Parameter-free Optimal Rates for Nonlinear Semi-Norm Contractions with Applications to Q-Learning

Ankur Naskar[1], **Gugan Thoppe**[1], **Vijay Gupta**[2]

[1]Computer Science and Automation, Indian Institute of Science, Bengaluru
[2]Electrical and Computer Engineering, Purdue University, IN, USA
ankurnaskar@iisc.ac.in, gthoppe@iisc.ac.in, gupta869@purdue.edu

Abstract

Algorithms for solving *nonlinear* fixed-point equations—such as average-reward *Q-learning* and *TD-learning*—often involve semi-norm contractions. Achieving parameter-free optimal convergence rates for these methods via Polyak–Ruppert averaging has remained elusive, largely due to the non-monotonicity of such semi-norms. We close this gap by (i.) recasting the averaged error as a linear recursion involving a nonlinear perturbation, and (ii.) taming the nonlinearity by coupling the semi-norm's contraction with the monotonicity of a suitably induced norm. Our main result yields the first parameter-free $\tilde{O}(1/\sqrt{t})$ optimal rates for Q-learning in both average-reward and exponentially discounted settings, where t denotes the iteration index. The result applies within a broad framework that accommodates synchronous and asynchronous updates, single-agent and distributed deployments, and data streams obtained either from simulators or along Markovian trajectories.

1 Introduction

Stochastic fixed-point iterations with contractive operators (Borkar 2009) permeate many fields. In Reinforcement Learning (RL), Temporal Difference (TD)[1] learning and Q-learning are canonical examples (Bertsekas and Tsitsiklis 1996; Sutton and Barto 2018; Szepesvári 2022). Game-theoretic approaches to decentralized learning and Nash-equilibrium computation adopt the same template (Sayin et al. 2021; Hu and Wellman 2003). In optimization, stochastic fixed-point iterations underpin methods for regularized least squares in over-parameterized models and for low-rank matrix completion (Li, Ma, and Zhang 2018; Marjanovic and Solo 2012). They are likewise of growing importance in nonlinear dynamical systems, where fixed-point theory has become an important tool for analysis and control design, particularly in robotics (Lohmiller and Slotine 1997; Manchester and Slotine 2014; Manchester, Tang, and Slotine 2017; Tsukamoto, Chung, and Slotine 2021).

Here, we focus on stochastic fixed-point iterations driven by *semi-norm* contractions. Their generic update rule is

$$Q_{t+1} = Q_t + \alpha_t \left[h(Q_t, y_t) - Q_t\right], \tag{1}$$

where (y_t) is an ergodic Markov chain on a state space $\mathcal{Y}$ with a unique stationary distribution η and $h : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ is such that the expected map $H(Q) = \mathbb{E}_{y\sim\eta} h(Q, y)$ is a semi-norm contraction. In other words, we have a semi-norm $\upsilon : \mathbb{R}^d \to \mathbb{R}$ and a factor $\beta \in [0, 1)$ so that

$$\upsilon(H(x) - H(y)) \leq \beta\, \upsilon(x - y) \quad \forall x, y \in \mathbb{R}^d. \tag{2}$$

A semi-norm is similar to a norm that need not satisfy the positive definiteness condition. Specifically, for all $x, y \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$, a semi-norm υ satisfies non-negativity ($\upsilon(x) \geq 0$), absolute homogeneity ($\upsilon(\lambda x) = |\lambda|\upsilon(x)$), and sub-additivity ($\upsilon(x+y) \leq \upsilon(x)+\upsilon(y)$), but not necessarily positive definiteness ($\upsilon(x) = 0 \not\Rightarrow x = 0$). As a result, every norm is a semi-norm, but the converse does not hold.

The role of semi-norm contractions in RL methods such as TD and Q-learning is fundamental. In these methods, the Bellman operator is contractive—under a semi-norm in the average-reward formulation and under a true norm (hence, also a semi-norm) in exponential discounting (Bertsekas and Tsitsiklis 1996). The generic recursion in (1) subsumes all variants of TD methods and Q-learning—including the ones with synchronous and asynchronous updates, single-agent and distributed deployments, and data drawn from generative models and Markovian trajectories.

Achieving optimal convergence rates in such algorithms is of fundamental interest. Several works have obtained such bounds for TD algorithms and Q-learning in their various forms. TD methods, for instance, have been looked at in (Dalal et al. 2018; Lakshminarayanan and Szepesvari 2018; Bhandari, Russo, and Singal 2018; Zhang, Zhang, and Maguluri 2021; Liu and Olshevsky 2023; Khodadadian et al. 2022; Dal Fabbro, Mitra, and Pappas 2023) and (Wang et al. 2024). On the other hand, tabular Q-learning has been the focus of (Even-Dar and Mansour 2003; Chen et al. 2021a; Zhang, Zhang, and Maguluri 2021) and (Chen et al. 2025). Table 1 summarizes the key differences in these results.

However, a key drawback of these results is that the optimal expected error bound, $\tilde{O}(1/\sqrt{t})$, is achieved only with a stepsize $\alpha_t = c/t$, where the constant c depends on unknown transition probabilities. Such a c is rarely available, making these rates effectively impractical.

For *linear* stochastic approximation, (Polyak and Juditsky 1992) and (Ruppert 1991) proposed a remarkable solution—now called Polyak–Ruppert averaging. Applied to an update

[1]Although Q-learning falls within the broader family of TD methods, we use TD exclusively for policy-evaluation algorithms.

	Reference	Distributed	Discounting	Asynchronous	Optimal rate	Universal stepsize
TD Learning	Dalal et al. (2018)	✗	Exp	✓	✗	✓
	Patil et al. (2023)	✗	Exp	✓	✓	✓
	Lakshminarayanan and Szepesvari (2018)	✗	Exp and Avg	✓	✓	✗
	Bhandari, Russo, and Singal (2018)	✗	Exp	✓	✓	✗
	Chen et al. (2021b)	✗	Exp	✓	✓	✗
	Durmus et al. (2025)	✗	Exp	✓	✓	✗
	Bhandari, Russo, and Singal (2021)	✗	Exp	✓	✓	✗
	Chen et al. (2025)	✗	Exp and Avg	✓	✓	✗
	Dal Fabbro, Mitra, and Pappas (2023)	✓	Exp	✓	✓	✗
	Khodadadian et al. (2022)	✓	Exp	✓	✓	✗
	Liu and Olshevsky (2023)	✓	Exp	✓	✓	✗
	Wang et al. (2024)	✓	Exp	✓	✓	✗
	Naskar et al. (2024)	✓	Exp and Avg	✓	✓	✓
Q-Learning	Even-Dar and Mansour (2003)	✗	Exp	✗	✗	✓
	Wainwright (2019)	✗	Exp	✗	✓	✗
	Zhang, Zhang, and Maguluri (2021)	✗	Avg	✗	✓	✗
	(Chen et al. 2021b)	✗	Exp	✓	✓	✗
	Li et al. (2023)	✗	Exp	✗	✓	✓
	Our Work	✓	Avg	✗	✓	✓
		✓	Exp	✓	✓	✓

Table 1: Comparison of our work with the existing literature on TD-learning and Q-learning algorithms. In the column labeled discounting, Exp refers to exponential, while Avg refers to average reward.

rule like (1), this procedure works in two steps:

1. **Generate base iterates**: Run (Q_t) with the parameter-free stepsize $\alpha_t = 1/(t+1)^\alpha$ for some $\alpha \in (1/2, 1)$.
2. **Average them**: Form the running mean $\bar{Q}_T = \frac{1}{T}\sum_{t=0}^{T-1} Q_t$, again without any problem-specific tuning.

