i \left(\prod_{j=0}^{i-2} c(S_j, A_j) \right) [(CP_{\pi_{\bar{c}}}) D(R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}} V - V)](S_{i-1}) \right] \\ &= \cdots \end{aligned}$$

$$= \mathbb{E}_{S_0\sim\kappa}\left[\mathbb{1}_{\{S_0=s\}}\beta^i[(CP_{\pi_{\bar{c}}})^i D(R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}}V - V)](S_0)\right]$$
$$= [\mathcal{K}(\beta CP_{\pi_{\bar{c}}})^i D(R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}}V - V)](s).$$

Therefore, we have

$$\begin{aligned}\bar{F}(V) &= \sum_{i=0}^{n-1}\mathcal{K}(\beta CP_{\pi_{\bar{c}}})^i D(R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}}V - V) + V \\ &= \left[I - \mathcal{K}\sum_{i=0}^{n-1}(\beta CP_{\pi_{\bar{c}}})^i D(I - \beta P_{\pi_{\bar{\rho}}})\right]V + \sum_{i=0}^{n-1}\mathcal{K}(\beta CP_{\pi_{\bar{c}}})^i DR_{\pi_{\bar{\rho}}}.\end{aligned} \tag{21}$$

(b) For any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$, we have

$$\begin{aligned}\left\|\bar{F}(V_1) - \bar{F}(V_2)\right\|_\infty &= \left\|\left[I - \mathcal{K}\sum_{i=0}^{n-1}(\beta CP_{\pi_{\bar{c}}})^i D(I - \beta P_{\pi_{\bar{\rho}}})\right](V_1 - V_2)\right\|_\infty \\ &\leq \left\|I - \mathcal{K}\sum_{i=0}^{n-1}(\beta CP_{\pi_{\bar{c}}})^i D(I - \beta P_{\pi_{\bar{\rho}}})\right\|_\infty \|V_1 - V_2\|_\infty\end{aligned}$$

For simplicity of notations, denote $G = I - \mathcal{K}\sum_{i=0}^{n-1}(\beta CP_{\pi_{\bar{c}}})^i D(I - \beta P_{\pi_{\bar{\rho}}})$. To evaluate the ℓ_∞-norm of G, we first show that G has non-negative entries. Note that G can be equivalently written by

$$\begin{aligned}G &= I - \mathcal{K}\sum_{i=0}^{n-1}(\beta CP_{\pi_{\bar{c}}})^i D(I - \beta P_{\pi_{\bar{\rho}}}) \\ &= I - \mathcal{K}D - \mathcal{K}\sum_{i=0}^{n-2}(\beta CP_{\pi_{\bar{c}}})^{i+1}D + \mathcal{K}\sum_{i=0}^{n-2}(\beta CP_{\pi_{\bar{c}}})^i \beta DP_{\pi_{\bar{\rho}}} + \mathcal{K}(\beta CP_{\pi_{\bar{c}}})^{n-1}\beta DP_{\pi_{\bar{\rho}}} \\ &= I - \mathcal{K}D + \mathcal{K}\sum_{i=0}^{n-2}(\beta CP_{\pi_{\bar{c}}})^i \beta(DP_{\pi_{\bar{\rho}}} - CP_{\pi_{\bar{c}}}D) + \mathcal{K}(\beta CP_{\pi_{\bar{c}}})^{n-1}\beta DP_{\pi_{\bar{\rho}}}.\end{aligned} \tag{22}$$

In view of Eq. (22), it remains to show that the matrix $DP_{\pi_{\bar{\rho}}} - CP_{\pi_{\bar{c}}}D$ has non-negative entries. For any $s, s' \in \mathcal{S}$, we have

$$\begin{aligned}[DP_{\pi_{\bar{\rho}}} - CP_{\pi_{\bar{c}}}D](s,s') &= D(s)P_{\pi_{\bar{\rho}}}(s,s') - C(s)P_{\pi_{\bar{c}}}(s,s')D(s') \\ &\geq D(s)P_{\pi_{\bar{\rho}}}(s,s') - C(s)P_{\pi_{\bar{c}}}(s,s') && (D(s') \leq 1) \\ &= D(s)\sum_{a\in\mathcal{A}}\pi_{\bar{\rho}}(a|s)P_a(s,s') - C(s)\sum_{a\in\mathcal{A}}\pi_{\bar{c}}(a|s)P_a(s,s') \\ &= \sum_{a\in\mathcal{A}}\left[\min\left(\bar{\rho}\pi_b(a|s), \pi(a|s)\right) - \min\left(\bar{c}\pi_b(a|s), \pi(a|s)\right)\right]P_a(s,s') \\ &\geq 0,\end{aligned} \tag{23}$$

where the last line follows from $\bar{c} \leq \bar{\rho}$.

Now since the matrix G has non-negative entries, we have

$$\begin{aligned}\|G\|_\infty &= \|G\mathbf{1}\|_\infty && (\mathbf{1} = (1,1,\cdots,1)^\top) \\ &= \left\|\left[I - \mathcal{K}\sum_{i=0}^{n-1}(\beta CP_{\pi_{\bar{c}}})^i D(I - \beta P_{\pi_{\bar{\rho}}})\right]\mathbf{1}\right\|_\infty\end{aligned}$$

$$
\begin{aligned}
&= \left\| \mathbf{1} - (1-\beta)\mathcal{K} \sum_{i=0}^{n-1} (\beta C P_{\pi_{\bar{c}}})^i D \mathbf{1} \right\|_\infty \\
&\leq 1 - \kappa_{\min}(1-\beta) \sum_{i=0}^{n-1} (\beta C_{\min})^i D_{\min} && (0 < C(s) \leq D(s) \leq 1 \text{ for all } s) \\
&= 1 - \kappa_{\min} \frac{(1-\beta)(1-(\beta C_{\min})^n) D_{\min}}{1 - \beta C_{\min}} \\
&= \gamma_2.
\end{aligned}
$$

It follows that $\left\|\bar{F}(V_1) - \bar{F}(V_2)\right\|_\infty \leq \|G\|_\infty \|V_1 - V_2\|_\infty \leq \gamma_2 \|V_1 - V_2\|_\infty$.

