On the Convergence Rates of Federated Q-Learning across Heterogeneous Environments

Anonymous authors
Paper under double-blind review

Abstract

Large-scale multi-agent systems are often deployed across wide geographic areas, where agents interact with heterogeneous environments. There is an emerging interest in understanding the role of heterogeneity in the performance of the federated versions of classic reinforcement learning algorithms. In this paper, we study synchronous federated Q-learning, which aims to learn an optimal Q-function by having K agents average their local Q-estimates per E iterations. We provide a characterization of the error evolution, which decays to zero as the number of iterations T increases. When $K(E-1)$ is below a certain threshold, similar to the homogeneous environment settings, there is a linear speed-up concerning K.

The slow convergence of having $E > 1$ turns out to be fundamental rather than an artifact of our analysis. We prove that, for a wide range of stepsizes, the ℓ_∞ norm of the error cannot decay faster than $\Theta_R\left(E/((1-\gamma)T)\right)$, where Θ_R only hides numerical constants and the specific choice of reward values. In addition, our experiments demonstrate that the convergence exhibits an interesting two-phase phenomenon. For any given stepsize, there is a sharp phase transition of the convergence: the error decays rapidly in the beginning yet later bounces up and stabilizes.

1 Introduction

Advancements in unmanned capabilities are rapidly transforming industries and national security by enabling fast-paced and versatile operations across domains such as advanced manufacturing (Park et al., 2019), autonomous driving (Kiran et al., 2021), and battlefields (Möhlenhof et al., 2021). Reinforcement learning (RL) – a cornerstone for unmanned capabilities – is a powerful machine learning method that aims to enable an agent to learn an optimal policy via interacting with its operating environment to solve sequential decision-making problems (Bertsekas & Tsitsiklis, 1996; Bertsekas, 2019). However, the ever-increasing complexity of the environment results in a high-dimensional state-action space, often imposing overwhelmingly high sample collection requirements on individual agents. This limited-data challenge becomes a significant hurdle that must be addressed to realize the potential of reinforcement learning.

In this paper, we study reinforcement learning within a federated learning framework (also known as Federated Reinforcement Learning (Qi et al., 2021; Jin et al., 2022; Woo et al., 2023)), wherein multiple agents independently collect samples and collaboratively train a common policy under the orchestration of a parameter server without disclosing the local data trajectories. A simple illustration can be found in Fig. 1. When the environments of all agents are homogeneous, it has been shown that the federated version of classic reinforcement learning algorithms can significantly alleviate the data collection burden on individual agents (Woo et al., 2023; Khodadadian et al., 2022) – error bounds derived therein exhibit a linear speedup in terms of the number of agents.

Moreover, by tuning the synchronization period E (i.e., the number of iterations between agent synchronization), the communication cost can be significantly reduced compared with $E = 1$ yet without significant performance degradation. However, many large-scale multi-agent systems are often deployed across wide geographic areas, resulting in agents interacting with heterogeneous environments. For instance, connected and autonomous vehicles (CAVs) operating in various regions of a metropolitan area encounter diverse

conditions such as varying traffic patterns, road infrastructure, and local regulations. The clients' federation must be managed in a way that ensures the learned policy is robust to environmental heterogeneity. There is an emerging interest in mathematically understanding the role of heterogeneity in the performance of the federated versions of classic reinforcement learning algorithms (Jin et al., 2022; Woo et al., 2023; Doan et al., 2019; Wang et al., 2023; Xie & Song, 2023) such as Q-learning, policy gradient methods, and temporal difference (TD) methods. In this paper, we study synchronous federated Q-learning (FQL) in the presence of environmental heterogeneity, which aims to learn an optimal Q-function by averaging local Q-estimates per E (where $E \geq 1$) update iterations on their local data. We leave the exploration of asynchronous Q-learning for future work. Federated Q-learning is a natural integration of FedAvg and Q-learning (Jin et al., 2022; Woo et al., 2023). The former is the most widely adopted classic federated learning algorithm (Kairouz et al., 2021; McMahan et al., 2017), and the latter is one of the most fundamental model-free reinforcement learning algorithms (Watkins & Dayan, 1992). Despite intensive study, the tight sample complexity of Q-learning in the single-agent setting was open until recently (Li et al., 2024). Similarly, the understanding of FedAvg is far from complete; a detailed discussion can be found in Section 2. A concise comparison of our work to the related work can be found in Table 1.

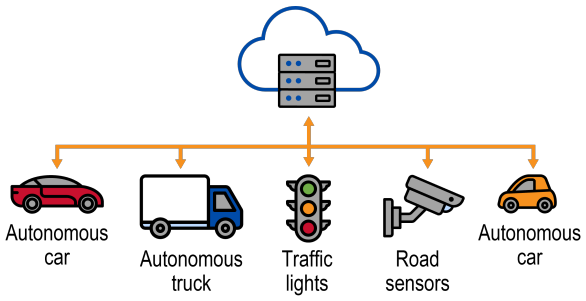


Figure 1: An illustration of a federated learning system.

Contributions. In this paper, we study synchronous federated Q-learning in the presence of environmental heterogeneity.

- We provide a characterization of the error evolution, which decays to zero as the number of iterations T increases. When $K(E-1)$ is below a threshold of $\widetilde{\mathcal{O}}\left((\kappa\epsilon)^{-1}(1-\gamma)^{-2}\right)$, similar to the homogeneous environment settings, there is a linear speed-up concerning K and the sample complexity is $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^5\epsilon^2}\right)$, matching the homogenous setting (Woo et al., 2023). In sharp contrast, when $K(E-1)$ is above the threshold, heterogeneous environments lead to significant performance degradation and results in a unique sample complexity of $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|\kappa E}{(1-\gamma)^3\epsilon}\right)$, where $\mathcal{S}$, $\mathcal{A}$ are the state and action sets, $\gamma \in (0,1)$ denotes the discount factor, and κ is a scalar characterizing the environment heterogeneity. Note E is also ϵ-dependent, making the sample complexity does not break the lower bound of $\Omega(\epsilon^{-2})$ for the single agent case.
- We prove that the convergence slowing down for $E > 1$ is fundamental. We show that the ℓ_∞ norm of the error cannot decay faster than $\Theta_R\left(\frac{E}{(1-\gamma)T}\right)$, where Θ_R only hides numerical constants and the specific choice of reward values. A practical implication of this impossibility result is that, eventually, having multiple local updates (i.e., $E > 1$) ends up consuming more samples (i.e., $E\times$ more) than using $E = 1$.
- Our numerical results illustrate that when the environments are heterogeneous and $E > 1$, there exists a sharp phase-transition of the error convergence for not too small stepsizes: The error decays rapidly in the beginning yet later bounces up and stabilizes. In addition, provided that the phase-transition time can be estimated, choosing different stepsizes for the two phases can lead to faster overall convergence for both constant and time-decaying stepsizes. We conjecture that this is because the error in phase 1 is mainly controlled by the initial error with impacting factor decay exponentially in time, and the error in phase 2 is dominated by the collective perturbation caused by environment heterogeneity and multiple local updates (i.e., $E > 1$).

Notably, all the asymptotic notations, e.g., $\mathcal{O}$ and $\widetilde{\mathcal{O}}$, unless otherwise specified, hide only numerical constants.

2 Related Work

Federated Learning. Federated learning is a communication-efficient distributed machine learning approach that enables training global models without sharing raw local data (McMahan et al., 2017; Kairouz et al.,

2021). Federated learning has been adopted in commercial applications that involve diverse edge devices such as autonomous vehicles (Du et al., 2020; Chen et al., 2021; Zeng et al., 2022; Posner et al., 2021; Peng et al., 2023), internet of things (Nguyen et al., 2019; Yu et al., 2020), industrial automation (Liu et al., 2020), healthcare (Yan et al., 2021; Sheller et al., 2019), and natural language processing (Yang et al., 2018; Ramaswamy et al., 2019). Multiple open-source frameworks and libraries are available such as FATE, Flower, OpenMinded-PySyft, OpenFL, TensorFlow Federated, and NVIDIA Clara.

FedAvg was proposed in the seminal work (McMahan et al., 2017), and has been one of the most widely implemented federated learning algorithms. It also has inspired many follow-up algorithms such as FedProx (Li et al., 2020b), FedNova (Wang et al., 2020), SCAFFOLD (Karimireddy et al., 2020), and adaptive federated methods (Deng et al., 2020). Despite intensive efforts, the theoretical understanding of FedAvg is far from complete. Most existing theoretical work on FedAvg overlooks the underlying data statistics at the agents, which often leads to misalignment of the pessimistic theoretical predictions and empirical success (Su et al., 2023; Pathak & Wainwright, 2020; Wang et al., 2022a;b). This theory and practice gap is studied in a recent work (Su et al., 2023) in the context of solving general non-parametric regression problems. It shows that the limiting points of the global model under FedAvg is one unbiased estimator of the underlying model that generates the data.