While this procedure yields a suboptimal convergence rate of $\tilde{O}(1/t^{\alpha/2})$ for the (Q_t) sequence, the averaged sequence $(\bar{Q}_t)$ attains the desired optimal rate of $\tilde{O}(1/\sqrt{t})$.

For TD-learning with linear function approximation—a special case of linear stochastic approximation—Polyak–Ruppert averaging already achieves parameter-free optimal rates, both for exponential discounting (Patil et al. 2023) and for average reward (Naskar et al. 2024).

Q-learning with Polyak–Ruppert averaging is challenging due to the algorithm's inherent nonlinearity. A recent breakthrough by Li et al. (2023) addresses this challenge for *synchronous Q-learning* with *exponential discounting*. In this setting, the Bellman operator is a contraction in the monotone $\|\cdot\|_\infty$ norm. The authors' key idea is to construct two auxiliary sequences, (L_t) and (U_t), whose Polyak–Ruppert averages bound the error in $(\bar{Q}_t)$ from below and above, respectively. Each auxiliary sequence follows the template of a linear stochastic approximation with a rapidly vanishing nonlinear remainder. Classical Polyak–Ruppert analysis then yields a $\tilde{O}(1/\sqrt{t})$ rate for both. By the monotonicity of $\|\cdot\|_\infty$, these bounds transfer directly to $\bar{Q}_t$.

For synchronous Q-learning with exponential discounting, the raw iterates (Q_t) already achieve the optimal rate for the stepsize $\alpha_t = \frac{1}{1+(1-\gamma)t}$, where γ is the (known) discount factor (Wainwright 2019, Corollary 3). Polyak-Ruppert averaging, therefore, is unnecessary in this setting if the goal is only to get parameter-free optimal rates.

For asynchronous Q-learning in exponential discounting—and for both synchronous and asynchronous versions under average reward—the contraction factor depends on unknown transition probabilities. Hence, parameter-free optimal rates in these cases remain open.

While the ideas of Li et al. (2023) extend to asynchronous Q-learning in exponential discounting (with some effort), they break down in the average-reward case. The challenge stems from the non-monotonicity of the span semi-norm

$$\|x\|_{\text{sp}} := \max_i x(i) - \min_i x(i), \tag{3}$$

under which the average-reward Bellman operator is contractive. Specifically, $0 \le x \le y$ does not imply $\|x\|_{\text{sp}} \le \|y\|_{\text{sp}}$. As a result, although auxiliary lower- and upper-bounding sequences can still be constructed and shown to achieve parameter-free optimal rates, $\|\cdot\|_{\text{sp}}$'s non-monotonicity prevents these rates from transferring to $(\bar{Q}_t)$.

Our work's **key highlights** are as follows:

1. We are the first to show that Polyak–Ruppert averaging achieves **parameter-free optimal rates** even for semi-norm contractive nonlinear fixed-point iterations.
2. Our result **applies directly to Q-learning in both the average-reward and exponentially discounted settings**, providing the first parameter-free optimal expected error rates of $\tilde{O}(1/\sqrt{t})$.
3. Our proof is novel and may be of independent interest. Specifically, our proof proceeds in two main steps: (i) decomposing the error in $\bar{Q}_t$ into a linear recursion with a nonlinear perturbation (following (Li et al. 2023)); and

(ii) showing rapid decay of the nonlinear term by exploiting the **contraction of the semi-norm** together with the **equivalence between its induced norm and a suitably chosen monotone norm**.

2 Problem Setup and Our Goal

Although our contributions are novel even in single-agent settings, for the sake of generality, we pose our fixed-point problem directly in a distributed architecture consisting of a central server and N agents. Each agent $i \in [N] := \{1, \dots, N\}$ has noisy access to an operator $H_i : \mathbb{R}^d \to \mathbb{R}^d$ that is a contraction with respect to a common semi-norm $\upsilon : \mathbb{R}^d \to \mathbb{R}$ with contraction factor β_i; see (2). Specifically, for any time instance $t \in \{0, 1, \dots\}$ and any query $Q \in \mathbb{R}^d$, agent i can compute $h_i(Q, y_i^t)$, where (y_i^t) is an ergodic Markov chain on a state space $\mathcal{Y}$ with a unique stationary distribution η_i, and the local function $h_i : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ satisfies

$$H_i(Q) = \mathbb{E}_{y \sim \eta_i} h_i(Q, y). \tag{4}$$

We assume that the Markov chains are independent across agents. The aim of this setup is to find a fixed point $Q^* \in \mathbb{R}^d$ of the averaged map $H : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$H(Q) = \frac{1}{N} \sum_i H_i(Q). \tag{5}$$

That is, a vector Q^* such that

$$H(Q^*) - Q^* \in E, \tag{6}$$

where

$$E = \{x \in \mathbb{R}^d : \upsilon(x) = 0\} \tag{7}$$

is the linear subspace where υ vanishes. Such a fixed point exists (and is unique modulo E) since the averaged operator $H : \mathbb{R}^d \to \mathbb{R}^d$ is itself a semi-norm contraction with a contraction factor

$$\beta := \max\{\beta_1, \dots, \beta_N\}. \tag{8}$$

A common way to obtain such a Q^* is for the agents to execute the distributed stochastic fixed-point iteration

$$Q_{t+1} = Q_t + \alpha_t \left[\frac{1}{N} \sum_{i=1}^{N} h_i(Q_t, y_t^i) - Q_t \right], \quad t \geq 0. \tag{9}$$

Our goal then is to determine whether the optimal $\tilde{O}(1/\sqrt{t})$ rate, with problem-independent stepsizes, holds for Polyak-Ruppert average sequence $(\bar{Q}_t)$ given by

$$\bar{Q}_{t+1} = \bar{Q}_t + \frac{1}{t+1} \left[Q_t - \bar{Q}_t \right], \quad t \geq 0. \tag{10}$$

Since our framework allows for semi-norm contractions, a positive result would instantly provide parameter-free optimal rates to all avatars of TD learning and Q-learning.

3 Main Result and RL Applications

In Section 3.1, we state our assumptions and present our main convergence-rate theorem for the Polyak–Ruppert average (10) of the stochastic fixed-point iteration in (9). Subsequently, in Section 3.2, as an illustration, we show how this result applies to Q-learning under both average reward and exponential discounting, yielding the first parameter-free optimal convergence rates for these methods.

3.1 Parameter-Free Optimal Convergence Rates for the Stochastic Fixed-Point Iteration (9)

We begin with the following fact about semi-norms.