(c) It is enough to show that $V_{\pi_{\bar{\rho}}}$ is a fixed-point of $\bar{F}$, the uniqueness follows from $\bar{F}$ being a contraction operator. Using the Bellman's equation $V_{\pi_{\bar{\rho}}} = R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}} V_{\pi_{\bar{\rho}}}$, we have by Eq. (21) that

$$
\bar{F}(V_{\pi_{\bar{\rho}}}) = V_{\pi_{\bar{\rho}}} + \mathcal{K} \sum_{i=0}^{n-1} (\beta C P_{\pi_{\bar{c}}})^i D (R_{\pi_{\bar{\rho}}} + \beta P_{\pi_{\bar{\rho}}} V_{\pi_{\bar{\rho}}} - V_{\pi_{\bar{\rho}}}) = V_{\pi_{\bar{\rho}}}.
$$

■

C.2 Proof of Theorem 3.2

We will apply Theorem 2.1 and Lemma 2.1 (2) to the V-trace algorithm. We begin by identifying the constants:

$$
\begin{aligned}
&A = A_1 + A_2 + 1 \leq 2(\bar{\rho}+1) z(\beta, \bar{c}),\ B = B_1 + B_2 = \bar{\rho} z(\beta, \bar{c}),\ \alpha_1 \leq 3,\ \alpha_2 \geq \frac{1-\gamma_2}{2},\ \alpha_3 \leq \frac{456 e \log(|\mathcal{S}|)}{1-\gamma_2} \\
&c_1 \leq (\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2,\ c_2 = 4(\bar{\rho}+1)^2 z(\beta, \bar{c})^2 (\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2.
\end{aligned}
$$

Now we apply Theorem 2.1 (2) (a). When $\epsilon_k = \epsilon$ for all $k \geq 0$ with properly chosen ϵ, then there exists $K_2 > 0$ such that the following inequality holds for all $k \geq K_2$:

$$
\begin{aligned}
\mathbb{E}[\|V_k - V_{\pi_{\bar{\rho}}}\|_\infty^2] &\leq \alpha_1 c_1 (1-\alpha_2 \epsilon)^{k-K_2} + \frac{\alpha_3 c_2}{\alpha_2} \epsilon t_\epsilon(\mathcal{M}_Y) \\
&\leq 3(\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2 \left(1 - \frac{1-\gamma_2}{2}\epsilon\right)^{k-K_2} \\
&\quad + \frac{3648 e \log(|\mathcal{S}|)}{(1-\gamma_2)^2} (\bar{\rho}+1)^2 z(\beta, \bar{c})^2 (\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2 \epsilon (t_\epsilon(\mathcal{M}_S) + n) \\
&= c_{V,1} \left(1 - \frac{1-\gamma_2}{2}\epsilon\right)^{k-K_2} + c_{V,2} \frac{\log(|\mathcal{S}|)}{(1-\gamma_2)^2} (\bar{\rho}+1)^2 z(\beta, \bar{c})^2 \epsilon (t_\epsilon(\mathcal{M}_S) + n),
\end{aligned}
$$

where $c_{V,1} = 3(\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2$ and $c_{V,2} = 3648 e (\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2$.

C.3 V-Trace with Diminishing Stepsizes

We here only present using linear stepsize that achieves the optimal convergence rate (Theorem 2.1 (2) (b) (iii)).

Theorem C.1. *Consider $\{V_k\}$ of Algorithm (10). Suppose Assumption 3.2 is satisfied and $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{4}{1-\gamma_2}$ and properly chosen h. Then there exists $K_2' > 0$ such that the following inequality holds for all $k \geq K_2'$:*

$$\mathbb{E}[\|V_k - V_{\pi_{\bar{\rho}}}\|_\infty^2] \leq c_{V,1}' \frac{K_2'+h}{k+h} + c_{V,2}' \frac{\log(|\mathcal{S}|)}{(1-\gamma_2)^3}(\bar{\rho}+1)^2 z(\beta,\bar{c})^2 \frac{t_k(\mathcal{M}_S)+n}{k+h},$$

where $c_{V,1}' = 3\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2$ and $c_{V,2}' = 233472e^2(\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2$.

Proof of Theorem C.1. The corresponding constants have been identified in the proof of Theorem 3.2. Now apply Theorem 2.1 (2) (c). When $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{4}{1-\gamma_2}$ and properly chosen h, there exists $K_2' > 0$ such that we have for all $k \geq K_2'$:

$$\begin{aligned}
\mathbb{E}[\|V_k - V_{\pi_{\bar{\rho}}}\|_\infty^2] &\leq \alpha_1 c_1 \left(\frac{K_2'+h}{k+h}\right)^{\alpha_2\epsilon} + \frac{8e\epsilon^2\alpha_3 c_2}{\alpha_2\epsilon - 1}\frac{t_k(\mathcal{M}_Y)}{k+h} \\
&\leq 3\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2 \frac{K_2'+h}{k+h} \\
&\quad + 233472e^2 \frac{\log(|\mathcal{S}|)}{(1-\gamma_2)^3}(\bar{\rho}+1)^2 z(\beta,\bar{c})^2 (\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2 \frac{t_k(\mathcal{M}_S)+n}{k+h} \\
&= c_{V,1}' \frac{K_2'+h}{k+h} + c_{V,2}' \frac{\log(|\mathcal{S}|)}{(1-\gamma_2)^3}(\bar{\rho}+1)^2 z(\beta,\bar{c})^2 \frac{t_k(\mathcal{M}_S)+n}{k+h},
\end{aligned}$$

where $c_{V,1}' = 3\|V_0 - V_{\pi_{\bar{\rho}}}\|_\infty + \|V_0\|_\infty + 1)^2$ and $c_{V,2}' = 233472e^2(\|V_{\pi_{\bar{\rho}}}\|_\infty + 1)^2$. ■

D n-step TD

In this section, we study the convergence bounds of the n-step TD-learning algorithm, which can be viewed as a special case of the V-trace algorithm with $\pi_b = \pi$ and $\bar{c} = \bar{\rho} = 1$. Therefore, one can directly apply Theorem 3.2 to this setting and obtain finite-sample bounds for n-step TD. However, we will show that due to on-policy sampling there are better properties (i.e., $\|\cdot\|_2$-contraction) of the n-step TD algorithm we can exploit, which enables us to obtain tighter bounds.

Observe that in the case of on-policy n-step TD, the update equation (10) simplifies to:

$$V_{k+1}(s) = V_k(s) + \epsilon_k \Gamma_3(V_k, S_k, A_k, ..., S_{k+n}) \tag{24}$$

when $s = S_k$, and $V_{k+1}(s) = V_k(s)$ otherwise, where $\Gamma_3(V_k, S_k, A_k, ..., S_{k+n}) = \sum_{i=0}^{n-1} \beta^i \mathcal{R}(S_{k+i}, A_{k+i}) + \beta^n V_k(S_{k+n}) - V_k(S_k)$ is the n-step temporal difference.