Reinforcement Learning. There has been extensive research on the convergence guarantees of reinforcement learning algorithms. A recent surge of work focuses on non-asymptotic convergence and the corresponding sample complexity for the single-agent setup. Bhandari et al. (2018) analyses non-asymptotic TD learning with linear function approximation (LFA) considering a variety of noise conditions, including noiseless, independent noise and Markovian noise. The results were extended to TD(λ) and Q-learning. Li et al. (2020a) investigates the sample complexity of asynchronous Q-learning with different families of learning rates. They also provide an extension of using variance reduction methods inspired by the seminal SVRG algorithm. Li et al. (2024) shows the sample complexity of Q-learning. Let $\mathcal{A}$ be the set of actions. When $|\mathcal{A}| = 1$, the sample complexity of synchronous Q-learning is sharp and minimax optimal, however, when $|\mathcal{A}| \geq 2$, it is shown that synchronous Q-learning has a lower bound which is not minimax optimal.

Multi-Agent RL. Yu et al. (2022) tests the performance of multi-agent Proximal Policy Optimization in four multi-agent testbeds, and it turns out the algorithm shows competitive results. The agents are homogeneous, i.e., they fully share the parameters. When the reward functions are different across agents, full parameter sharing is not effective. Christianos et al. (2021) proposes a selective parameter sharing technique, which automatically partitions agents so that they can benefit from the parameter sharing. Zhong et al. (2024) further proposes provably correct heterogeneous-agent algorithms, which allow agents to have different policy functions, and the algorithms show superior effectiveness and stability in various challenging benchmarks.

Federated Reinforcement Learning. Woo et al. (2023) provides sample complexity guarantees for both synchronous and asynchronous distributed Q-learning and reveals that given the same transition probability (i.e., homogeneous environment) for all agents, they can speed up the convergence process linearly by collaboratively learning the optimal Q-function. Salgia & Chi (2025) further develops a new algorithm which uses variance reduction techniques to estimate the Q-functions, and quantization for the Q-function to save the communication cost. They trade off between sample complexity and communication complexity and achieve a better sample complexity compared to Woo et al. (2023) under the same homogeneous setting. Doan et al. (2019) investigates the distributed Temporal Difference (TD) algorithm TD(0) with LFA under the setting of multi-agent MDP, where multiple agents act in a shared environment and each agent has its own reward function. They provide a finite-time analysis of this algorithm that with constant stepsize, the estimates of agents can converge to a neighborhood around optimal solutions and asymptotically converge to the optimal solutions. Khodadadian et al. (2022) studies on-policy federated TD learning, off-policy federated TD learning, and federated Q-learning of homogeneous environment and reward with Markovian noise. The sample complexity derived exhibits linear speedup with respect to the number of agents. Heterogeneous environments are considered in Jin et al. (2022); Wang et al. (2023); Xie & Song (2023); Zhang et al. (2023). Jin et al. (2022) studies federated Q-learning and policy gradient methods under the setting of different known transition probabilities for each agent. Yet, no state sampling is considered. Wang et al. (2023) proposes FedTD(0) with LFA dealing with the environmental and reward heterogeneity of MDPs. They rigorously prove that in a low-heterogeneity regime, there is a linear convergence speedup in the number of agents.

Work	RL Algorithm	Hetero-geneity	Optimality	Lower bound	Sampling	Finite-time	Task
Wang et al. (2023)	TD(0)	✔	✗	✗	✔	✔	Pred
Xie & Song (2023)	Policy Gradient	✔	✗	✗	✔	✗	Pred, Plan
Zhang et al. (2024)	SARSA	✔	✗	✗	✔	✔	Pred, Plan
Khodadadian et al. (2022)	TD, Q-Learning	✗	✔	✗	✔	✔	Pred, Plan
Jin et al. (2022)	Q-Learning, Policy Gradient	✔	✔	✗	✗	✔	Pred, Plan
Woo et al. (2023)	Q-Learning	✗	✔	✗	✔	✔	Pred, Plan
Zheng et al. (2023)	Q-Learning	✗	✔	✗	✔	✔	Pred, Plan
Our work	Q-Learning	✔	✔	✔	✔	✔	Pred, Plan

Table 1: Comparison of various works in the context of FRL. Pred and Plan stand for prediction (policy evaluation) and planning (policy optimization), respectively.

Xie & Song (2023) uses KL-divergence to penalize the deviation of local update from the global policy, and they prove that under the setting of heterogeneous environments, the local update is beneficial for global convergence using their method. Zhang et al. (2024) proposes FedSARSA using the classic on-policy RL algorithm SARSA with linear function approximation (LFA) under the setting of heterogeneous environments and rewards. They theoretically prove that the algorithm can converge to the near-optimal solution. Neither Xie & Song (2023) nor Zhang et al. (2024) characterize sample complexity.

Technical Comparisons with Woo et al. (2023); Zhang et al. (2024); Wang et al. (2023). While Zhang et al. (2024) and Wang et al. (2023) examined federated versions of TD learning and SARSA, our paper uniquely addresses federated Q-Learning, emphasizing its distinct theoretical and applicational advantages for optimal policy learning and planning—areas inadequately covered by the former methods. Specifically, the upper bound in Zhang et al. (2024) and Wang et al. (2023) do not indicate how fundamentally the convergence rates are impacted by the heterogeneity κ and synchronization period E, and the upper bounds does not decay to 0 as $T \to \infty$. In contrast, our upper bound converges to 0 as $T \to \infty$. Furthermore, we derived a lower bound on the convergence rates, showing the fundamental limitation of multiple local updates (i.e., $E > 1$) in the presence of environmental heterogeneity. To the best of our knowledge, this is the first result of its kind.

Our analysis of Theorem 1 builds upon the roadmap established by Woo et al. (2023), but adapting their analysis to our setting introduces significant challenges. In Woo et al. (2023), all agents operate in a homogeneous environment, meaning each of the K agents shares the same underlying transition distribution. This homogeneity allows the concentration bound on the difference between the true transition distribution and sampled estimates to become arbitrarily small as the number of samples increases. However, in our setting, each agent has its own environment with a distinct transition distribution. This heterogeneity introduces a perturbation term in the error upper bound that does not decrease with additional samples. Additionally, when $E > 1$ and $\kappa > 0$, the term involving $\kappa(E-1)$ in the upper bound becomes the dominant term, resulting a unique sample complexity. Further technical details and implications of these adjustments are provided in Corollary 2.

3 Preliminary on Q-Learning

Markov decision process. A Markov decision process (MDP) is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \gamma, R \rangle$, where $\mathcal{S}$ represents the set of states, $\mathcal{A}$ represents the set of actions, the transition probability $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to [0,1]$ provides the probability distribution over the next states given a current state s and action a, the reward function $R : \mathcal{S} \times \mathcal{A} \to [0,1]$ assigns a reward value to each state-action pair, and the discount factor $\gamma \in (0,1)$ models the preference for immediate rewards over future rewards. It is worth noting that $\mathcal{P} = \{P(\cdot \mid s,a)\}_{s \in \mathcal{S}, a \in \mathcal{A}}$ is a collection of $|\mathcal{S}| \times |\mathcal{A}|$ probability distributions over $\mathcal{S}$, one for each state-action pair (s,a).

Policy, value function, Q-Function, and optimality. A policy π specifies the action-selection strategy and is defined by the mapping $\pi : \mathcal{S} \to \Delta(A)$, where $\pi(a \mid s)$ denotes the probability of choosing action a when in state s. For a given policy π, the value function $V^\pi : \mathcal{S} \to \mathbb{R}$ measures the expected total discounted reward starting from state s:

$$V^\pi(s) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t,a_t)} \left[\sum_t \gamma^t R(s_t, a_t) \mid s_0 = s \right], \quad \forall s \in \mathcal{S}.$$

The state-action value function, or Q-function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, evaluates the expected total discounted reward from taking action a in state s and then following policy π:

$$Q^\pi(s,a) = R(s,a) \, + \, \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t,a_t)} \left[\sum_t \gamma^t \mathcal{R}(s_t, a_t) \mid s_0 = s, a_0 = a \right], \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

An optimal policy π^* is one that maximizes the value function for every state, that is $\forall s \in \mathcal{S}, V^{\pi^*}(s) \geq V^\pi(s)$ for any other $\pi \neq \pi^*$. Such a policy ensures the highest possible cumulative reward. The optimal value function V^* (shorthand for V^{π^*}) and the optimal Q-function Q^* (shorthand for Q^{π^*}) are defined under the optimal policy π^*.