Lemma 3.1. (Chen et al. 2025, Proposition 2.1) *For any semi-norm $\upsilon : \mathbb{R}^d \to \mathbb{R}$, one can define an induced norm $\|\cdot\|$ such that, for all $Q \in \mathbb{R}^d$,*

$$\upsilon(Q) = \min_{e \in E} \|Q - e\|, \tag{11}$$

where E is as defined in (7)*.*

Since all norms are equivalent in finite dimensions, given any monotone norm[2] $\|\cdot\|_m$, there exist $c_\ell, c_u > 0$ such that, for all Q,

$$c_\ell \|Q\|_m \leq \|Q\| \leq c_u \|Q\|_m. \tag{12}$$

Next, we state our assumptions. For all $t \geq 0$, let $(\mathcal{F}_t)$ be the filtration sequence of σ-fields defined by

$$\mathcal{F}_t := \sigma\left(\{Q_0\} \cup \{y_k^i : i \in [N], k < t\}\right). \tag{13}$$

$\mathcal{A}_1$ **Local update functions:** For each agent $i \in [N]$, there exist $b_i : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^d$ and $A_i : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}^{d \times d}$ and constants $C_{\mathrm{A}}, C_{\mathrm{b}}, C_* > 0$ such that $h_i(Q, y) = b_i(Q, y) + A_i(Q, y)Q$ and the following holds:

(a) $A_i(Q, y)e = e$ for all $e \in E$ (see (7)).

(b) $A_i(Q, y)Q \geq A_i(Q', y)Q$ for all $Q, Q' \in \mathbb{R}^d$, where the inequality is in a coorindate sense.

(c) $\|A_i(Q, y)\| \leq C_{\mathrm{A}}$ and $\|b_i(Q, y)\| \leq C_{\mathrm{b}}$, where $\|\cdot\|$ is the induced norm (see (11)).

(d) For any semi-norm fixed point Q^* satisfying (6), $A_i(Q, y) = A_i(Q^*, y)$ and $b_i(Q, y) = b_i(Q^*, y)$ if $\upsilon(Q - Q^*) < C_*$.

$\mathcal{A}_2$ **Noise:** There exist constants $C_{\mathrm{E}} > 0$ and $\rho \in (0, 1)$ such that, for agent $i \in [N]$ and time $\tau \in \{0, 1, \cdots, t-1\}$, the Markov chain (y_t^i) satisfies

$$\|\mathbb{P}(y_t^i = \cdot | \mathcal{F}_{t-\tau}) - \eta_i\|_{\mathrm{TV}} \leq C_{\mathrm{E}}\, \rho^\tau,$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance.

$\mathcal{A}_3$ **Raw iterate convergence:** There exists a constant $C_Q > 0$ such that, for all $t \geq t_*$,

$$\mathbb{E}\upsilon(Q_t - Q^*)^2 \leq C_Q\, \tau_t\, \alpha_t,$$

where $\tau_t := \min\{\tau > 0 : \rho^\tau \leq \alpha_t^2\}$ and $t_* := \min\{t > 0 : t \geq 2\tau_t\}$.

Remark 3.2. *Assumption $\mathcal{A}_1$ holds for any piece-wise linear update rule, including TD-learning and Q-learning in both exponential discounting and average-reward settings. At this point, it is unclear if $\mathcal{A}_1$ will hold beyond these cases.*

Remark 3.3. *Assumption $\mathcal{A}_2$ is called the geometric mixing time assumption and is standard for algorithms using Markovian sampling (Durmus et al. 2025; Zhang, Zhang, and Maguluri 2021; Chen et al. 2025).*

Remark 3.4. *The condition in Assumption $\mathcal{A}_3$ has been derived for several variants of TD and Q-learning algorithms; see the references in Table 1 for details.*

[2] For all $x, y \in \mathbb{R}^d$, if $0 \leq x \leq y$ then $\|x\|_m \leq \|y\|_m$.

Our main result can now be stated as follows.

Theorem 3.5. *Consider the stochastic fixed-point iteration* (9) *under Assumptions $\mathcal{A}_1$–$\mathcal{A}_3$. Choose the stepsize $\alpha_t = \frac{1}{(t+1)^\alpha}$ for some $\alpha \in (1/2, 1)$. Let τ_t and t_* be as in Assumption $\mathcal{A}_3$, and let ρ be as in Assumption $\mathcal{A}_2$. Then, for $T > t_*$, the Polyak-Ruppert averages given in* (10) *satisfy*

$$\mathbb{E}\upsilon(\bar{Q}_T - Q^*) \leq \frac{C^{(1)}\sqrt{\tau_T}}{\sqrt{NT}} + \frac{C^{(2)}\ln(T)}{T^\alpha},$$

where constants $C^{(1)}$ and $C^{(2)}$ are as in [Table 2] (Naskar, Thoppe, and Gupta 2025).

Remark 3.6 (Optimal Convergence Rate)**.** *The factor τ_T grows only logarithmically with T, i.e., $\tau_T = O(\ln T)$. Separately, since $\alpha > \frac{1}{2}$, the second term is $o(1/T)$ and thus asymptotically negligible. Hence, $\mathbb{E}\upsilon(\bar{Q}_T - Q^*) = \tilde{O}(1/\sqrt{T})$, which implies that this result gives the optimal expected error rate up to logarithmic factors.*

Remark 3.7 (Linear Speedup)**.** *The leading term in our bound has a factor of $1/\sqrt{N}$, implying that the iteration complexity improves linearly with the number N of agents.*

Remark 3.8 (Parameter-Free Stepsize)**.** *The exponent $\alpha \in (1/2, 1)$ in the stepsize $\alpha_t = 1/(t+1)^\alpha$ can be fixed universally; it requires no knowledge of problem-specific parameters such as the contraction coefficient or the mixing-time.*

3.2 Reinforcement Learning Applications

We next use Theorem 3.5 to obtain parameter-free optimal convergence rates for Q-learning. Our analysis here targets two specific settings in which such rates are still lacking: the synchronous variant with average reward and the asynchronous version with exponential discounting (empirical verification with synthetic problem instances can be found in [Appendix](Naskar, Thoppe, and Gupta 2025)).

For any set U, denote by $\Delta(U)$ the collection or set of probability distributions on U.

Synchronous Average-reward Q-learning: We first show how our setup from Section 2 specializes for this algorithm. Each agent i controls a Markov Decision Process (MDP) $\mathcal{M}_i := (\mathcal{S}, \mathcal{A}, \mathcal{P}_i, \mathcal{R}_i)$, where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are common to all the agents, while $\mathcal{P}_i : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $\mathcal{R}_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are agent-specific transition and reward functions, respectively. For a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the average reward of agent i is

$$r_i^\pi := \liminf_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \mathcal{R}_i(s_t, a_t)\right], \tag{14}$$

where a_t and s_{t+1} are sampled from $\pi(\cdot|s_t)$ and $\mathcal{P}_i(\cdot|s_t, a_t)$, respectively. Agent i's *optimal policy* is

$$\pi_i^* := \arg\max_\pi (r_i^\pi). \tag{15}$$

In average reward, one computes the optimal *differential Q-value function* Q_i^* to find the optimal policy; Q_i^* is defined as follows. Let $\mathcal{T}_i^J : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ satisfying

$$\mathcal{T}_i^J Q := \mathcal{R}_i + \sum_{k=1}^{J-1} (\mathcal{P}_i^{\pi_Q})^k \mathcal{R}_i + (\mathcal{P}_i^{\pi_Q})^J Q \tag{16}$$

denote agent i's *differential J-step Bellman operator*, where $\pi_Q(s) = \arg\max_a Q(s, a)$ and, for $s, s' \in \mathcal{S}$ and $a, a' \in \mathcal{A}$,

$$\mathcal{P}_i^{\pi_Q}(s', a'|s, a) = \begin{cases} \mathcal{P}_i(s'|s, a) & \text{if } a' \in \pi_Q(s'), \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Choosing $J \geq |\mathcal{S}|$ guarantees that $\mathcal{T}_i^J$ is a contraction in the span semi-norm $\|\cdot\|_{\text{sp}}$ (see the discussion below (Zhang, Zhang, and Maguluri 2021, Assumption 3)). Now, for $\upsilon = \|\cdot\|_{\text{sp}}$, the subspace E from (7) equals $\{c\mathbb{1} : c \in \mathbb{R}\}$, where $\mathbb{1}$ is the all ones vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Therefore, there exists a vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ satisfying

$$\mathcal{T}_i^J Q - Q \in \{c\mathbb{1} : c \in \mathbb{R}\}. \tag{18}$$

We denote such a vector by Q_i^*. The optimal policy π_i^* is greedy with respect to Q_i^*, i.e., $\pi_i^* = \pi_{Q_i^*}$.