An important idea in the n-step TD is to use the parameter n to adjust the bootstrapping effect. When $n = 0$, Eq. (24) is the standard TD-learning update, which corresponds to extreme bootstrapping. When $n = \infty$, Eq. (24) is the Monte Carlo method for estimating V_π, which corresponds to no bootstrapping. A long-standing question in RL is about the efficiency of bootstrapping, i.e., the choice of n that leads to the optimal performance of the algorithm.

In the following sections, we will establish finite-sample convergence bounds of the n-step TD. By evaluating the resulting bounds as a function of n, we provide theoretical insight into the bias-variance trade-off in terms of n. To achieve that, we need the following assumption.

Assumption D.1. The Markov chain $\mathcal{M}_{\mathcal{S}} = \{S_k\}$ induced by the target policy π is irreducible and aperiodic.

Assumption D.1 implies that $\{S_k\}$ has a unique stationary distribution (denoted by $\kappa \in \Delta^{|\mathcal{S}|}$), and the geometric mixing property [26].

D.1 Properties of the n-step TD Algorithm

To begin with, we rewrite the update equation (24) in the form of the SA algorithm studied in Section 2. Let $Y_k = (S_k, A_k, ..., S_{k+n-1}, A_{k+n-1}, S_{k+n})$ for all $k \geq 0$. It is clear that $\{Y_k\}$ is a Markov chain, whose state-space is denoted by $\mathcal{Y}$. Define an operator $F : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{Y} \mapsto \mathbb{R}^{|\mathcal{S}|}$ by

$$[F(V, y)](s) = \mathbb{1}_{\{s_0=s\}} \left(\sum_{i=0}^{n-1} \beta^i \mathcal{R}(s_i, a_i) + \beta^n V_k(s_n) - V_k(s_0) \right) + V(s), \ \forall\, s \in \mathcal{S}.$$

Then the n-step TD algorithm (24) can be equivalently written by

$$V_{k+1} = V_k + \epsilon_k (F(V_k, Y_k) - V_k).$$

We next establish the properties of the n-step TD algorithm in the following proposition. Let $\mathcal{K} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ be a diagonal matrix with diagonal entries $\{\kappa(s)\}_{s\in\mathcal{S}}$, and let $\kappa_{\min} = \min_{s\in\mathcal{S}} \kappa(s)$.

Proposition D.1. *Under Assumption D.1, the n-step TD algorithm (24) has the following properties:*

(1) The operator F satisfies for all $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$ and $y \in \mathcal{Y}$:

(a) $\|F(V_1, y) - F(V_2, y)\|_2 \leq 3\|V_1 - V_2\|_2$.

(b) $\|F(\mathbf{0}, y)\|_2 \leq \frac{1}{1-\beta}$.

(2) The Markov chain $\{Y_k\}$ has a unique stationary distribution, denoted by μ. Moreover, there exists $C_3 > 0$ and $\sigma_3 \in (0, 1)$ such that $\max_{y\in\mathcal{Y}} \|P^{k+n}(y, \cdot) - \mu(\cdot)\|_{TV} \leq C_3\sigma_3^k$ for all $k \geq 0$.

(3) Define the expected operator $\bar{F} : \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ of F by $\bar{F}(V) = \mathbb{E}_{Y\sim\mu}[F(V, Y)]$ for all $V \in \mathbb{R}^{|\mathcal{S}|}$. Then

(a) $\bar{F}$ is a linear operator given by $\bar{F}(V) = \left[I - \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i(I - \beta P_\pi)\right] V + \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i R_\pi$.

(b) $\bar{F}$ is a contraction mapping with respect to $\|\cdot\|_p$ for any $p \in [1, \infty]$, with a common contraction factor $\gamma_3 := 1 - \kappa_{\min}(1 - \beta^n)$.

(c) $\bar{F}$ has a unique fixed-point V_π.

Proof of Proposition D.1. (1) (a) For any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$ and $y \in \mathcal{Y}$, we have

$$\begin{aligned}
&\|F(V_1, y) - F(V_2, y)\|_2^2 \\
={}& \sum_{s\in\mathcal{S}} \left[\mathbb{1}_{\{s_0=s\}} \left(\beta^n(V_1(s_n) - V_2(s_n)) - (V_1(s_0) - V_2(s_0))\right) + V_1(s) - V_2(s)\right]^2 \\
\leq{}& \sum_{s\in\mathcal{S}} \left[\mathbb{1}_{\{s_0=s\}}(\beta^n + 1)\|V_1 - V_2\|_2 + \|V_1 - V_2\|_2\right]^2 \\
\leq{}& 4\sum_{s\in\mathcal{S}} \mathbb{1}_{\{s_0=s\}}\|V_1 - V_2\|_2^2 + \|V_1 - V_2\|_2^2 + 4\sum_{s\in\mathcal{S}} \mathbb{1}_{\{s_0=s\}}\|V_1 - V_2\|_2^2 \\
\leq{}& 9\|V_1 - V_2\|_2^2.
\end{aligned}$$

Hence we have $\|F(V_1, y) - F(V_2, y)\|_2 \leq 3\|V_1 - V_2\|_2$.

(b) For any $y \in \mathcal{Y}$, we have

$$\|F(\mathbf{0}, y)\|_2^2 = \sum_{s\in\mathcal{S}} \mathbb{1}_{\{s_0=s\}} \left(\sum_{i=0}^{n-1} \beta^i \mathcal{R}(s_i, a_i)\right)^2 \leq \sum_{s\in\mathcal{S}} \mathbb{1}_{\{s_0=s\}} \left(\sum_{i=0}^{n-1} \beta^i\right)^2 \leq \frac{1}{(1-\beta)^2}.$$

Hence we have $\|F(\mathbf{0}, y)\|_2 \leq \frac{1}{1-\beta}$.

(2) The proof is identical to that of Proposition 3.1 (2).

(3) (a) Since n-step TD is a special case of V-trace, we can directly apply Proposition C.1 (3) (a) here. Observe that when $\pi = \pi_b$ and $\bar{c} = \bar{\rho} = 1$, we have $C = D = I$ and $P_{\pi_{\bar{c}}} = P_{\pi_{\bar{\rho}}} = P_\pi$. Hence we have

$$\bar{F}(V) = \left[I - \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i(I - \beta P_\pi)\right] V + \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i R_\pi.$$

(b) For any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$ and $p \geq 1$, we have

$$\begin{aligned}
\|\bar{F}(V_1) - \bar{F}(V_2)\|_p &= \left\|\left[I - \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i(I - \beta P_\pi)\right](V_1 - V_2)\right\|_p \\
&\leq \left\|I - \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i(I - \beta P_\pi)\right\|_p \|V_1 - V_2\|_p.
\end{aligned}$$

For simplicity of notations, we denote $G = I - \mathcal{K}\sum_{i=0}^{n-1}(\beta P_\pi)^i(I - \beta P_\pi)$. Since G has non-negative entries (established in the proof of Proposition C.1 (3) (b)), we have

$$\|G\|_\infty = \|G\mathbf{1}\|_\infty = \left\|\mathbf{1} - \kappa\sum_{i=0}^{n-1}\beta^i(1-\beta)\right\|_\infty = 1 - \kappa_{\min}(1-\beta^n).$$

Moreover, using the fact that κ is the stationary distribution of P_π (i.e., $\kappa^\top P_\pi = \kappa^\top$), we have

$$\|G\|_1 = \|\mathbf{1}^\top G\|_\infty = \left\|\mathbf{1}^\top - \kappa^\top\sum_{i=0}^{n-1}\beta^i(1-\beta)\right\|_\infty = 1 - \kappa_{\min}(1-\beta^n).$$

To proceed, we need the following lemma.