The Bellman optimality equation for the value function and state-value function are:

$$V^*(s) = \max_a [R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s')]$$
$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q^*(s',a').$$

Q-learning. Q-learning (Watkins & Dayan, 1992) is a model-free reinforcement learning algorithm that aims to learn the value of actions of all states by updating Q-values through iterative exploration of the environment, ultimately converging to the optimal state-action function. Based on the Bellman optimality equation for the state-action function, the update rule for Q-Learning is formulated as:

$$Q_{t+1}(s,a) = (1-\lambda) Q_t(s,a) + \lambda [R(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_t(s',a')], \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where s' is sampled from the environment or the transition probability and λ is the stepsize.

4 Federated Q-learning

The federated learning system consists of one parameter server (PS) and K agents. The K agents are deployed in possibly heterogeneous yet independent environments. The K agents are modeled as Markov Decision Processes (MDPs) with $\mathcal{M}_k = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}^k, \gamma, R \rangle$ for $k = 1, \cdots, K$, where $\mathcal{P}^k = \{P^k(\cdot \mid s,a)\}_{s \in \mathcal{S}, a \in \mathcal{A}}$ are the collection of probability distributions that can be heterogeneous across agents. In the synchronous setting, each agent k has access to a generative model, and generates a new state sample for each (s,a) via $s_t^k(s,a) \sim P^k(\cdot \mid s,a)$, i.e., $\mathbb{P}\left\{s_t^k(s,a) = s'\right\} = P^k(s' \mid s,a)$ for all $s' \in \mathcal{S}$, independently across state-action pairs (s,a). For each (s,a), the global environment $\bar{P}(\cdot \mid s,a)$ (Jin et al., 2022) is defined as

$$\bar{P}(s' \mid s,a) = \frac{1}{K} \sum_{k=1}^{K} P^k(s' \mid s,a), \forall s' \tag{1}$$

with the corresponding global MDP defined as $\mathcal{M}_g = \langle \mathcal{S}, \mathcal{A}, \bar{\mathcal{P}}, \gamma, R \rangle$. Define transition heterogeneity κ as

$$\sup_{k,s,a} \left\| \bar{P}(\cdot \mid s,a) - P^k(\cdot \mid s,a) \right\|_\infty := \kappa. \tag{2}$$

Let Q^* denote the optimal Q-function of the global MDP. By the Bellman optimality equation, we have,

$$\forall (s,a), Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \bar{P}(s' \mid s,a) V^*(s'), \tag{3}$$

where $V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a)$ is the optimal value function.

The goal of federated Q-learning is to have the K agents collaboratively learn Q^*. We consider synchronous federated Q-learning, which is a natural integration of FedAvg and Q-learning (Woo et al., 2023; Jin et al., 2022) – described in Algorithm 1. Every agent initializes its local Q^k estimate as Q_0 and performs standard synchronous Q-learning based on the locally collected samples $s_t^k(s,a)$. Whenever $t+1 \mod E = 0$, through the parameter server, the K agents average their local estimate of Q; that is, all agents report their $Q_{t+\frac{1}{2}}^k$ to the parameter server, which computes the average and sends back to agents.

Algorithm 1 Synchronous Federated Q-Learning

Inputs: discount factor γ, E, total iteration T, step-size λ, initial estimate Q_0

1: **for** $k \in [K]$ **do**
2: $\quad Q_0^k = Q_0$
3: **end for**
4: **for** $t = 0$ to $T-1$ **do**
5: $\quad$ **for** $k \in [K]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
6: $\quad\quad Q_{t+\frac{1}{2}}^k(s,a) = (1-\lambda)Q_t^k(s,a) + \lambda\left(R(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_t^k(s_t^k(s,a),a')\right)$
7: $\quad\quad$ **if** $(t+1) \mod E = 0$ **then**
8: $\quad\quad\quad Q_{t+1}^k = \frac{1}{K}\sum_{k=1}^K Q_{t+\frac{1}{2}}^k$
9: $\quad\quad$ **else**
10: $\quad\quad\quad Q_{t+1}^k = Q_{t+\frac{1}{2}}^k$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: **end for**
14: **return** $Q_T = \frac{1}{K}\sum_{k=1}^K Q_T^k$

5 Main Results

With a little abuse of notation, let the matrix $P^k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ represent the transition kernel of the MDP of agent k with the (s,a)-th row being $P^k(\cdot \mid s,a) \in \mathbb{R}^{|\mathcal{S}|}$ – the transition probability of the state-action pair (s,a). For ease of exposition, we write $P^k(\cdot \mid s,a) = P^k(s,a)$ as the state transition probability at the state-action pair (s,a) when its meaning is clear from the context.

5.1 Main Convergence Results.

Let $\widetilde{P}_t^k \in \{0,1\}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ denote the local empirical transition matrix at the t-th iteration, defined as

$$\widetilde{P}_t^k(s' \mid s,a) = \mathbf{1}\{s' = s_t^k(s,a)\}.$$

Denoting $\widetilde{P}_i^k V^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times 1}$ with the (s,a)-th entry as $\widetilde{P}_i^k(s,a)V^* = \sum_{s' \in \mathcal{S}} \widetilde{P}_i^k(s'|s,a)V^*(s')$. Let $\bar{Q}_{t+1} := \frac{1}{K}\sum_{k=1}^K Q_{t+1}^k$. From lines 6, 8, and 10 of Algorithm 1, it follows that

$$\bar{Q}_{t+1} = \frac{1}{K}\sum_{k=1}^K \left((1-\lambda)Q_t^k + \lambda(R + \gamma \widetilde{P}_t^k V_t^k)\right),$$

where $V_t^k(s) := \max_{a \in \mathcal{A}} Q_t^k(s,a)$ for all $s \in \mathcal{S}$. Define

$$\Delta_{t+1} := Q^* - \bar{Q}_{t+1}, \text{ and } \Delta_0 := Q^* - Q_0. \tag{4}$$

The error iteration Δ_t is captured in the following lemma.

Lemma 1 (Error iteration)**.** *For any $t \geq 0$,*

$$\Delta_{t+1} = (1-\lambda)^{t+1}\Delta_0 + \gamma\lambda \sum_{i=0}^t (1-\lambda)^{t-i} \frac{1}{K}\sum_{k=1}^K (\bar{P} - \widetilde{P}_i^k)V^* + \gamma\lambda \sum_{i=0}^t (1-\lambda)^{t-i} \frac{1}{K}\sum_{k=1}^K \widetilde{P}_i^k (V^* - V_i^k). \tag{5}$$

To show the convergence of $\|\Delta_{t+1}\|_\infty$, we bound each of the three terms in the right-hand-side of (5). The following lemma is a coarse upper bound of errors.

Lemma 2. *Choosing $R(s,a) \in [0,1]$ for each state-action pair (s,a), and choose $0 \leq Q_0(s,a) \leq \frac{1}{1-\gamma}$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, then $0 \leq Q_t^k(s,a) \leq \frac{1}{1-\gamma}$, $0 \leq Q^*(s,a) \leq \frac{1}{1-\gamma}$,*

$$\left\|Q^* - Q_t^k\right\|_\infty \leq \frac{1}{1-\gamma}, \quad \text{and } \left\|V^* - V_t^k\right\|_\infty \leq \frac{1}{1-\gamma}, \quad \forall\, t \geq 0, \text{and } k \in [K]. \tag{6}$$

With the choice of Q_0 in Lemma 2, the first term in (5) can be bounded as $\left\|(1-\lambda)^{t+1}\Delta_0\right\|_\infty \leq (1-\lambda)^{t+1}\frac{1}{1-\gamma}$. In addition, as detailed in the proof of Lemma 4 and Theorem 1, the boundedness in Lemma 2 enables us to bound the second term in (5) via invoking the Hoeffding's inequality. It remains to bound the third term in (5), for which we follow the analysis roadmap of Woo et al. (2023) by a two-step procedure that is described in Lemma 3 and Lemma 4. Let

$$\Delta_t^k = Q^* - Q_t^k, \text{ and } \chi(t) = t - (t \mod E), \tag{7}$$

i.e., Δ_t^k is the local error of agent k, and $\chi(t)$ is the most recent synchronization iteration of t.