As shown above (8), the averaged map $\mathcal{T}^J := \frac{1}{N}\sum_{i=1}^N \mathcal{T}_i^J$ is also a semi-norm contraction under $\|\cdot\|_{\text{sp}}$. Hence, there exists a $Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ such that

$$\mathcal{T}^J Q^* - Q^* \in \{c\mathbb{1} : c \in \mathbb{R}\}. \tag{19}$$

To compute such a Q^*, distributed synchronous J-step average-reward Q-learning can be used, and it is a natural extension of the single-agent variant in (Zhang, Zhang, and Maguluri 2021, Algorithm 2). Its update rule is a special case of (9). We now specify the local operator h_i and the local Markov chain (y_t^i)—because of synchronous updates, the latter reduces to IID samples. Specifically, set $d = |\mathcal{S}||\mathcal{A}|$ and $\mathcal{Y} = (\mathcal{S}^J)^{\mathcal{S}\times\mathcal{A}}$; every $y \in \mathcal{Y}$ is a map that assigns a J-step sequence of states to every state-action pair. Then, (y_t^i) is generated as follows: at every iteration $t \geq 0$, and for every state-action pair (s, a), $y_t^i(s, a)$ is a sequence of states $(s_{t,1}^i, \ldots, s_{t,J}^i)$, where $s_{t,1}^i \sim \mathcal{P}_i(\cdot|s, a)$, and $s_{t,k+1}^i \sim \mathcal{P}(\cdot|s_{t,k}^i, \pi_{Q_t}(s_{t,k}^i))$, $k \geq 1$. On the other hand, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $y(s, a) = (s_1, \ldots, s_J)$, we have

$$h_i(Q, y)(s, a) := \mathcal{R}_i(s, a) + \sum_{k=1}^{J-1} \mathcal{R}_i(s_k, \pi_Q(s_k)) + \max_{a'} Q(s_J, a'). \tag{20}$$

Our next main result obtains parameter-free optimal convergence rates for the Polyak-Ruppert average $(\bar{Q}_t)$ of the synchronous average-reward Q-learning.

Theorem 3.9 (Synchronous Average-Reward Q-learning)**.** *For each $i \in [N]$, the local function h_i and the Markov chain $\{(y_t^i) : i \in [N]\}$ satisfy the conditions $\mathcal{A}_1$ and $\mathcal{A}_2$. Moreover, the raw iterates satisfy $\mathcal{A}_3$. Hence, the conclusion of Theorem 3.5 holds, i.e., $\mathbb{E}\|\bar{Q}_T - Q^*\|_{\text{sp}} = \tilde{O}(1/\sqrt{NT})$.*

Remark 3.10. *(Zhang, Zhang, and Maguluri 2021) derive optimal convergence rates for the synchronous average-reward Q-learning in the single-agent case. However, the stepsize α_t there depends on $\mathcal{T}_i^J$'s contraction factor, which is unknown as it depends on $\mathcal{P}_i$, $i \in [N]$. In contrast, we obtain optimal rates with parameter-free stepsizes.*

Remark 3.11. *Let $\varepsilon_p, \varepsilon_r > 0$ be such that, for every pair of agents i, j and every state-action pair (s, a),*

$$\|\mathcal{P}_i(\cdot|s,a) - \mathcal{P}_j(\cdot|s,a)\|_{\text{TV}} \le \varepsilon_p \|\mathcal{P}_i(\cdot|s,a)\|_{\text{TV}}$$

and $\|\mathcal{R}_i - \mathcal{R}_j\|_\infty \le \varepsilon_r$. Then, we can show that for each agent i, $\|Q^* - Q_i^*\| \le H_{\text{Avg}}(\varepsilon_p, \varepsilon_r)$, implying that $(\bar{Q}_T)$ converges to Q_i^* modulo the heterogeneity gap, i.e., $H_{\text{Avg}}(\varepsilon_p, \varepsilon_r)$. This gap goes to 0 as $\varepsilon_p + \varepsilon_r \to 0$.

Asynchronous Exponentially-Discounted Q-Learning: The setup here is similar to one in the average-reward case with an additional parameter $\gamma \in (0,1)$, called the discount factor. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, agent i's Q-value function $Q_i^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is now defined as

$$Q_i^\pi(s,a) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}_i(s_t, a_t) \middle| s_0 = s, a_0 = a \right], \quad (21)$$

where, for $t \ge 1$, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim \mathcal{P}_i(\cdot|s_t, a_t)$. Also, agent i's optimal policy π_i^* is one that satisfies

$$Q_i^{\pi_i^*}(s,a) \ge Q_i^\pi(s,a), \quad \forall \pi \text{ and } \forall s, a.$$

Each agent i computes π_i^* by estimating $Q_i^{\pi_i^*}$, which is a fixed-point of the Bellman operator $\mathcal{T}_i$ given by $\mathcal{T}_i Q = \mathcal{R}_i + \gamma \mathcal{P}_i^{\pi_Q} Q$, where the greedy policy π_Q and the transition matrix $\mathcal{P}_i^{\pi_Q}$ have the same meaning as in the average-reward case. For each agent i, $\mathcal{T}_i$ is a γ-contraction in $\|\cdot\|_\infty$ and has the unique fixed point $Q_i^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$; importantly, $\pi_i^* = \pi_{Q_i^*}$. As in (8), the averaged map $\mathcal{T} := \frac{1}{N}\sum_i \mathcal{T}_i$ also is a γ-contraction and has a unique fixed point $Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

One way to find this Q^* is to use distributed asynchronous exponentially-discounted Q-learning, which we describe next. Let μ be a fixed behavior policy. Then, the above algorithm is obtained from (9) by taking $d = |\mathcal{S}||\mathcal{A}|$, $\mathcal{Y} = \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, and letting

$$h_i(Q, s, a, s') = Q - Q(s,a)\mathbb{1}_{(s,a)} + [\mathcal{R}_i(s,a) + \gamma \max_{a'} Q(s', a')]\mathbb{1}_{(s,a)}$$

and $y_t^i = (s_t^i, a_t^i, s_{t+1}^i)$, where for each agent i and each $t \ge 0$, $a_t^i \sim \mu(\cdot|s_t^i)$ and $s_{t+1}^i \sim \mathcal{P}_i(\cdot|s_t^i, a_t^i)$.