Lemma D.1. *Let $G \in \mathbb{R}^{d\times d}$ be a matrix with non-negative entries. Then we have for all $p \in [1, \infty]$:*

$$\|G\|_p \leq \|G\|_1^{1/p}\|G\|_\infty^{1-1/p}.$$

Proof of Lemma D.1. The result clearly holds when $p = 1$ or $p = \infty$. Now consider $p \in (1, \infty)$. Using the definition of induced matrix norm, we have for any $x \neq 0$:

$$\begin{aligned}
\|Gx\|_p^p &= \sum_{i=1}^d \left(\sum_{j=1}^d G_{ij}x_j\right)^p \\
&= \sum_{i=1}^d [G\mathbf{1}]_i^p \left(\sum_{j=1}^d \frac{G_{ij}}{[G\mathbf{1}]_i}x_j\right)^p \\
&\leq \sum_{i=1}^d [G\mathbf{1}]_i^{p-1}\sum_{j=1}^d G_{ij}x_j^p && \text{(Jensen's inequality)}
\end{aligned}$$

$$\leq \|G\|_\infty^{p-1} \sum_{j=1}^d x_j^p \sum_{i=1}^d G_{ij}$$
$$= \|G\|_\infty^{p-1} \sum_{j=1}^d x_j^p [\mathbf{1}^\top G]_j$$
$$\leq \|G\|_\infty^{p-1} \|G\|_1 \|x\|_p^p.$$

It follows that $\|G\|_p \leq \|G\|_1^{1/p} \|G\|_\infty^{1-1/p}$. ■

Using Lemma D.1 and we have

$$\|G\|_p \leq \|G\|_1^{1/p} \|G\|_\infty^{1-1/p} \leq 1 - \kappa_{\min}(1-\beta^n) = \gamma_3.$$

Therefore, we have $\|\bar{F}(V_1) - \bar{F}(V_2)\|_2 \leq \gamma_3 \|V_1 - V_2\|_2$. Hence the operator $\bar{F}(\cdot)$ is a contraction mapping with respect to $\|\cdot\|_2$, with contraction factor γ_3.

(c) The proof is identical to that of Proposition C.1 (3) (c).

■

D.2 Finite-Sample Convergence Bounds of n-Step TD

Next, we will use the $\|\cdot\|_2$-contraction property to derive finite-sample convergence bounds of Algorithm (24). Note that Lemma 2.1 (1) is applicable in this case. Let $\kappa_{\min} = \min_{s\in\mathcal{S}} \kappa(s)$.

Theorem D.1. *Consider $\{V_k\}$ of Algorithm (24). Suppose that Assumption D.1 is satisfied, and $\epsilon_k \equiv \epsilon$ with properly chosen ϵ. Then there exists $K_3 > 0$ such that the following inequality holds for all $k \geq K_3$:*

$$\mathbb{E}[\|V_k - V_\pi\|_2^2] \leq \hat{c}_1 \left(1 - (1-\gamma_3)\epsilon\right)^{k-K_3} + \hat{c}_2 \frac{\epsilon(t_\epsilon(\mathcal{M}_S) + n)}{(1-\beta)^2(1-\gamma_3)},$$

where $\hat{c}_1 = (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2$ and $\hat{c}_2 = 228(4\|V_\pi\|_2 + 1)^2$.

Proof of Theorem D.1. We will apply Theorem and Lemma 2.1 (1) to the n-step TD algorithm. We begin by identifying the constants:

$$A = A_1 + A_2 + 1 = 4,\ B = B_1 + B_2 = \frac{1}{1-\beta},\ \alpha_1 \leq 1,\ \alpha_2 \geq 1 - \gamma_3,\ \alpha_3 \leq 228$$
$$c_1 \leq (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2,\ c_2 = \frac{1}{(1-\beta)^2}(4\|V_\pi\|_2 + 1)^2.$$

Now apply Theorem 2.1 (2) (a). When $\epsilon_k = \epsilon$ for all $k \geq 0$ with properly chosen ϵ, there exists $K_3 > 0$ such that we have for all $k \geq K_3$:

$$\begin{aligned}
&\mathbb{E}[\|V_k - V_\pi\|_2^2] \\
&\leq \alpha_1 c_1 (1-\alpha_2\epsilon)^{k-K_3} + \frac{\alpha_3 c_2}{\alpha_2} \epsilon t_\epsilon(\mathcal{M}_Y) \\
&\leq (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2 (1 - (1-\gamma_3)\epsilon)^{k-K_3} + \frac{228}{1-\gamma_3} \frac{1}{(1-\beta)^2} (4\|V_\pi\|_2 + 1)^2 \epsilon (t_\epsilon(\mathcal{M}_S) + n) \\
&= \hat{c}_1 (1 - (1-\gamma_3)\epsilon)^{k-K_3} + \hat{c}_2 \frac{\epsilon(t_\epsilon(\mathcal{M}_S) + n)}{(1-\gamma_3)(1-\beta)^2},
\end{aligned}$$

where $\hat{c}_1 = (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2$ and $\hat{c}_2 = 228(4\|V_\pi\|_2 + 1)^2$. ■

Similar to V-trace, the bias-variance trade-off in choosing n is also present in n-step TD. To see this, observe from Theorem D.1 that the bias is of the size $(1-\Theta(1-\beta^n))^k$ while the variance is of the size $\Theta(\frac{n}{1-\beta^n})$, where we used the explicit expression of γ_3 in Proposition D.1.

For n-step TD with diminishing stepsize, we here only present the result for using linear diminishing stepsize that achieves the optimal convergence rate (Theorem 2.1 (2) (b) (iii)).