Lemma 3. *If $t \mod E = 0$, then $\left\|\frac{1}{K}\sum_{k=1}^K \widetilde{P}_t^k(V^* - V_t)\right\|_\infty \leq \|\Delta_t\|_\infty$. Otherwise,*

$$\left\|\frac{1}{K}\sum_{k=1}^K \widetilde{P}_t^k(V^* - V_t^k)\right\|_\infty \leq \left\|\Delta_{\chi(t)}\right\|_\infty + 2\lambda\frac{1}{K}\sum_{k=1}^K\sum_{t'=\chi(t)}^{t-1}\left\|\Delta_{t'}^k\right\|_\infty + \gamma\lambda\frac{1}{K}\sum_{k=1}^K\max_{s,a}\left|\sum_{t'=\chi(t)}^{t-1}\left(\widetilde{P}_{t'}^k(s,a) - \bar{P}(s,a)\right)V^*\right|.$$

where we use the convention that $\sum_{t'=\chi(t)}^{\chi(t)-1}\left\|\Delta_{t'}^k\right\|_\infty = 0$.

Lemma 4. *Choose $\lambda \leq \frac{1}{E}$. For any $\delta \in (0,1)$, with probability at least $(1-\delta)$,*

$$\left\|\Delta_i^k\right\|_\infty \leq \left\|\Delta_{\chi(i)}\right\|_\infty + \frac{3\gamma}{1-\gamma}\lambda(E-1)\kappa + \frac{3\gamma}{1-\gamma}\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}, \forall\ i \leq T, k \in [K]. \tag{8}$$

To bound the ℓ_∞ norm of the third term in (5), we first invoke Lemma 3, followed by Lemma 4. It is worth noting that directly applying Lemma 4 can also lead to a valid error bound yet the resulting bound will not decay as T increases with proper choice of stepsizes.

Both Lemma 3 and Lemma 4 are non-trivial adaptations of the approach in Woo et al. (2023) due to the absence of a common optimal action for any given state in heterogeneous environments. Moreover, in the homogeneous setting, each agent draws samples from the same true transition distribution, allowing concentration inequalities to bound the discrepancy between the true distribution and sampled estimates. However, this line of reasoning does not go through in the presence of environmental heterogeneity. When $\kappa > 0$, each of the K agents has its own transition distribution, and the discrepancy is captured by the environmental heterogeneity parameter κ.

Theorem 1 (Convergence)**.** *Choose $E - 1 \leq \frac{1}{\lambda}\min\{\frac{1-\gamma}{4\gamma}, \frac{1}{K}\}$ and $\lambda \leq \frac{1}{E}$. For any $\delta \in (0, \frac{1}{3})$, with probability at least $1 - 3\delta$, it holds that*

$$\|\Delta_T\|_\infty \leq \frac{4}{(1-\gamma)^2}\exp\left\{-\frac{1}{2}\sqrt{(1-\gamma)\lambda T}\right\} + \frac{14\gamma^2}{(1-\gamma)^2}\lambda(E-1)\kappa + \frac{16}{(1-\gamma)^2}\sqrt{\frac{\lambda}{K}\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}.$$

The first term of Theorem 1 is the standard error bound in the absence of environmental heterogeneity and sampling noises. The second term arises from environmental heterogeneity. It is clear that when $E = 1$, the environmental heterogeneity does not negatively impact the convergence. The last term results from the randomness in sampling.

Remark 1 (Eventual zero error). It is common to choose the stepsize λ based on the time horizon T. Let $\lambda = g(T)$ be a non-increasing function of T, and other parameters be fixed with respect to T. As long as $\lambda = g(T)$ decay in T, terms 2 and 3 in Theorem 1 will go to 0 as T increases. In addition, when $\lambda = \omega(1/T)$, the first term will decay to 0. Conversely, the convergence bounds in Zhang et al. (2024) and Wang et al. (2023) do not decay to 0.

There is a tradeoff in the convergence rates of the first term and the remaining terms – the slower λ decay in T leads to faster decay in the first term but slower in the remaining terms. Forcing these terms to decay around the same speed leads to slow overall convergence. Corollary 1 follows immediately from Theorem 1 via carefully choosing λ to balance the decay rates of different terms.

Corollary 1. *Choose* $(E-1) \leq \min \frac{1}{\lambda}\{\frac{1-\gamma}{4\gamma}, \frac{1}{K}\}$, *and* $\lambda = \frac{4\log^2(TK)}{T(1-\gamma)}$. *Let* $T \geq E$. *For any* $\delta \in (0, \frac{1}{3})$, *with probability at least* $1-3\delta$,

$$\|\Delta_T\|_\infty \leq \frac{4}{(1-\gamma)^2 TK} + \frac{32}{(1-\gamma)^{2.5}} \frac{\log(TK)}{\sqrt{TK}} \sqrt{\log \frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} + \frac{56\log^2(TK)}{(1-\gamma)^3} \frac{E-1}{T} \kappa.$$

Remark 2. Intuitively, both terms 1 and 2 decay as if there are TK iterations. In fact, the decay rate of the sampling noises in Corollary 1, with respect to TK, is the minimax optimal up to polylog factors (Vershynin, 2018). The decay of the third term is controlled by environmental heterogeneity when $E > 1$. In sharp contrast to the homogeneous settings, larger E significantly slows down the convergence of this term. We show in the next subsection that this slow convergence is fundamental.

Remark 3 (Communication cost and convergence). From Corollary 1, by choosing $E = \widetilde{\Theta}(\sqrt{T})$, and other parameters are fixed with respect to T, we can reach the same error bound of $\widetilde{\mathcal{O}}(1/\sqrt{T})$ with communication cost of $\widetilde{\mathcal{O}}(\sqrt{T})$, which is better than $\widetilde{\mathcal{O}}(T)$.

Corollary 2. *Choose* $E-1 \leq \frac{1}{\lambda}\min\{\frac{1-\gamma}{4\gamma}, \frac{1}{K}\}$ *and* $\lambda \leq \frac{1}{E}$, *and define* $x_1 = \frac{4096 \log \frac{|\mathcal{S}||\mathcal{A}|K}{\delta} \log^2(\frac{(1-\gamma)^2\epsilon}{8})}{K(1-\gamma)^5\epsilon^2}$,$x_2 = \frac{168\kappa(E-1)\gamma^2}{(1-\gamma)^3\epsilon} \log^2(\frac{(1-\gamma)^2\epsilon}{12})$, *and* $x_3 = \frac{9216}{K(1-\gamma)^5\epsilon^2} \log \frac{|\mathcal{S}||\mathcal{A}|K}{\delta} \log^2(\frac{(1-\gamma)^2\epsilon}{12})$,

- *When* $\kappa = 0$ *or* $E = 1$, *for any* $\delta \in (0, \frac{1}{3})$, *with probability at least* $1-3\delta$, *it holds that*

$$\|\Delta_T\|_\infty \leq \epsilon,$$

  when $T \geq \exp\{-W_{-1}(-\frac{1}{x_1})\}$, *and* $\lambda = \frac{\epsilon^2(1-\gamma)^4 K}{2304 \log \frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}$, *where* W_{-1} *is the Lambert* W *function. The resulting sample complexity is* $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^5\epsilon^2}\right)$.

- *When* $\kappa > 0$ *and* $E > 1$,
  - *If* $K(E-1) \geq \frac{55 \log \frac{|\mathcal{S}||\mathcal{A}|K}{\delta}}{\kappa\gamma^2\epsilon(1-\gamma)^2}$, *for any* $\delta \in (0, \frac{1}{3})$, *with probability at least* $1-3\delta$, *it holds that*

$$\|\Delta_T\|_\infty \leq \epsilon,$$

    when $T \geq \exp\{-W_{-1}(-\frac{1}{x_2})\}$ *and* $\lambda = \frac{\epsilon(1-\gamma)^2}{42\kappa(E-1)\gamma^2}$. *The sample complexity is* $\widetilde{\mathcal{O}}\left(\frac{\kappa|\mathcal{S}||\mathcal{A}|E}{(1-\gamma)^3\epsilon}\right)$.
  - *If* $K(E-1) \leq \frac{55 \log \frac{|\mathcal{S}||\mathcal{A}|K}{\delta}}{\kappa\gamma^2\epsilon(1-\gamma)^2}$, *for any* $\delta \in (0, \frac{1}{3})$, *with probability at least* $1-3\delta$, *it holds that*

$$\|\Delta_T\|_\infty \leq \epsilon,$$

    when $T \geq \exp\{-W_{-1}(-\frac{1}{x_3})\}$ *and* $\lambda = \frac{\epsilon^2(1-\gamma)^4 K}{2304 \log \frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}$. *The sample complexity is* $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^5\epsilon^2}\right)$.