Our next main result establishes parameter-free rates for the Polyak-Ruppert average $(\bar{Q}_t)$ of the exponentially discounted Q-learning algorithm described above.

Theorem 3.12 (Asynchronous Exponentially-Discounted Q-learning)**.** *The assumptions $\mathcal{A}_1$–$\mathcal{A}_3$ hold. Consequently, $\mathbb{E}\|\bar{Q}_T - Q^*\|_\infty = \tilde{O}(1/\sqrt{NT})$.*

4 Proof Outline

In this section, we sketch our proofs for Theorems 3.5, 3.9, and 3.12. The detailed proofs can be found in the appendix of the extended version (Naskar, Thoppe, and Gupta 2025).

4.1 Proof (Sketch) of Theorem 3.5

Our proof proceeds through four key steps:

1. Derive a recursion for the raw nonlinear iterate error, involving a linear core and a nonlinear perturbation.
2. Extend this decomposition to the Polyak–Ruppert average, expressing the nonlinear perturbation as a sum of matrix–vector products.
3. **Leverage semi-norm contraction** and geometric mixing ($\mathcal{A}_2$) to bound the average of the linear components.
4. Bound the nonlinear term involving the sum of matrix-vector products via three sub-steps:
   (a) Bound the semi-norm of the sum by a sum of matrix-vector semi-norm products, and then **bound the semi-norm of each vector by its induced norm**.
   (b) **Use semi-norm contraction** to obtain a uniform bound on the matrix semi-norm.
   (c) **Invoke the induced norm's equivalence with and a monotone norm** to prove that the overall nonlinear term vanishes rapidly, yielding the desired convergence rate.

We now describe the above four steps in detail.

Step 1 (Raw Error Decomposition). Using the definitions of b_i and A_i from Assumption $\mathcal{A}_1$, we introduce a few notations. For $t \ge 0$ and $Q \in \mathbb{R}^d$, define the d-dimensional vectors $\hat{b}_t^Q = \frac{1}{N}\sum_{i=1}^N b_i(Q, y_t^i)$ and $b^Q = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{y\sim\eta_i}[b_i(Q,y)]$, and the $d \times d$ matrices

$$\hat{A}_t^Q = \frac{1}{N}\sum_{i=1}^N A_i(Q, y_t^i), \quad A^Q = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{y\sim\eta_i}[A_i(Q,y)].$$

Using (4), (5), and the structural form of h_i in Assumption $\mathcal{A}_1$, it follows that H has the form $H(Q) = b^Q + A^Q Q$. Since Q^* is a fixed point of H, we then have from (6) that

$$b^{Q^*} + A^{Q^*} Q^* = H(Q^*) = Q^* + e^* \quad (22)$$

for some $e^* \in E$. Finally, for $t \ge 0$, define

$$\Delta_t := Q_t - (Q^* + e^*) \quad \text{and} \quad \bar{\Delta}_t := \bar{Q}_t - (Q^* + e^*).$$

Since $e^* \in E$, we have $v(\bar{\Delta}_T) = v(\bar{Q}_T - Q^*)$.

The following result decomposes the raw iterate error Δ_t into the desired linear and non-linear components.

Lemma 4.1. *For any $t \ge 0$, we have*

$$\Delta_{t+1} = \left[\mathbb{I} - \alpha_t\left(\mathbb{I} - \hat{A}_t^{Q^*}\right)\right]\Delta_t + \alpha_t\omega_t + \alpha_t\xi_t, \quad (23)$$

where $\omega_t := \left[\hat{b}_t^{Q^*} - b^{Q^*}\right] + \left[\hat{A}_t^{Q^*} - A^{Q^*}\right]Q^* + e^*$,

$$\xi_t := \left[\hat{b}_t^{Q_t} - \hat{b}_t^{Q^*}\right] + \left[\hat{A}_t^{Q_t} - \hat{A}_t^{Q^*}\right]Q_t.$$

Remark 4.2. *The $\alpha_t\xi_t$ term in* (23) *stems from the nonlinearity of the fixed-point iteration* (9)*. If this term is omitted,* (23) *reduces to a standard linear stochastic approximation.*

Step 2 (Polyak-Ruppert Error Decomposition). We now use the error decomposition achieved in (23) to derive a similar decomposition for $\bar{\Delta}_t$—the Polyak-Ruppert average of Δ_t—in Lemma 4.3 below. This refined decomposition underpins the subsequent analysis, allowing us to exploit the contraction property of the semi-norm alongside the equivalence between the induced norm and any monotone norm.

For $0 \le t_1 \le t_2$, let $\Gamma_{t_1:t_2} := \prod_{t=t_1}^{t_2-1}\left[\mathbb{I} - \alpha_t[\mathbb{I} - \hat{A}_t^{Q^*}]\right]$.

Lemma 4.3. *Let t_* be defined as in Theorem 3.5. Then, for $T \geq t_*$,*

$$\bar{\Delta}_T = \bar{\Delta}_T^{\text{trans}} + \bar{\Delta}_T^{\text{init}} + \bar{\Delta}_T^{\text{noise}} + \bar{\Delta}_T^{\text{nonlin}}, \tag{24}$$

where

$$\bar{\Delta}_T^{\text{trans}} := \left(\tfrac{1}{T}\sum_{t=0}^{t_*-1}\Delta_t\right), \quad \bar{\Delta}_T^{\text{init}} := \tfrac{1}{T}\sum_{t=t_*}^{T-1}\Gamma_{0:t}\Delta_0,$$

and

$$\bar{\Delta}_T^{\text{noise}} := \frac{1}{T}\sum_{t=t_*}^{T-1}\sum_{k=0}^{t-1}\alpha_k\Gamma_{k+1:t}\omega_k,$$

$$\bar{\Delta}_T^{\text{nonlin}} := \frac{1}{T}\sum_{t=t_*}^{T-1}\sum_{k=0}^{t-1}\alpha_k\Gamma_{k+1:t}\xi_k.$$

In turn, we have that

$$\mathbb{E}\upsilon(\bar{\Delta}_T) \leq \boldsymbol{A}_T^{\text{trans}} + \boldsymbol{A}_T^{\text{init}} + \boldsymbol{A}_T^{\text{noise}} + \boldsymbol{A}_T^{\text{nonlin}}, \tag{25}$$

where

$$\boldsymbol{A}_T^{\text{trans}} := \frac{1}{T}\sum_{t=0}^{t_*-1}\mathbb{E}\upsilon(\Delta_t), \qquad \boldsymbol{A}_T^{\text{init}} := \mathbb{E}\upsilon(\bar{\Delta}_T^{\text{init}}),$$

$$\boldsymbol{A}_T^{\text{noise}} := \mathbb{E}\upsilon(\bar{\Delta}_T^{\text{noise}}), \qquad \boldsymbol{A}_T^{\text{nonlin}} := \mathbb{E}\upsilon(\bar{\Delta}_T^{\text{nonlin}}).$$

Remark 4.4. *The superscripts* trans*,* init*,* noise*, and* nonlin *denote, respectively, the transient, initial-condition-based, noise-induced, and nonlinear components.*

Remark 4.5. *In the spirit of Remark 4.2, the $\boldsymbol{A}_T^{\text{nonlin}}$ term in* (25) *stems from the nonlinearity in* (9)*. If this nonlinear contribution were omitted,* (25) *would reduce to a standard decomposition in linear stochastic-approximation literature analyses, e.g., see (Durmus et al. 2025, equation (16)).*

Step 3 (Bounding the Averaged Linear Components). Now, we use the semi-norm contraction to bound the terms $\mathbf{A}_T^{\text{trans}}$, $\mathbf{A}_T^{\text{init}}$, and $\mathbf{A}_T^{\text{noise}}$, that are present in (25).