Theorem D.2. *Consider $\{V_k\}$ of Algorithm (24). Suppose that Assumption D.1 is satisfied and $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{2}{1-\gamma_3}$ and properly chosen h. Then there exists $K_3' > 0$ such that the following inequality holds for all $k \geq K_3'$:*

$$\mathbb{E}[\|V_k - V_\pi\|_c^2] \leq \hat{c}_1' \frac{K_3'+h}{k+h} + \hat{c}_2' \frac{t_k(\mathcal{M}_S)+n}{(1-\gamma_3)^2(1-\beta)^2(k+h)},$$

where $\hat{c}_1' = (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2$ and $\hat{c}_2' = 7296e(4\|V_\pi\|_2+1)^2$.

Proof of Theorem D.2. The constants are already identified in the proof of Theorem D.1. Apply Theorem 2.1) (2) (b) (iii), when $\epsilon_k = \frac{\epsilon}{k+h}$ with $\epsilon = \frac{2}{1-\gamma_3}$ and properly chosen h, there exists $K_3' > 0$ such that we have for all $k \geq K_3'$:

$$\begin{aligned}
&\mathbb{E}[\|V_k - V_\pi\|_c^2] \\
&\leq \alpha_1 c_1 \left(\frac{K_3'+h}{k+h}\right)^{\alpha_2\epsilon} + \frac{8e\epsilon^2\alpha_3 c_2}{\alpha_2\epsilon - 1}\frac{t_k(\mathcal{M}_Y)}{k+h} \\
&\leq (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2 \frac{K_3'+h}{k+h} + \frac{7296e}{(1-\gamma_3)^2}\frac{1}{(1-\beta)^2}(4\|V_\pi\|_2+1)^2\frac{t_k(\mathcal{M}_Y)}{k+h} \\
&= \hat{c}_1' \frac{K_3'+h}{k+h} + \hat{c}_2' \frac{t_k(\mathcal{M}_S)+n}{(1-\gamma_3)^2(1-\beta)^2(k+h)},
\end{aligned}$$

where $\hat{c}_1' = (\|V_0 - V_\pi\|_2 + \|V_0\|_2 + 4)^2$ and $\hat{c}_2' = 7296e(4\|V_\pi\|_2+1)^2$. ■

E TD(λ)

E.1 Properties of the TD(λ) Algorithm

To utilize our SA results, we begin by rewriting the update equation of the TD(λ) algorithm (11) in the form of the stochastic iterative algorithm (6). For ease of exposition, we consider only using constant stepsize in the TD(λ) algorithm, i.e., $\epsilon_k = \epsilon$ for all $k \geq 0$.

For any $k \geq 0$, let $Y_k = (S_0, ..., S_k, A_k, S_{k+1})$ (which takes value in $\mathcal{Y}_k := \mathcal{S}^{k+2} \times \mathcal{A}$), and define a time-varying operator $F_k : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{Y}_k \mapsto \mathbb{R}^{|\mathcal{S}|}$ by $[F_k(V, y)](s) = [F_k(V, s_0, ..., s_k, a_k, s_{k+1})](s) = \Gamma_4(V, s_k, a_k, s_{k+1}) \sum_{i=0}^k (\beta\lambda)^{k-i} \mathbb{1}_{\{s_i = s\}} + V(s)$ for all $s \in \mathcal{S}$. Note that the sequence $\{Y_k\}$ is not a Markov chain since it has a time-varying state-space. Using the notations of $\{Y_k\}$ and $F_k(\cdot,\cdot)$, we can rewrite the update equation of the TD(λ) algorithm by

$$V_{k+1} = V_k + \epsilon\left(F_k(V_k, Y_k) - V_k\right). \tag{25}$$

Although Eq. (25) is similar to the update equation for SA algorithm (6), since the sequence $\{Y_k\}$ is not a Markov chain and the operator $F_k(\cdot,\cdot)$ is time-varying, our Theorem 2.1 is not directly applicable.

To overcome this difficulty, let us carefully look at the operator $F_k(\cdot,\cdot)$. Although $F_k(V_k, Y_k)$ depends on the whole trajectory of states visited before (through the term $\sum_{i=0}^k (\beta\lambda)^{k-i}\mathbb{1}_{\{S_i=s\}}$), due to the geometric factor $(\beta\lambda)^{k-i}$, the states visited during the early stage of the iteration are not important.

Inspired by this observation, we define the truncated sequence $\{Y_k^\tau\}$ of $\{Y_k\}$ by $Y_k^\tau = (S_{k-\tau}, ..., S_k, A_k, S_{k+1})$ for all $k \geq \tau$, where τ is a *fixed* non-negative integer. Note that the random process $\mathcal{M}_Y = \{Y_k^\tau\}$ is now a Markov chain, whose state-space is denoted by $\mathcal{Y}_\tau$. Similarly, we define the truncated operator $F_k^\tau : \mathbb{R}^{|\mathcal{S}|} \times \mathcal{Y}_\tau \mapsto \mathbb{R}^{|\mathcal{S}|}$ of $F_k(\cdot,\cdot)$ by $[F_k^\tau(V, s_{k-\tau}, ..., s_k, a_k, s_{k+1})](s) = \Gamma_4(V, s_k, a_k, s_{k+1}) \sum_{i=k-\tau}^k (\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}} + V(s)$ for all $s \in \mathcal{S}$. Using the above notations, we can further rewrite the update equation (25) by

$$V_{k+1} = V_k + \epsilon\left(F_k^\tau(V_k, Y_k^\tau) - V_k\right) + \underbrace{\epsilon\left(F_k(V_k, Y_k) - F_k^\tau(V_k, Y_k^\tau)\right)}_{\text{The Error Term}}. \tag{26}$$

Now, we want to argue that when the truncation level τ is large enough, the last term on the RHS of the previous equation is negligible compared to the other two terms. In fact, we have the following result.

Lemma E.1. *For all $k \geq 0$ and $\tau \in [0, k]$, denote $y = (s_0, ..., s_k, a_k, s_{k+1})$ and $y_\tau = (s_{k-\tau}, ..., s_k, a_k, s_{k+1})$. Then the following inequality holds for all $V \in \mathbb{R}^{|\mathcal{S}|}$: $\|F_k^\tau(V, y_\tau) - F_k(V, y)\|_2 \leq \frac{(\beta\lambda)^{\tau+1}}{1-\beta\lambda}(1 + 2\|V\|_2)$.*