Remark 4 (Sample complexity on K and E and conditional linear speedup.). From Corollary 2, we can conclude that when the setting is homogeneous (i.e., $\kappa = 0$) or the agents communicate every step (i.e., $E = 1$), the sample complexity $\widetilde{\mathcal{O}}\left(\frac{|\mathcal{S}||\mathcal{A}|}{K(1-\gamma)^5\epsilon^2}\right)$ matches the one in Woo et al. (2023). On the other hand, when the setting is heterogeneous (i.e., $\kappa > 0$) and $E > 1$, it is evident that if the total computation steps per synchronization are sufficiently small, i.e., $K(E-1) \leq \widetilde{\mathcal{O}}((\kappa\epsilon)^{-1}(1-\gamma)^{-2})$, the sample complexity also matches the one in the homogeneous setting, where there is a linear speedup. Otherwise, the sample complexity $\widetilde{\mathcal{O}}\left(\frac{\kappa|\mathcal{S}||\mathcal{A}|E}{(1-\gamma)^3\epsilon}\right)$ increases with E, meaning that multiple local rounds only consume more samples (i.e., E-times more) samples without achieving linear speedup.

5.2 On the Fundamentals of Convergence Slowdown for $E > 1$ in Heterogeneous Environments.

Theorem 2. *Let $Q_0 = \mathbf{0}$. For any even $K \geq 2$, there exists a collection of $\{(\mathcal{S}, \mathcal{A}, \mathcal{P}^k, R, \gamma) : k \in [K]\}$ where $|\mathcal{S}| = 2$, $|\mathcal{A}| = 1$, and $R := \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$ fixed for all MDPs in the collection, such that, for $E \geq 2$ and time-invariant stepsize $\lambda \leq \frac{1}{1+\gamma}$,*

$$\|\Delta_T\|_\infty \geq c_R \frac{E}{(1-\gamma)T},$$

when $T/E \in \mathbb{N}$ and $T \geq E \cdot \max\left\{\exp\{\frac{4E(\gamma+2)}{(1+\gamma)\gamma^2(E-1)}\}, \exp\left\{-W_{-1}\left(-\frac{1-\gamma}{2(1+\gamma)}\right)\right\}\right\}$, where W_{-1} is the Lambert W function, $c_R = \frac{1}{2}\min\left\{\frac{|r_1+r_2|}{e}, |r_1 - r_2|\right\}$ when $r_1 \neq r_2$ and $c_R = \frac{|r_1+r_2|}{2e}$ otherwise.

Proof Sketch. Below we provide the proof sketch of Theorem 2. The full proof is deferred to Appendix F.

The eventual slow rate convergence is due to the heterogeneous environments $\mathcal{P}^k$ regardless of the cardinality of the action space. In particular, we prove the slow rate when the action space is a singleton, in which case the Q-function coincides with the V-function. The process is also known as the Markov reward process. According to Algorithm 1, when $(t+1) \bmod E \neq 0$, we have

$$Q_{t+1}^k = \left((1-\lambda)I + \lambda\gamma P^k\right) Q_t^k + \lambda R.$$

Following Algorithm 1, we let z denote the z-th synchronization round, and obtain the following recursion between two synchronization rounds:

$$\Delta_{(z+1)E} = \bar{A}^{(E)}\Delta_{zE} + \left(\left(I - \bar{A}^{(E)}\right) - \left(I + \bar{A}^{(1)} + \ldots \bar{A}^{(E-1)}\right)\left(I - \bar{A}^{(1)}\right)\right) Q^*, \tag{9}$$

where $\bar{A}^{(\ell)} \triangleq \frac{1}{K}\sum_{k=1}^{K}(A^k)^\ell$ and $A^k \triangleq (1-\lambda)I + \lambda\gamma P^k$. While the first term on the right-hand side of (9) decays rapidly to zero, the second term is non-vanishing due to environment heterogeneity for $E \geq 2$. Specifically, to ensure the rapid decay of the first term, it is necessary to select a stepsize $\lambda = \widetilde{\Omega}(\frac{1}{zE})$. However, this choice results in the dominating residual error from the second term, which increases linearly with $\lambda E = \widetilde{\Omega}(1/z)$.

Next, we instantiate the analyses by constructing the set $\mathcal{P}^k$ over two states and an even number of clients with

$$P^{2k-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P^{2k} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \text{for } k \in \mathbb{N}. \tag{10}$$

Applying the formula of $\bar{A}^{(\ell)}$ yields the following eigen-decomposition:

$$\bar{A}^{(\ell)} = \alpha_\ell(I - \bar{P}) + \beta_\ell \bar{P},$$

where $\bar{P} = \frac{1}{2}\mathbf{1}\mathbf{1}^\top$, $\alpha_\ell \triangleq \frac{1}{2}(\nu_1^\ell + \nu_2^\ell)$, $\beta_\ell \triangleq \nu_2^\ell$, $\nu_1 \triangleq 1 - (1+\gamma)\lambda$, and $\nu_2 \triangleq 1 - (1-\gamma)\lambda$. For this instance of $\mathcal{P}_k$, the error evolution (9) reduces to $\Delta_{(z+1)E} = \left(\alpha_E(I - \bar{P}) + \beta_E\bar{P}\right)\Delta_{zE} + \kappa_E(I - \bar{P})Q^*$ with $\kappa_E \triangleq -\frac{\gamma}{2}\left(\frac{1-\nu_2^E}{1-\gamma} - \frac{1-\nu_1^E}{1+\gamma}\right)$, which further yields the following full error recursion:

$$\Delta_{zE} = \left(\alpha_E^z(I - \bar{P}) + \beta_E^z\bar{P}\right)\Delta_0 + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E(I - \bar{P})Q^*.$$

Starting from $Q_0 = 0$, the error can be decomposed into

$$\Delta_{zE} = \beta_E^z\bar{P}Q^* + \left(\alpha_E^z + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E\right)(I - \bar{P})Q^*. \tag{11}$$

The two terms of the error are orthogonal and both non-vanishing. Therefore, it remains to lower bound the maximum magnitude of two coefficients irrespective of the stepsize λ. To this end, we analyze two regimes of λ separated by a threshold $\lambda_0 \triangleq \frac{\log(T/E)}{(1-\gamma)T}$:

- Slow rate due to small stepsize when $\lambda \leq \lambda_0$. Since β_E^z decreases as λ increases,

$$\beta_E^z \geq (1-(1-\gamma)\lambda_0)^{zE} = \left(1 - \frac{\log z}{zE}\right)^{zE} \geq \frac{E}{eT}.$$

- Slow rate due to environment heterogeneity when $\lambda \geq \lambda_0$. We show that

$$\left|\alpha_E^z + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E\right| \geq \frac{E}{(1-\gamma)T}.$$

We conclude that at least one component of the error in (11) must be slower than the rate $\Omega(E/T)$.

Remark 5. The explicit calculations are based on a set $\mathcal{P}^k$ over a pair of states. Nevertheless, the evolution (9) is generally applicable. Similar analyses can be extended to scenarios involving more than two states, provided that the sequence of matrices $\bar{A}^{(\ell)}$ is simultaneously diagonalizable. For instance, the construction of the transition kernels in (10) can be readily extended to multiple states if the set $\mathcal{S}$ can be partitioned into two different classes. The key insight is the non-vanishing residual on the right-hand side of (9) when $E \geq 2$ due to the environment heterogeneity.

5.3 Discussion on Time-varying Stepsize

Although using time-varying stepsize is common and simple when implementing the algorithm, it is not easy to transfer from current time-invariant stepsize analysis to time-varying stepsize analysis. This is because in the time-invariant stepsize analysis we are dealing with a function of one variable, however, in the time-varying case, we are dealing with a function of T variables.

For example, in our lower bound analysis, we picked a threshold λ_0 and showed that no matter λ is greater or smaller than λ_0, the convergence rate is greater than $\Theta_R\left(E/((1-\gamma)T)\right)$, and we can claim we have covered all the cases. However, for time-varying stepsize, the number of stepsizes is T, and it is not easy to generalize a similar result by just considering several cases because each stepsize gives an additional dimension. Even if we know the sequence is decaying, without specifying a particular family of stepsizes, it is not possible to divide it into several cases as we did for time-invariant stepsize.

We conjecture that both approaches lead to comparable residual error levels over extended training. For example, the stepsizes used in Figure 4a and Figure 4b are $\frac{1}{\sqrt{T}}$ and $\frac{1}{\sqrt{t+1}}$, respectively. While the time-decaying stepsize appears to have faster initial convergence due to its larger values, we observe that as t increases, the convergence rates of the two strategies seem to align, suggesting a similar asymptotic behavior.