First, by applying the discrete Gronwall inequality on the Δ_t expansion in (23) and then using the fact that t_* is a constant, we get the following bound on $\mathbf{A}_T^{\text{trans}}$.

Lemma 4.6 (Bounding $\mathbf{A}_T^{\text{trans}}$). *For $T > t_*$, $\boldsymbol{A}_T^{\text{trans}} \leq \frac{C_\Delta t_*}{T}$, where $C_\Delta > 0$ is as defined in [Table 2] (Naskar, Thoppe, and Gupta 2025).*

Next, we get bounds on $\mathbf{A}_T^{\text{init}}$ and $\mathbf{A}_T^{\text{noise}}$. For $t \geq 0$, let $\mathbb{E}[\cdot|\mathcal{F}_t]$ be denoted by $\mathbb{E}_t[\cdot]$. We proceed with a bound on $\mathbb{E}_{t_1}\upsilon(\Gamma_{t_1:t_2})$, which is given in Lemma 4.7 below. To prove this result, we follow the recipe in (Durmus et al. 2025, Proposition 7), which proves a similar result for *linear stochastic iterations* with a fixed stepsize α. A key difference is that their result relies on a *norm bound* $\|\mathbb{I} - \alpha[\mathbb{I} - A]\| < 1-\alpha(1-\lambda)$ for some $\lambda > 0$, whereas our Lemma 4.7 uses an analogous semi-norm bound. The proof proceeds by dividing $\Gamma_{t_1:t_2}$ into blocks of a fixed size $h > 0$. Specifically, we define the ℓ-th block as

$$Y_\ell := \prod_{t=t(\ell-1)}^{t(\ell)-1}\left[\mathbb{I} - \alpha_t[\mathbb{I} - \hat{A}_t^{Q^*}]\right] \text{ for } \ell \leq \left\lfloor \tfrac{t_2-t_1}{h} \right\rfloor,$$

with $t(0) = t_1$ and $t(\ell) = t(\ell-1) + h - 1$. To bound $\upsilon(\Gamma_{t_1:t_2})$, we then bound each block. To get the latter, block Y_ℓ is decomposed into a contractive part—bounded using the semi-norm contraction, and a remainder part—bounded using the geometric mixing (Assumption $\mathcal{A}_\mathbf{2}$). In particular, we write $Y_\ell = \left[\mathbb{I} - \alpha_{\ell-1:\ell}[\mathbb{I} - A^{Q^*}]\right] + R_\ell$, where $\alpha_{\ell-1:\ell} := \sum_{s=t(\ell-1)}^{t(\ell)-1}\alpha_s$, and R_ℓ is a suitably defined remainder term involving $(\hat{A}_t^{Q^*} - A^{Q^*})$-type terms. The semi-norm contractive property of H implies that $\upsilon(A^{Q^*}) \leq \beta$. Together with $\alpha_{\ell-1:\ell} \geq h\alpha_{t(\ell)}$

$$\upsilon(\mathbb{I} - \alpha_{\ell-1:\ell}[\mathbb{I} - A^{Q^*}]) \leq 1 - h\alpha_{t(\ell)}(1-\beta).$$

Separately, we can show—using geometric mixing $\mathcal{A}_\mathbf{2}$ and the bounds in $\mathcal{A}_\mathbf{1}$(c)—that

$$\mathbb{E}_{t_1}\upsilon(R_\ell) \leq \alpha_{t(\ell)}\Big(\frac{C_E C_{\text{A}}}{1-\rho}\Big) + O(\alpha_{\ell(t-1)}^2).$$

By combining the above two expressions, we get

$$\mathbb{E}_{t_1}\upsilon(Y_\ell) \leq 1 - \beta_h\alpha_{t(\ell)} + O(\alpha_{t(\ell-1)}^2), \tag{26}$$

where h can be chosen to ensure that $\beta_h := h(1-\beta) - \frac{C_{\text{E}}C_{\text{A}}}{1-\rho} > 0$. The following bound follows from (26).

Lemma 4.7. *Let $C_\Gamma > 0$ be as defined in [Table 2] (Naskar, Thoppe, and Gupta 2025). For $0 \leq t_1 < t_2$,*

$$\mathbb{E}_{t_1}\upsilon(\Gamma_{t_1:t_2}) \leq C_\Gamma\, e^{-\beta_h\sum_{\ell=0}^{\lfloor (t_2-t_1)/h\rfloor}\alpha_{t_1+\ell h}}.$$

The following bound on $\mathbf{A}_T^{\text{init}}$ now follows from Lemma 4.7.

Lemma 4.8 (Bounding $\mathbf{A}_T^{\text{init}}$). *For $T > t_*$, $\boldsymbol{A}_T^{\text{init}} \leq \frac{\xi_\Gamma\upsilon(\Delta_0)}{T}$.*

The bound on the noise-induced term $\mathbf{A}_T^{\text{noise}}$ follows a recipe similar to (Durmus et al. 2025, Section 2, Propositions 8 and 10), and proceeds as follows. First, we decompose $\mathbf{A}_T^{\text{noise}}$ into two terms: the first term comprises a sum of products between noise ω_t and semi-norm contractive matrices $[\mathbb{I} - \alpha_t[\mathbb{I} - A^{Q^*}]]$—to be handled by the contractive nature of $[\mathbb{I}-\alpha_t[\mathbb{I}-A^{Q^*}]]$ and the geometric mixing of ω_t—while the second term involves a sum of products between random matrices $\hat{A}_t^{Q^*}$, contractive matrices $[\mathbb{I}-\alpha_t[\mathbb{I}-A^{Q^*}]]$, and matrix noise of the form $[\hat{A}_t^{Q^*} - A^{Q^*}]$—to be split into appropriately spaced intervals to invoke independence between $\hat{A}_t^{Q^*}$ and the matrix noise, and then bounded using geometric mixing.

The following lemma provides a bound on $\mathbf{A}_T^{\text{noise}}$.

Lemma 4.9 (Bounding $\mathbf{A}_T^{\text{noise}}$). *For $T > t_*$,*

$$\boldsymbol{A}_T^{\text{noise}} \leq \frac{C_1^{noise}\sqrt{\tau_T}}{\sqrt{NT}} + \frac{C_2^{noise}\ln(T)}{T^\alpha}.$$

where C_1^{noise}, C_2^{noise} are as in [Table 2] (Naskar, Thoppe, and Gupta 2025).