Proof of Lemma E.1. For any $V \in \mathbb{R}^{|\mathcal{S}|}$ and $(s_0, ..., s_k, a_k, s_{k+1})$, we have by definition of the operators $F_k^\tau(\cdot,\cdot)$ and $F_k(\cdot,\cdot)$ that

$$\begin{aligned}
&\|F_k^\tau(V, s_{k-\tau}, ..., s_k, a_k, s_{k+1}) - F_k(V, s_0, ..., s_k, a_k, s_{k+1})\|_2^2 \\
&= \sum_{s\in\mathcal{S}}\left[(\mathcal{R}(s_k, a_k) + \beta V(s_{k+1}) - V(s_k))\sum_{i=0}^{k-\tau-1}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right]^2 \\
&\leq (1 + 2\|V\|_2)^2 \sum_{s\in\mathcal{S}}\left[\sum_{i=0}^{k-\tau-1}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right]^2 \\
&= (\beta\lambda)^{2(\tau+1)}(1 + 2\|V\|_2)^2 \sum_{s\in\mathcal{S}}\left[\sum_{i=0}^{k-\tau-1}(\beta\lambda)^{k-\tau-1-i}\mathbb{1}_{\{s_i=s\}}\right]^2 \\
&\leq \frac{(\beta\lambda)^{2(\tau+1)}}{1-\beta\lambda}(1 + 2\|V\|_2)^2 \sum_{s\in\mathcal{S}}\sum_{i=0}^{k-\tau-1}(\beta\lambda)^{k-\tau-1-i}\mathbb{1}_{\{s_i=s\}} \qquad \text{(Cauchy Schwarz inequality)} \\
&= \frac{(\beta\lambda)^{2(\tau+1)}}{1-\beta\lambda}(1 + 2\|V\|_2)^2 \sum_{i=0}^{k-\tau-1}(\beta\lambda)^{k-\tau-1-i}\sum_{s\in\mathcal{S}}\mathbb{1}_{\{s_i=s\}} \\
&= \frac{(\beta\lambda)^{2(\tau+1)}}{(1-\beta\lambda)^2}(1 + 2\|V\|_2)^2.
\end{aligned}$$

The result follows by taking the square root on both sides of the previous inequality. ■

Lemma E.1 indicates that the error term in Eq. (26) is indeed geometric small. Suppose we ignore that error term, the update equation becomes $V_{k+1} \approx V_k + \epsilon_k(F_k^\tau(V_k, Y_k^\tau) - V_k)$. Since the random process $\mathcal{M}_Y = \{Y_k^\tau\}$ is a Markov chain, once we establish the required properties for the truncated operator $F_k^\tau(\cdot,\cdot)$, our SA results become applicable.

From now on, we will choose $\tau = \min\{k \geq 0 : (\beta\lambda)^{k+1} \leq \epsilon\} \leq \frac{\log(1/\epsilon)}{\log(1/(\beta\lambda))}$, where ϵ is the constant stepsize we use. This implies that the error term in Eq. (26) is of the order $O(\epsilon^2)$. Under this choice of τ, we next investigate the properties of the operator $F_k^\tau(\cdot,\cdot)$ and the random process $\{Y_k^\tau\}$ in the following proposition. Let $\mathcal{K}\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ be a diagonal matrix with diagonal entries $\{\kappa(s)\}_{s\in\mathcal{S}}$, and let $\kappa_{\min}=\min_{s\in\mathcal{S}}\kappa(s)$.

Proposition E.1. *Suppose Assumption D.1 is satisfied. Then we have the following results.*

(1) For any $k\geq\tau$, the operator $F_k^\tau(\cdot,\cdot)$ satisfies $\|F_k^\tau(V_1,y)-F_k^\tau(V_2,y)\|_2\leq\frac{3}{1-\beta\lambda}\|V_1-V_2\|_2$, and $\|F_k^\tau(\mathbf{0},y)\|_2\leq\frac{1}{1-\beta\lambda}$ for any $V_1,V_2\in\mathbb{R}^{|\mathcal{S}|}$ and $y\in\mathcal{Y}_\tau$.

(2) The Markov chain $\{Y_k^\tau\}_{k\geq\tau}$ has a unique stationary distribution, denoted by μ. Moreover, there exists $C_4>0$ and $\sigma_4\in(0,1)$ such that $\max_{y\in\mathcal{Y}_\tau}\|P^{k+\tau+1}(y,\cdot)-\mu(\cdot)\|_{TV}\leq C_4\sigma_4^k$ for all $k\geq 0$.

(3) For any $k\geq\tau$, define the expected operator $\bar{F}_k^\tau:\mathbb{R}^{|\mathcal{S}|}\mapsto\mathbb{R}^{|\mathcal{S}|}$ by $\bar{F}_k^\tau(V)=\mathbb{E}_{Y\sim\mu}[F_k^\tau(V,Y)]$. Then

(a) $\bar{F}_k^\tau$ is a linear operator given by $\bar{F}_k^\tau(V)=\left(I-\mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i(I-\beta P_\pi)\right)V+\mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^iR_\pi$.

(b) $\bar{F}_k^\tau$ is a contraction mapping with respect to $\|\cdot\|_p$ for any $p\in[1,\infty]$, with a common contraction factor $\gamma_4=1-\kappa_{\min}\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}$.

(c) $\bar{F}_k^\tau$ has a unique fixed-point V_π.

Proof of Proposition E.1. (1) For any $V_1,V_2\in\mathbb{R}^{|\mathcal{S}|}$ and $y\in\mathcal{Y}_\tau$, we have

$$
\begin{aligned}
&\|F_k^\tau(V_1,y)-F_k^\tau(V_2,y)\|_2^2\\
=&\sum_{s\in\mathcal{S}}\left[(\beta(V_1(s_{k+1})-V_2(s_{k+1}))-(V_1(s_k)-V_2(s_k)))\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}+V_1(s)-V_2(s)\right]^2\\
=&\sum_{s\in\mathcal{S}}(V_1(s)-V_2(s))^2+\sum_{s\in\mathcal{S}}\left[(\beta(V_1(s_{k+1})-V_2(s_{k+1}))-(V_1(s_k)-V_2(s_k)))\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right]^2\\
&+2\sum_{s\in\mathcal{S}}\left[(\beta(V_1(s_{k+1})-V_2(s_{k+1}))-(V_1(s_k)-V_2(s_k)))\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right](V_1(s)-V_2(s))\\
\leq&\|V_1-V_2\|_2^2+4\|V_1-V_2\|_2^2\sum_{s\in\mathcal{S}}\left[\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right]^2+2\|V_1-V_2\|_2^2\sum_{s\in\mathcal{S}}\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\\
\leq&\|V_1-V_2\|_2^2+4\|V_1-V_2\|_2^2\left(\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\right)\sum_{s\in\mathcal{S}}\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}+2\left(\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\right)\|V_1-V_2\|_2^2\\
=&\|V_1-V_2\|_2^2+4\|V_1-V_2\|_2^2\left(\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\right)^2+2\left(\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\right)\|V_1-V_2\|_2^2\\
=&\|V_1-V_2\|_2^2\left[1+2\left(\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\right)\right]^2\\
\leq&\frac{9}{(1-\beta\lambda)^2}\|V_1-V_2\|_2^2.
\end{aligned}
$$

It follows that $\|F_k^\tau(V_1,y)-F_k^\tau(V_2,y)\|_2\leq\frac{3}{1-\beta\lambda}\|V_1-V_2\|_2$.