6 Experiments

Description of the setup. In our experiments, we consider $K = 20$ agents (Jin et al., 2022), each interacting with an independently and randomly generated 5×5 maze environment $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}^k, R, \gamma \rangle$ for $k \in \{1, 2, \cdots, 20\}$. The state set $\mathcal{S}$ contains 25 cells that the agent is currently in. The action set contains 4 actions $\mathcal{A} = \{\text{left}, \text{up}, \text{right}, \text{down}\}$. Thus, $|\mathcal{S}| \times |\mathcal{A}| = 100$. We choose $\gamma = 0.99$. For ease of verifying our theory, each entry of the reward $R \in \mathbb{R}^{100}$ is sampled from $\text{Bern}(p = 0.05)$, which slightly departs from a typical maze environment wherein only two state-action pairs have nonzero rewards. We choose this reward so that $\|\Delta_0\|_\infty \approx 100 = \frac{1}{1-\gamma}$, which is the coarse upper bound of $\|\Delta_t\|_\infty$ for all t. For each agent k, its state transition probability vectors $\mathcal{P}^k$ are constructed on top of standard state transition probability vectors of maze environments incorporated with a drifting probability 0.1 in each non-intentional action as in `WindyCliff` (Jin et al., 2022; Paul et al., 2019). In this way, the environment heterogeneity lies not only in the differences of the non-zero probability values (Jin et al., 2022; Paul et al., 2019) but also in the probability supports (i.e., the locations of non-zero entries). Our construction is more challenging: The environment heterogeneity κ as per (2) of our environment construction was calculated to be 1.2. Yet, the largest environment heterogeneity of the `WindyCliff` construction in Jin et al. (2022) is about 0.31.

We choose $Q_0 = \mathbf{0} \in \mathbb{R}^{100}$. All numerical results are based on 5 independent runs to capture the variability. The dark lines represent the mean of the runs, while the shaded areas around each line illustrate the range obtained by adding and subtracting one standard deviation from the mean. The maximum time duration is $T = 20,000$ in our experiment since it is sufficient to capture the characteristics of the training process.

Convergence behavior and two-phase phenomenon. We demonstrate through numerical simulations that our analysis aligns with the observed behaviors. For algorithms with a time-invariant stepsize, convergence requires sufficiently small stepsizes and a sufficiently large number of iterations T.

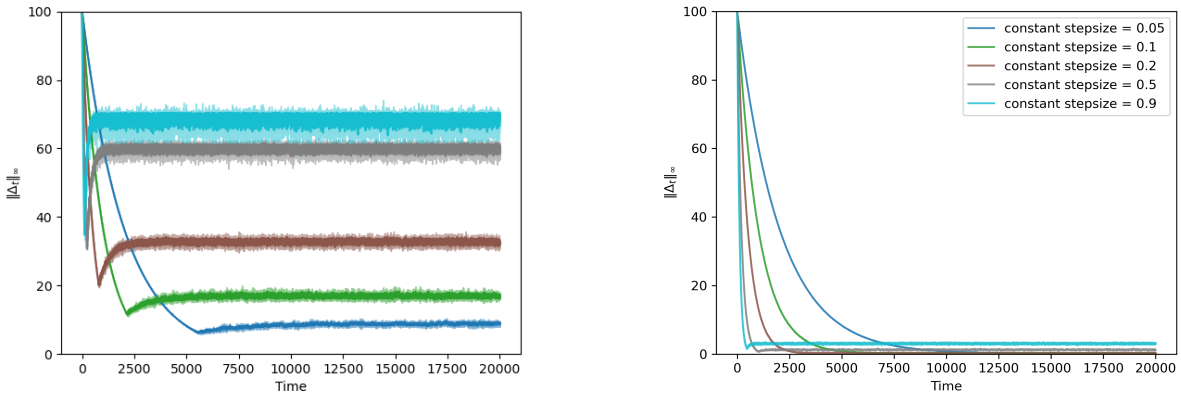
To explore the impact of stepsizes on convergence, we use $\lambda \in \{0.9, 0.5, 0.2, 0.1, 0.05\}$, spanning a range within $(0, 1)$. As shown in Figure 2a, these stepsizes are not sufficiently small, leading to a two-phase phenomenon: the ℓ_∞-norm of $\Delta_t = Q^* - \bar{Q}_t$ has a rapid decay in the first phase followed by a bounce back in the second phase. This phenomenon is distinctive to heterogeneous settings. In contrast, Figure 2b indicates that in homogeneous environments, no drastic bounce occurs, irrespective of the stepsize. Note that if multiple plots on the same page share the same legend, we display the legend only once for clarity.

Figure 4a (light blue curve) demonstrates that with a sufficiently small stepsize, such as $\lambda = \frac{1}{\sqrt{T}}$, the error continuously decreases, reaching approximately 24 by iteration 20,000.

A useful practice implication of our results is that: While constant stepsizes are often used in reinforcement learning problems because of the great performance in applications as described in Sutton & Barto (2018), they suffer significant performance degradation in the presence of environmental heterogeneity.

Impacts of the synchronization period E. In homogeneous settings (refer to Figure 5 in Appendix G.1), the synchronization period E has negligible impact, consistent with prior findings in the literature (Woo et al., 2023; Khodadadian et al., 2022). However, under heterogeneous conditions, larger E values lead to increased final error across the five constant stepsizes, as depicted in Figure 3 and Figure 2a. This degradation persists even with time-decaying stepsizes $\lambda_t = \frac{1}{\sqrt{t+1}}$, as shown in Figure 6. We hypothesize that larger E values require either smaller or more rapidly decaying stepsizes to mitigate the degradation caused by increased synchronization periods.

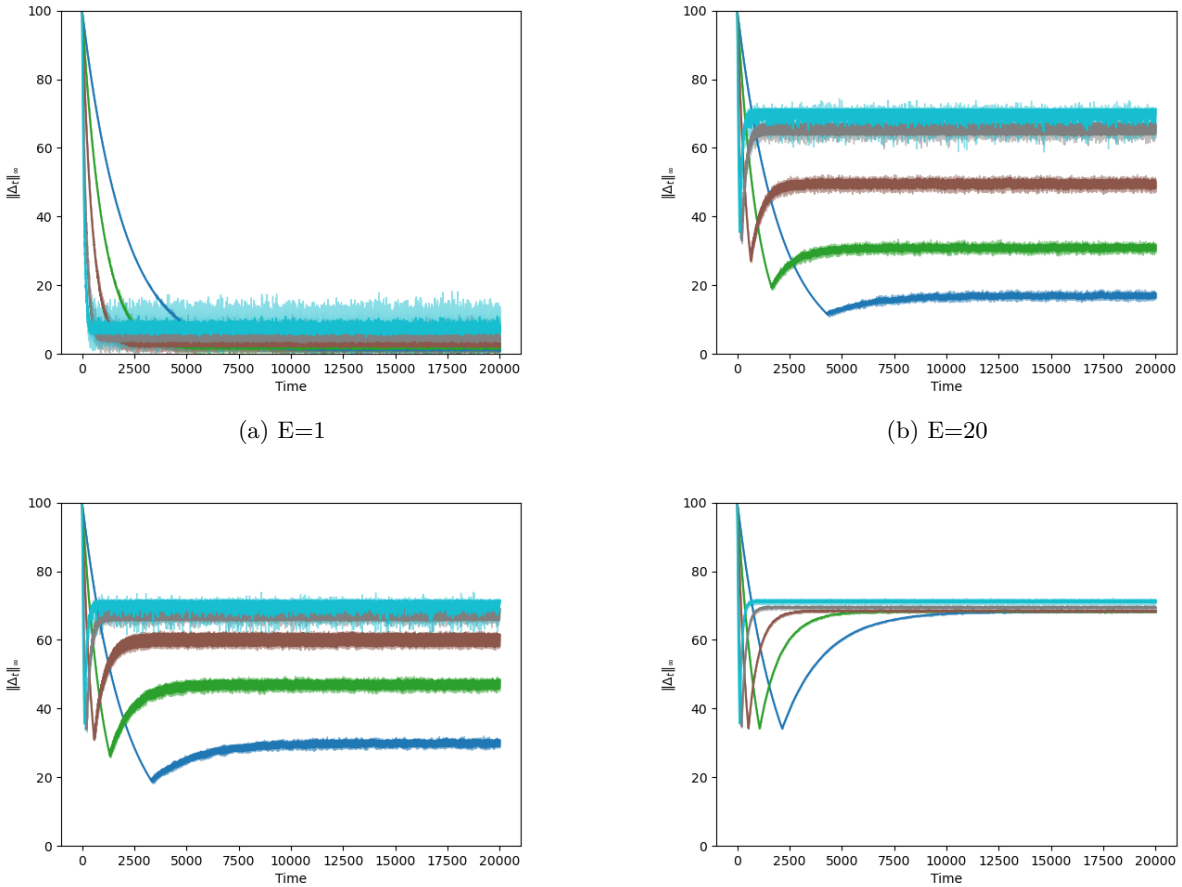
Potential utilization of the two-phase phenomenon. As shown in Figures 2a and 3, in the presence of environmental heterogeneity, the smaller the stepsizes, the smaller error $\|\Delta_t\|_\infty$ can reach and less significant of the error bouncing in the second phase. In our preliminary experiments, we tested small stepsizes $\lambda = 1/T^\alpha$ for $\alpha \in \{0.4, 0.5, \cdots, 1\}$, which eventually lead to small errors yet at the cost of being extremely slow. Among these choices, $\lambda = 1/\sqrt{T}$ has the fastest convergence performance yet is still ≈ 24 up to iteration 20,000.