Step 4 (Analysis of the Nonlinear term). Lemma 4.11 below provides a bound for the term $\mathbf{A}_T^{\text{nonlin}}$ caused by the nonlinear perturbation. To get that bound, we first switch the double sum in the $\bar{\Delta}_T^{\text{nonlin}}$ expression from Lemma 4.3 to get

$$\bar{\Delta}_T^{\text{nonlin}} = \frac{1}{T}\sum_{k=0}^{T-2}\left(\alpha_k\sum_{t=k+1}^{T-1}\Gamma_{k+1:t}\right)\xi_k = \frac{1}{T}\sum_{k=0}^{T-2}M_k^T\xi_k,$$

where $M_{t_1}^{t_2} := \alpha_{t_1} \sum_{s=t_1+1}^{t_2-1} \Gamma_{t_1+1:s}$, for $0 \le t_1 < t_2$. Now, applying the semi-norm triangle inequality, we get

$$\mathbf{A}_T^{\text{nonlin}} \le \frac{1}{T} \sum_{k=0}^{T-2} \mathbb{E}\left[\upsilon(M_k^T)\, \upsilon(\xi_k)\right]. \tag{27}$$

Now, to bound (27), we show that $\mathbb{E}_k \upsilon(M_k^T) = O(1)$ and that $\mathbb{E}\upsilon(\xi_k)$ decays at a sufficiently fast rate. The $\upsilon(M_k^T)$ bound, given below, follows from Lemma 4.7.

Lemma 4.10. *For $0 \le t_1 < t_2$, we have $\mathbb{E}_{t_1}\upsilon(M_{t_1}^{t_2}) \le K_\Gamma$, where the constant $K_\Gamma > 0$ is as defined in [Table 2] (Naskar, Thoppe, and Gupta 2025).*

Next, we examine the decay rate of $\upsilon(\xi_k)$. Given that Assumption $\mathcal{A}_\mathbf{3}$ states $\mathbb{E}\upsilon(Q_k - Q^*)^2 = \tilde{O}(\alpha_k)$, it is enough for us to show that that $\upsilon(\xi_k) = O(\upsilon(Q_k - Q^*)^2)$. To do this, we utilize the following sandwiching relation obtained from $\mathcal{A}_\mathbf{1}$.(b):

$$0 \le \xi_k - [\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}] \le [\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}][Q_k - Q^*]. \tag{28}$$

(Li et al. 2023) exploits a similar sandwiching relation (without the $[\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}]$ term) to bound a certain remainder term. However, their work concerns norm contractive operators, where the monotonicity of their norm permits them to translate (28) to a desired norm-based inequality. This tactic fails for us since our semi-norm $\upsilon(\cdot)$ lacks this monotonicity property. To overcome this hurdle, we turn to the induced norm $\|\cdot\|$ given in Lemma 3.1. Then, we choose and fix a suitable monotone norm $\|\cdot\|_m$ to translate (28) and get

$$\|\xi_k\|_m \le \|\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}\|_m + \|[\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}][Q_k - Q^*]\|_m.$$

Combing this with $\upsilon(\xi_k) \le \|\xi_k\|$ and (12) gives

$$\upsilon(\xi_k) \le \frac{c_u}{c_\ell}\|\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}\| + \frac{c_u}{c_\ell}\|[\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}][Q_k - Q^*]\|. \tag{29}$$

To conclude **Step 4**, we show that the RHS in (29) is $O(\upsilon(Q_k - Q^*)^2)$. For the first term, we use the following fact from Assumptions $\mathcal{A}_\mathbf{1}$.(b) and $\mathcal{A}_\mathbf{1}$.(d):

$$\|\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}\| \begin{cases} = 0 & \text{if } \upsilon(Q_k - Q^*) < C_*, \\ \le 2C_{\text{b}} & \text{otherwise.} \end{cases}$$

This implies that

$$\|\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}\| \le (2C_{\text{b}}/C_*^2)\, \upsilon(Q_k - Q^*)^2. \tag{30}$$

Although a similar argument proves that $\|\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}\| = O(\upsilon(Q_k - Q^*)^2)$, applying the naive inequality $\|Bx\| \le \|B\|\|x\|$ for the second term in (29) yields an expression involving $\|x\|$—in our case, $\|Q_k - Q^*\|$—a potentially unbounded term!

Instead, we need a bound involving $\upsilon(Q_k - Q^*)$. Let $e_k := \arg\min_{e \in E} \|Q_k - Q^* - e\|$. Since $e_k \in E$, Assumption $\mathcal{A}_\mathbf{1}$.(a) implies $[\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}]e_k = 0$. Hence,

$$[\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}]\,[Q_k - Q^*] = [\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}]\,[Q_k - Q^* - e_k].$$

Additionally, from Lemma 3.1, we know $\|Q_k - Q^* - e_k\| = \upsilon(Q_k - Q^*)$. With this modification in mind, we use arguments similar to (30) to get

$$\|\hat{A}_k^{Q_k} - \hat{A}_k^{Q^*}\| \le (2C_{\text{A}}/C_*)\, \upsilon(Q_k - Q^*). \tag{31}$$

Then, we combine (29), (30), and (31) to get the desired bound on $\upsilon(\xi_t)$:

$$\begin{aligned} &\upsilon(\xi_k) \\ &\le \frac{c_u}{c_\ell}\Big(\|\hat{b}_k^{Q_k} - \hat{b}_k^{Q^*}\| + \|\hat{A}_k^{Q_t} - \hat{A}_k^{Q^*}\|\,\|Q_k - Q^* - e_k\|\Big) \\ &\le \frac{2c_u}{c_\ell}\Big(\frac{C_{\text{A}}}{C_*} + \frac{C_{\text{b}}}{C_*^2}\Big)\upsilon(Q_k - Q^*)^2. \end{aligned}$$

Together with Assumption $\mathcal{A}_\mathbf{3}$, the above expression implies that $\mathbb{E}\upsilon(\xi_k) = \tilde{O}(\alpha_k)$. The following lemma is a consequence of this fact combined with Lemma 4.10 and (27).

Lemma 4.11 (Bounding $\mathbf{A}_T^{\text{nonlin}}$). *For $T > t_*$,*

$$\boldsymbol{A}_T^{\text{nonlin}} \le \frac{2c_u K_\Gamma C_Q}{c_\ell(1-\alpha)}\Big(\frac{C_{\text{b}}}{C_*^2} + \frac{C_{\text{A}}}{C_*}\Big)\Big(\frac{\tau_T}{T^\alpha}\Big).$$

Finally, putting together the bounds on $\mathbf{A}_T^{\text{trans}}$, $\mathbf{A}_T^{\text{init}}$, $\mathbf{A}_T^{\text{noise}}$, and $\mathbf{A}_T^{\text{nonlin}}$ from Lemma 4.6, 4.8, 4.9, and 4.11, respectively, gives the desired bound in Theorem 3.5. □