Similarly, for any $y\in\mathcal{Y}_\tau$, we have

$$
\begin{aligned}
\|F_k^\tau(\mathbf{0},y)\|_2^2=&\sum_{s\in\mathcal{S}}\left[\mathcal{R}(s_k,a_k)\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right]^2\\
\leq&\sum_{s\in\mathcal{S}}\left[\sum_{i=k-\tau}^{k}(\beta\lambda)^{k-i}\mathbb{1}_{\{s_i=s\}}\right]^2
\end{aligned}
$$

$$\leq \left(\sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i}\right)^2 \qquad \text{(Cauchy Schwarz)}$$
$$\leq \frac{1}{(1-\beta\lambda)^2}.$$

It follows that $\|F_k^\tau(\mathbf{0}, y)\|_2 \leq \frac{1}{1-\beta\lambda}$.

(2) The proof is identical to that of Propositon 3.1 (2).

(3) (a) For any $V \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$, we have

$$\begin{aligned}
[\bar{F}_k^\tau(V)](s) &= \mathbb{E}_{Y\sim\mu}\left[[F_k^\tau(V,Y)](s)\right] \\
&= \mathbb{E}_{Y\sim\mu}\left[(\mathcal{R}(S_k,A_k) + \beta V(S_{k+1}) - V(S_k)) \sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i} \mathbb{1}_{\{S_i=s\}}\right] + V(s) \\
&= \mathbb{E}_{Y\sim\mu}\left[\sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i} \mathbb{1}_{\{S_i=s\}} \mathbb{E}\left[(\mathcal{R}(S_k,A_k) + \beta V(S_{k+1}) - V(S_k)) \mid S_k, S_{k-1}, ..., S_0\right]\right] + V(s) \\
&= \mathbb{E}_{Y\sim\mu}\left[\sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i} \mathbb{1}_{\{S_i=s\}} (R_\pi(S_k) + \beta[P_\pi V](S_k) - V(S_k))\right] + V(s) \\
&= \sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i} \sum_{s_0\in\mathcal{S}} \kappa(s_0) P_\pi^i(s_0,s) \sum_{s'\in\mathcal{S}} P_\pi^{k-i}(s,s')(R_\pi(s') + \beta[P_\pi V](s') - V(s')) + V(s) \\
&= \kappa(s) \sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i} \sum_{s'\in\mathcal{S}} P_\pi^{k-i}(s,s')(R_\pi(s') + \beta[P_\pi V](s') - V(s')) + V(s) \\
&= \kappa(s) \sum_{i=k-\tau}^{k} (\beta\lambda)^{k-i} [P_\pi^{k-i}(R_\pi + \beta P_\pi V - V)](s) + V(s).
\end{aligned}$$

It follows that

$$\begin{aligned}
\bar{F}_k^\tau(V) &= \mathcal{K} \sum_{i=k-\tau}^{k} (\beta\lambda P_\pi)^{k-i}(R_\pi + \beta P_\pi V - V) + V \\
&= \mathcal{K} \sum_{i=0}^{\tau} (\beta\lambda P_\pi)^i (R_\pi + \beta P_\pi V - V) + V \\
&= \left[I - \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i (I - \beta P_\pi)\right] V + \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i R_\pi.
\end{aligned}$$

(b) For any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$ and $p \in [1,\infty]$, we have

$$\begin{aligned}
\|\bar{F}_k^\tau(V_1) - \bar{F}_k^\tau(V_2)\|_p &= \left\|\left[I - \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i (I - \beta P_\pi)\right](V_1 - V_2)\right\|_p \\
&\leq \left\|I - \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i (I - \beta P_\pi)\right\|_p \|V_1 - V_2\|_p.
\end{aligned}$$

Denote $G = I - \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i (I - \beta P_\pi)$. It remains to provide an upper bound on $\|G\|_p$. Since

$$G = I - \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i + \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i \beta P_\pi$$

$$= I - \mathcal{K} - \mathcal{K}\sum_{i=1}^{\tau}(\beta\lambda P_\pi)^i + \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i \beta P_\pi$$
$$= I - \mathcal{K} - \mathcal{K}\sum_{i=0}^{\tau-1}(\beta\lambda P_\pi)^{i+1} + \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i \beta P_\pi$$
$$= I - \mathcal{K} + \mathcal{K}\sum_{i=0}^{\tau-1}(\beta\lambda P_\pi)^i \beta P_\pi(1-\lambda) + \mathcal{K}(\beta\lambda P_\pi)^\tau \beta P_\pi,$$

the matrix $G_{\lambda,\tau}$ has non-negative entries. Therefore, we have

$$\|G_{\lambda,\tau}\|_\infty = \|G_{\lambda,\tau}\mathbf{1}\|_\infty = \left\|\mathbf{1} - \kappa\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}\right\|_\infty = 1 - \kappa_{\min}\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}$$

and

$$\|G_{\lambda,\tau}\|_1 = \|\mathbf{1}^\top G_{\lambda,\tau}\|_\infty = \left\|\mathbf{1}^\top - \kappa^\top\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}\right\|_\infty = 1 - \kappa_{\min}\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}.$$

It then follows from Lemma D.1 that

$$\|G_{\lambda,\tau}\|_p \le \|G_{\lambda,\tau}\|_1^{1/p}\|G_{\lambda,\tau}\|_\infty^{1-1/p} \le 1 - \kappa_{\min}\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}.$$

Hence the operator $F_k^\tau(\cdot,\cdot)$ is a contraction with respect to $\|\cdot\|_p$, with a common contraction factor $\gamma_4 = 1 - \kappa_{\min}\frac{(1-\beta)(1-(\beta\lambda)^{\tau+1})}{1-\beta\lambda}$.