(a) Heterogeneous environments $E = 10$.

(b) Homogeneous environments $E = 10$.

Figure 2: The ℓ_∞ error of different constant stepsizes under the heterogeneous and homogenous settings.



(a) E=1

(b) E=20

(c) E=40

(d) E=∞

Figure 3: Convergence behavior for constant stepsizes $(0.05, 0.1, 0.2, 0.5, 0.9)$ under various synchronization intervals E $(1, 20, 40, \infty)$. In heterogeneous settings, higher E and larger λ lead to higher residual errors.

Let t_0 be the iteration at which the error trajectory $\|\Delta_t\|_\infty$ switches from phase 1 to phase 2. Provided that t_0 can be estimated, choosing different stepsizes for the two phases can lead to faster overall convergence, compared with using the same stepsize throughout.

Figure 4a illustrates two-phase training with different phase 1 stepsizes and phase 2 stepsize $\lambda = 1/\sqrt{T}$ compared with using $\lambda = 1/\sqrt{T}$ throughout. Overall, using $\lambda = 1/\sqrt{T}$ throughout leads to the slowest convergence, highlighting the benefits of the two-phase training strategy. Among all two-phase stepsize choices, the stepsize of 0.05 in the first phase results in a longer phase 1 duration ($t_0 = 5550$) but the lowest final error (2.75327), suggesting a better convergence. We further test the convergence performance with respect to different target error levels, details can be found in Appendix G.3.

We also evaluated the two-phase training strategy using various time-decaying step sizes, including $\frac{1}{\sqrt{t+1}}$, $\frac{c+1}{t+c}$, $\frac{1}{t+1}$, and $\frac{1}{(t+1)^{0.7}}$. In all cases, Figure 4 shows the two-phase training has an advantage.

We leave the estimation and characterization of t_0 for future work.

References

Dimitri Bertsekas. *Reinforcement learning and optimal control*, volume 1. Athena Scientific, 2019.

Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, 1996.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pp. 1691–1692. PMLR, 2018.

Jin-Hua Chen, Min-Rong Chen, Guo-Qiang Zeng, and Jia-Si Weng. Bdfl: a byzantine-fault-tolerance decentralized federated learning method for autonomous vehicle. *IEEE Transactions on Vehicular Technology*, 70(9):8639–8652, 2021.

Filippos Christianos, Georgios Papoudakis, Muhammad A Rahman, and Stefano V Albrecht. Scaling multi-agent reinforcement learning with selective parameter sharing. In *International Conference on Machine Learning*, pp. 1989–1998. PMLR, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed td (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 1626–1635. PMLR, 2019.

Zhaoyang Du, Celimuge Wu, Tsutomu Yoshinaga, Kok-Lim Alvin Yau, Yusheng Ji, and Jie Li. Federated learning for vehicular internet of things: Recent advances and open issues. *IEEE Open Journal of the Computer Society*, 1:45–61, 2020.

Hao Jin, Yang Peng, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37. PMLR, 2022.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under markovian sampling. In *International Conference on Machine Learning*, pp. 10997–11057. PMLR, 2022.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020a.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13172–13179, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Thies Möhlenhof, Norman Jansen, and Wiam Rachid. Reinforcement learning environment for tactical networks. In *2021 International Conference on Military Communication and Information Systems (ICMCIS)*, pp. 1–8. IEEE, 2021.

Thien Duc Nguyen, Samuel Marchal, Markus Miettinen, Hossein Fereidooni, N Asokan, and Ahmad-Reza Sadeghi. Dïot: A federated self-learning anomaly detection system for iot. In *2019 IEEE 39th International conference on distributed computing systems (ICDCS)*, pp. 756–767. IEEE, 2019.

In-Beom Park, Jaeseok Huh, Joongkyun Kim, and Jonghun Park. A reinforcement learning approach to robust scheduling of semiconductor manufacturing facilities. *IEEE Transactions on Automation Science and Engineering*, 17(3):1420–1431, 2019.

Reese Pathak and Martin J Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in neural information processing systems*, 33:7057–7066, 2020.

Supratik Paul, Michael A Osborne, and Shimon Whiteson. Fingerprint policy optimisation for robust reinforcement learning. In *International Conference on Machine Learning*, pp. 5082–5091. PMLR, 2019.

Muzi Peng, Jiangwei Wang, Dongjin Song, Fei Miao, and Lili Su. Privacy-preserving and uncertainty-aware federated trajectory prediction for connected autonomous vehicles. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11141–11147, 2023. doi: 10.1109/IROS55552.2023.10341638.

Jason Posner, Lewis Tseng, Moayad Aloqaily, and Yaser Jararweh. Federated learning in vehicular networks: Opportunities and solutions. *IEEE Network*, 35(2):152–159, 2021.

Jiaju Qi, Qihao Zhou, Lei Lei, and Kan Zheng. Federated reinforcement learning: techniques, applications, and open challenges. *Intelligence & Robotics*, 2021. doi: 10.20517/ir.2021.02. URL `https://doi.org/10.20517%2Fir.2021.02`.

Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

Sudeep Salgia and Yuejie Chi. The sample-communication complexity trade-off in federated q-learning. *Advances in Neural Information Processing Systems*, 37:39694–39747, 2025.

Micah J Sheller, G Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pp. 92–104. Springer, 2019.

Lili Su, Jiaming Xu, and Pengkun Yang. A non-parametric view of fedavg and fedprox: Beyond stationary points. *Journal of Machine Learning Research*, 24(203):1–48, 2023.

Richard S. Sutton and Andrew G. Barto. *Chapter 2.5 Tracking a Nonstationary Problem, Reinforcement Learning: An Introduction*, chapter 8, pp. 33. The MIT Press, 2018.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Chunnan Wang, Xiang Chen, Junzhe Wang, and Hongzhi Wang. Atpfl: Automatic trajectory prediction model design under federated learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6563–6572, June 2022a.

Han Wang, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33: 7611–7623, 2020.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022b.

Christopher Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992. URL `https://api.semanticscholar.org/CorpusID:208910339`.

Jiin Woo, Gauri Joshi, and Yuejie Chi. The blessing of heterogeneity in federated q-learning: Linear speedup and beyond. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 37157–37216. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/woo23a.html`.

Zhijie Xie and Shenghui Song. FedKL: Tackling Data Heterogeneity in Federated Reinforcement Learning by Penalizing KL Divergence. *IEEE Journal on Selected Areas in Communications*, 41(4):1227–1242, April 2023. ISSN 1558-0008. doi: 10.1109/JSAC.2023.3242734. URL `https://ieeexplore.ieee.org/abstract/document/10038492?casa_token=yGyMDlnL_FsAAAAA:hqNvzWEb6yVKwTZVdHKLVnorDg07AWx4uujDsLLTTY_7unjr1ew8Yv4_UUAWfCx3X1b9wHNySP8`.

Bingjie Yan, Jun Wang, Jieren Cheng, Yize Zhou, Yixian Zhang, Yifan Yang, Li Liu, Haojiang Zhao, Chunjuan Wang, and Boyi Liu. Experiments of federated learning for covid-19 chest x-ray images. In *Advances in Artificial Intelligence and Security: 7th International Conference, ICAIS 2021, Dublin, Ireland, July 19-23, 2021, Proceedings, Part II 7*, pp. 41–53. Springer, 2021.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

Tianlong Yu, Tian Li, Yuqiong Sun, Susanta Nanda, Virginia Smith, Vyas Sekar, and Srinivasan Seshan. Learning context-aware policies from multiple smart homes via federated multi-task learning. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pp. 104–115. IEEE, 2020.

Tengchan Zeng, Omid Semiari, Mingzhe Chen, Walid Saad, and Mehdi Bennis. Federated learning on the road autonomous controller design for connected and autonomous vehicles. *IEEE Transactions on Wireless Communications*, 21(12):10407–10423, 2022.

Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *arXiv preprint arXiv:2401.15273*, 2024.

Shangtong Zhang, Remi Tachet Des Combes, and Romain Laroche. On the convergence of sarsa with linear function approximation. In *International Conference on Machine Learning*, pp. 41613–41646. PMLR, 2023.