4.2 Proof (Sketch) for Applications to Q-learning

In this section, we discuss the proof sketch for Theorem 3.9. Similar arguments are used to prove Theorem 3.12. The detailed proofs are provided in [Appendix](Naskar, Thoppe, and Gupta 2025). We show that the J-step average-reward Q-learning, as a special case of (9), satisfies conditions $\mathcal{A}_\mathbf{1}$–$\mathcal{A}_\mathbf{3}$. Due to IID sampling, $\mathcal{A}_\mathbf{2}$ holds trivially. Whereas $\mathcal{A}_\mathbf{3}$ follows from (Zhang, Zhang, and Maguluri 2021, Eq. B7). Additionally, one can show that there exist stochastic matrices $\{\hat{\mathcal{P}}_k^{\pi_Q}\}_{k=1}^J$ such that $h_i(Q, y) = b_i(Q, y) + A_i(Q, y)Q$, where $A_i = \hat{\mathcal{P}}_J^{\pi_Q}$ and $b_i = \sum_{k=0}^{J-1} \hat{\mathcal{P}}_k^{\pi_Q}\mathcal{R}_i$. Then, A_i's and b_i's satisfy condition $\mathcal{A}_\mathbf{1}$.(a)–(d). Conditions $\mathcal{A}_\mathbf{1}$.(a) and $\mathcal{A}_\mathbf{1}$.(c) follow from the stochasticity of $\{\hat{\mathcal{P}}_k^{\pi_Q}\}$. Condition $\mathcal{A}_\mathbf{1}$.(b) follows from the greedy property of π_Q. Lastly, $\mathcal{A}_\mathbf{1}$.(d) relies on the fact that the greedy policy is piece-wise constant on $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

5 Conclusions and Future Directions

We prove parameter-free convergence guarantees for nonlinear stochastic fixed-point iterations whose mean operators are only semi-norm contractions. In particular, we show that the existing parameter-free (but suboptimal) bounds for raw iterates (see $\mathcal{A}_\mathbf{3}$) can be improved to the optimal $\tilde{O}(1/T)$ rate by applying Polyak–Ruppert averaging.

As an application, we obtain the first parameter-free optimal rates for synchronous average-reward Q-learning and asynchronous exponentially discounted Q-learning.

Our results assume tabular models, full communication, and honest worker measurements. Future work should relax these assumptions to handle function approximation, limited bandwidth, and adversarial data.

Acknowledgements

The work was partially supported by the ARO grant W911NF2310111, and by Purdue University and DST-SERB's Overseas Visiting Doctoral Fellowship. Gugan Thoppe's research is supported in part by grants from the Walmart Centre for Tech Excellence at IISc, the Indo-French Centre for the Promotion of Advanced Research Grant (CEFIPRA; Project 7102-1), Kotak IISc AI-ML Centre's FinTech Grant, DST-SERB's Core Research Grant (CRG/2021/008330), and the Pratiksha Trust Young Investigator Award.

References

Bertsekas, D. P.; and Tsitsiklis, J. N. 1996. *Neuro-Dynamic Programming*. Belmont, MA: Athena Scientific.

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*, 1691–1692. PMLR.

Bhandari, J.; Russo, D.; and Singal, R. 2021. A Finite Time Analysis of Temporal Difference Learning with Linear Function Approximation. *Operations Research*, 69(3): 950–973.

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

Chen, Z.; Maguluri, S. T.; Shakkottai, S.; and Shanmugam, K. 2021a. Finite-sample analysis of off-policy td-learning via generalized bellman operators. *Advances in Neural Information Processing Systems*, 34: 21440–21452.

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

Chen, Z.; Zhang, S.; Zhang, Z.; Haque, S. U.; and Maguluri, S. T. 2025. A non-asymptotic theory of seminorm lyapunov stability: From deterministic to stochastic iterative algorithms. *arXiv preprint arXiv:2502.14208*.

Dal Fabbro, N.; Mitra, A.; and Pappas, G. J. 2023. Federated td learning over finite-rate erasure channels: Linear speedup under markovian sampling. *IEEE Control Systems Letters*, 7: 2461–2466.

Dalal, G.; Szorenyi, B.; Thoppe, G.; and Mannor, S. 2018. Finite Sample Analyses for TD(0) With Function Approximation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32.

Durmus, A.; Moulines, E.; Naumov, A.; and Samsonov, S. 2025. Finite-time high-probability bounds for Polyak–Ruppert averaged iterates of linear stochastic approximation. *Mathematics of Operations Research*, 50(2): 935–964.

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

Hu, J.; and Wellman, M. P. 2003. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov): 1039–1069.

Khodadadian, S.; Sharma, P.; Joshi, G.; and Maguluri, S. T. 2022. Federated Reinforcement Learning: Linear Speedup Under Markovian Sampling. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 10997–11057. PMLR.

Lakshminarayanan, C.; and Szepesvari, C. 2018. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International conference on artificial intelligence and statistics*, 1347–1355. PMLR.

Li, X.; Yang, W.; Liang, J.; Zhang, Z.; and Jordan, M. I. 2023. A statistical analysis of Polyak-Ruppert averaged Q-learning. In *International Conference on Artificial Intelligence and Statistics*, 2207–2261. PMLR.

Li, Y.; Ma, T.; and Zhang, H. 2018. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, 2–47. PMLR.

Liu, R.; and Olshevsky, A. 2023. Distributed TD (0) with almost no communication. *IEEE Control Systems Letters*, 7: 2892–2897.

Lohmiller, W.; and Slotine, J.-J. 1997. Applications of contraction analysis. In *Proceedings of the 36th IEEE Conference on Decision and Control*, volume 2, 1044–1049. IEEE.

Manchester, I. R.; and Slotine, J.-J. E. 2014. Output-feedback control of nonlinear systems using control contraction metrics and convex optimization. In *2014 4th Australian control conference (AUCC)*, 215–220. IEEE.

Manchester, I. R.; Tang, J. Z.; and Slotine, J.-J. E. 2017. Unifying robot trajectory tracking with control contraction metrics. In *Robotics Research: Volume 2*, 403–418. Springer.

Marjanovic, G.; and Solo, V. 2012. On l_q optimization and matrix completion. *IEEE Transactions on signal processing*, 60(11): 5714–5724.

Naskar, A.; Thoppe, G.; and Gupta, V. 2025. Parameter-free Optimal Rates for Nonlinear Semi-Norm Contractions with Applications to Q-Learning. arXiv:2508.05984.

Naskar, A.; Thoppe, G.; Koochakzadeh, A.; and Gupta, V. 2024. Federated TD Learning in Heterogeneous Environments with Average Rewards: A Two-timescale Approach with Polyak-Ruppert Averaging. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 387–393. IEEE.

Patil, G.; Prashanth, L.; Nagaraj, D.; and Precup, D. 2023. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, 5438–5448. PMLR.

Polyak, B. T.; and Juditsky, A. B. 1992. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4): 838–855.

Ruppert, D. 1991. *Stochastic approximation*. In Handbook of Sequetial Analysis.

Sayin, M.; Zhang, K.; Leslie, D.; Basar, T.; and Ozdaglar, A. 2021. Decentralized Q-learning in zero-sum Markov games. *Advances in Neural Information Processing Systems*, 34: 18320–18334.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. MIT Press.

Szepesvári, C. 2022. *Algorithms for reinforcement learning*. Springer nature.

Tsukamoto, H.; Chung, S.-J.; and Slotine, J.-J. E. 2021. Contraction theory for nonlinear stability analysis and learning-based control: A tutorial overview. *Annual Reviews in Control*, 52: 135–169.

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

Wang, H.; Mitra, A.; Hassani, H.; Pappas, G. J.; and Anderson, J. 2024. Federated TD Learning with Linear Function Approximation under Environmental Heterogeneity. *Transactions on Machine Learning Research*.

Zhang, S.; Zhang, Z.; and Maguluri, S. T. 2021. Finite Sample Analysis of Average-Reward TD Learning and Q-Learning. *Advances in Neural Information Processing Systems*, 34: 1230–1242.