(c) It is enough to show that V_π is a fixed-point of $\bar{F}_k^\tau(\cdot)$, the uniqueness follows from $\bar{F}_k^\tau(\cdot)$ being a contraction. Using the Bellman's equation $R_\pi + \beta P_\pi V_\pi - V_\pi = 0$, we have

$$\bar{F}_k^\tau(V_\pi) = \mathcal{K}\sum_{i=0}^{\tau}(\beta\lambda P_\pi)^i(R_\pi + \beta P_\pi V_\pi - V_\pi) + V_\pi = V_\pi.$$

■

E.2 Proof of Theorem 3.3

We will exploit the $\|\cdot\|_2$-contraction property of the operator $\bar{F}_k^\tau(\cdot)$ provided in Proposition E.1. Let $M(x) = \|x\|_2^2$ be our Lyapunov function. Using the update equation (26), and we have for all $k \ge 0$:

$$\begin{aligned}
&\|V_{k+1} - V_\pi\|_2^2 \\
=& \|V_k - V_\pi\|_2^2 + \underbrace{2\epsilon(V_k - V_\pi)^\top\left(\bar{F}_k^\tau(V_k) - V_k\right)}_{①} + \underbrace{2\epsilon(V_k - V_\pi)^\top\left(F_k^\tau(V_k, Y) - \bar{F}_k^\tau(V_k)\right)}_{②} \\
&+ \underbrace{\epsilon^2\|F_k^\tau(V_k, Y) - V_k\|_2^2}_{③} + \underbrace{\epsilon^2\|F_k(V_k, Y_k) - F_k^\tau(V_k, Y)\|_2^2}_{④} \\
&+ \underbrace{2\epsilon(V_k - V_\pi)^\top\left(F_k(V_k, Y_k) - F_k^\tau(V_k, Y)\right)}_{⑤} + \underbrace{2\epsilon\left(F_k^\tau(V_k, Y) - V_k\right)^\top\left(F_k(V_k, Y_k) - F_k^\tau(V_k, Y)\right)}_{⑥}.
\end{aligned} \tag{27}$$

The terms ①, ②, and ③ correspond to the terms T_1, T_3, and T_4 in Eq. (12), and hence can be controlled in the same way as provded in Lemmas A.1, A.4, and A.5. The proof is omitted. As for the terms ③, ④, and ⑤, we can easily use Lemma E.2 along with the Cauchy Schwarz inequality to bound them, which gives the following result.

Lemma E.2. *The following inequalities hold:*

(1) ④ $\leq \frac{8\epsilon^2}{(1-\beta\lambda)^2}\|V_k - V_\pi\|_2^2 + \frac{2\epsilon^2}{(1-\beta\lambda)^2}(4\|V_\pi\|_2+1)^2$ *for all* $k \geq \tau$.

(2) ⑤ $\leq \frac{16\epsilon^2}{(1-\beta\lambda)}\|V_k - V_\pi\|_2^2 + \frac{4\epsilon^2}{(1-\beta\lambda)}(4\|V_\pi\|_2+1)^2$ *for all* $k \geq \tau$.

(3) ⑥ $\leq \frac{64\epsilon^2}{(1-\beta\lambda)^2}\|V_k - V_\pi\|_2^2 + \frac{4\epsilon^2}{(1-\beta\lambda)^2}(4\|V_\pi\|_2+1)^2$ *for all* $k \geq \tau$.

Proof of Lemma E.2. (1) For all $k \geq \tau$, we have

$$\begin{aligned}
④ &= \epsilon^2 \|F_k(V_k, Y_k) - F_k^\tau(V_k, Y_k^\tau)\|_2^2 \\
&\leq \frac{\epsilon^2(\beta\lambda)^{2(\tau+1)}}{(1-\beta\lambda)^2}(2\|V_k\|_2+1)^2 && \text{(Lemma E.1)} \\
&\leq \frac{\epsilon^4}{(1-\beta\lambda)^2}(2\|V_k - V_\pi\|_2 + 2\|V_\pi\|_2 + 1)^2 \\
&\leq \frac{8\epsilon^2}{(1-\beta\lambda)^2}\|V_k - V_\pi\|_2^2 + \frac{2\epsilon^2}{(1-\beta\lambda)^2}(4\|V_\pi\|_2+1)^2.
\end{aligned}$$

(2) For all $k \geq \tau$, we have

$$\begin{aligned}
⑤ &= 2\epsilon(V_k - V_\pi)^\top \left(F_k(V_k, Y_k) - F_k^\tau(V_k, Y_k^\tau)\right) \\
&\leq 2\epsilon\|V_k - V_\pi\|_2 \|F_k(V_k, Y_k) - F_k^\tau(V_k, Y_k^\tau)\|_2 \\
&\leq \frac{2\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)}\|V_k - V_\pi\|_2(2\|V_k\|_2+1) && \text{(Proposition E.1 (1))} \\
&\leq \frac{2\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)}(2\|V_k - V_\pi\|_2 + 2\|V_\pi\|_2+1)^2 \\
&\leq \frac{16\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)}\|V_k - V_\pi\|_2^2 + \frac{4\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)}(4\|V_\pi\|_2+1)^2 \\
&\leq \frac{16\epsilon^2}{(1-\beta\lambda)}\|V_k - V_\pi\|_2^2 + \frac{4\epsilon^2}{(1-\beta\lambda)}(4\|V_\pi\|_2+1)^2, . && \text{(The choice of } \tau)
\end{aligned}$$

(3) For all $k \geq \tau$, we have

$$\begin{aligned}
⑥ &= 2\epsilon\left(F_k^\tau(V_k, Y_k^\tau) - V_k\right)^\top \left(F_k(V_k, Y_k) - F_k^\tau(V_k, Y_k^\tau)\right) \\
&\leq 2\epsilon\|F_k^\tau(V_k, Y_k^\tau) - V_k\|_2 \|F_k(V_k, Y_k) - F_k^\tau(V_k, Y_k^\tau)\|_2 \\
&\leq \frac{2\epsilon(\beta\lambda)^{\tau+1}}{1-\beta\lambda}\left(\frac{3}{1-\beta\lambda}\|V_k\|_2 + \frac{1}{1-\beta\lambda} + \|V_k\|_2\right)(2\|V_k\|_2+1) \\
&\leq \frac{2\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)^2}(4\|V_k\|_2+1)(2\|V_k\|_2+1) \\
&\leq \frac{2\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)^2}(4\|V_k - V_\pi\|_2 + 4\|V_\pi\|_2+1)^2 \\
&\leq \frac{64\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)^2}\|V_k - V_\pi\|_2^2 + \frac{4\epsilon(\beta\lambda)^{\tau+1}}{(1-\beta\lambda)^2}(4\|V_\pi\|_2+1)^2 \\
&\leq \frac{64\epsilon^2}{(1-\beta\lambda)^2}\|V_k - V_\pi\|_2^2 + \frac{4\epsilon^2}{(1-\beta\lambda)^2}(4\|V_\pi\|_2+1)^2. && \text{(The choice of } \tau)
\end{aligned}$$

■

Substitute the upper bounds we have for the terms ① to ⑥ into Eq. (27), and we have the one-step contractive inequality for the TD(λ) algorithm. Repeatedly using that inequality and we obtain the desired finite-sample convergence bounds.