Zhong Zheng, Fengyu Gao, Lingzhou Xue, and Jing Yang. Federated q-learning: Linear regret speedup with low communication cost. *arXiv preprint arXiv:2312.15023*, 2023.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32):1–67, 2024. URL `http://jmlr.org/papers/v25/23-0488.html`.

Appendices

A Proof of Lemma 1

The update of Δ_{t+1} is as follows:

$$\begin{aligned}
\Delta_{t+1} &= Q^* - \bar{Q}_{t+1} \\
&= \frac{1}{K}\sum_{k=1}^{K}(Q^* - ((1-\lambda)Q_t^k + \lambda(R + \gamma \widetilde{P}_t^k Q_t^k))) \\
&= \frac{1}{K}\sum_{k=1}^{K}((1-\lambda)(Q^* - Q_t^k) + \lambda(Q^* - R - \gamma \widetilde{P}_t^k V_t^k)) \\
&= (1-\lambda)\Delta_t + \gamma\lambda \frac{1}{K}\sum_{k=1}^{K}(\bar{P}V^* - \widetilde{P}_t^k V_t^k) \\
&= (1-\lambda)\Delta_t + \frac{\gamma\lambda}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_t^k)V^* + \frac{\gamma\lambda}{K}\sum_{k=1}^{K}\widetilde{P}_t^k(V^* - V_t^k) \\
&= (1-\lambda)^{t+1}\Delta_0 + \gamma\lambda\sum_{i=0}^{t}(1-\lambda)^{t-i}\frac{1}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_i^k)V^* \\
&\quad + \gamma\lambda\sum_{i=0}^{t}(1-\lambda)^{t-i}\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_i^k(V^* - V_i^k),
\end{aligned}$$

recalling that $\Delta_0 = Q^* - Q_0$.

B Proof of Lemma 2

We first show $0 \leq Q_t^k(s,a) \leq \frac{1}{1-\gamma}$ by inducting on t. When $t = 0$, this is true by the choice of Q_0. Suppose that $0 \leq Q_{t-1}^k(s,a) \leq \frac{1}{1-\gamma}$ for any state-action pair (s,a) and any client k. Let's focus on time t. When t is not a synchronization iteration (i.e., $t+1 \mod E \neq 0$), we have

$$\begin{aligned}
Q_t^k(s,a) &= (1-\lambda)Q_{t-1}^k(s,a) + \lambda(R(s,a) + \gamma\widetilde{P}_t^k(s,a)V_{t-1}^k) \\
&\leq \frac{1-\lambda}{1-\gamma} + \lambda(R(s,a) + \gamma\widetilde{P}_t^k(s,a)V_{t-1}^k) \\
&\overset{(a)}{\leq} \frac{1-\lambda}{1-\gamma} + \lambda(1 + \frac{\gamma}{1-\gamma}) \\
&\leq \frac{1}{1-\gamma} - \frac{\lambda}{1-\gamma} + \frac{\lambda}{1-\gamma} \\
&= \frac{1}{1-\gamma},
\end{aligned}$$

where inequality (a) holds because for any $s, V_{t-1}^k(s) = \max_{a\in\mathcal{A}} Q_{t-1}^k(s,a) \leq \frac{1}{1-\gamma}$ by the inductive hypothesis, and each element of $\widetilde{P}_t^k(s,a) \in [0,1]$. Then $\widetilde{P}_t^k(s,a)V_{t-1}^k \leq \|\widetilde{P}_t^k(s,a)\|_1 \left\|V_{t-1}^k\right\|_\infty \leq \frac{1}{1-\gamma}$ by Hölder's inequality. Similarly, we can show the case when t is a synchronization iteration.

With the above argument, we can also show that $0 \leq Q^*(s,a) \leq \frac{1}{1-\gamma}$ for any state-action pair (s,a). Therefore, we have that $\left\|Q^* - Q_t^k\right\|_\infty \leq \frac{1}{1-\gamma}$.

Next, we show that bound on $\left\|V^* - V_t^k\right\|_\infty$.

$$\begin{aligned}
\left\|V^* - V_t^k\right\|_\infty &= \max_{s\in\mathcal{S}} \left|V^*(s) - V_t^k(s)\right| \\
&= \max_{s\in\mathcal{S}} \left|\max_{a\in\mathcal{A}} Q^*(s,a) - \max_{a'\in\mathcal{A}} Q_t^k(s,a')\right| \\
&\leq \max_{s\in\mathcal{S},a\in\mathcal{A}} \left|Q^*(s,a) - Q_t^k(s,a)\right| \\
&= \left\|Q^* - Q_t^k\right\|_\infty \\
&\leq \frac{1}{1-\gamma}.
\end{aligned}$$

C Proof of Lemma 3

When $t \mod E = 0$, i.e., i is a synchronization round, $Q_t^k = Q_t^{k'}$ for any pair of agents $k, k' \in [K]$. Hence,

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(s,a)(V^* - V_t^k) &= \left(\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(s,a)\right)(V^* - \bar{V}_t) \\
&\leq \|\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(s,a)\|_1 \left\|V^* - \bar{V}_t\right\|_\infty \\
&\leq \left\|Q^* - \bar{Q}_t\right\|_\infty \\
&= \|\Delta_t\|_\infty . \qquad (12)
\end{aligned}$$

For general t, we have

$$\begin{aligned}
\left\|\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(V^* - V_t^k)\right\|_\infty &= \left\|\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(V^* - V_{\chi(t)}^k + V_{\chi(t)}^k - V_i^k)\right\|_\infty \\
&\leq \left\|\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(V^* - V_{\chi(t)}^k)\right\|_\infty + \left\|\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(V_{\chi(t)}^k - V_t^k)\right\|_\infty \\
&\leq \left\|\Delta_{\chi(t)}\right\|_\infty + \left\|\frac{1}{K}\sum_{k=1}^{K} \widetilde{P}_t^k(V_{\chi(t)}^k - V_t^k)\right\|_\infty \quad \text{by (12)} \\
&\leq \left\|\Delta_{\chi(t)}\right\|_\infty + \frac{1}{K}\sum_{k=1}^{K} \left\|V_{\chi(t)}^k - V_t^k\right\|_\infty . \qquad (13)
\end{aligned}$$

For any state $s \in \mathcal{S}$, we have

$$\begin{aligned}
&V_t^k(s) - V_{\chi(t)}^k(s) \\
&= Q_t^k(s, a_t^k(s)) - Q_{\chi(t)}^k(s, a_{\chi(t)}^k(s)) \\
&\overset{(a)}{\leq} Q_t^k(s, a_t^k(s)) - Q_{\chi(t)}^k(s, a_t^k(s)) \\
&= Q_t^k(s, a_t^k(s)) - Q_{t-1}^k(s, a_t^k(s)) + Q_{t-1}^k(s, a_t^k(s)) - Q_{t-2}^k(s, a_t^k(s)) \\
&\quad + \cdots + Q_{\chi(t)+1}^k(s, a_t^k(s)) - Q_{\chi(t)}^k(s, a_t^k(s)). \qquad (14)
\end{aligned}$$

where inequality (a) holds because $Q_{\chi(t)}^k(s, a_t^k(s)) \leq Q_{\chi(t)}^k(s, a_{\chi(t)}^k(s))$.

For each t' such that $\chi(t) \leq t' \leq t$, it holds that,

$$
\begin{aligned}
&Q^k_{t'+1}(s, a^k_t(s)) - Q^k_{t'}(s, a^k_t(s)) \\
&= (1-\lambda) Q^k_{t'}(s, a^k_t(s)) + \lambda(R(s, a^k_t(s)) + \gamma \widetilde{P}^k_{t'}(s, a^k_t(s)) V^k_{t'}) - Q^k_{t'}(s, a^k_t(s)) \\
&\overset{(a)}{=} -\lambda Q^k_{t'}(s, a^k_t(s)) + \lambda \left(Q^*(s, a^k_t(s)) - R(s, a^k_t(s)) - \gamma \bar{P}(s, a^k_t(s)) V^* + R(s, a^k_t(s)) + \gamma \widetilde{P}^k_{t'}(s, a^k_t(s)) V^k_{t'} \right) \\
&= \lambda \Delta^k_{t'}(s, a^k_t(s)) + \gamma\lambda \left((\widetilde{P}^k_{t'}(s, a^k_t(s)) - \bar{P}(s, a^k_t(s))) V^* + \widetilde{P}^k_{t'}(s, a^k_t(s))(V^k_{t'} - V^*) \right) \\
&\leq 2\lambda \left\| \Delta^k_{t'} \right\|_\infty + \gamma\lambda \left(\widetilde{P}^k_{t'}(s, a^k_t(s)) - \bar{P}(s, a^k_t(s)) \right) V^*,
\end{aligned}
$$

where equality (a) follows from the Bellman equation (3). Thus,

$$
\begin{aligned}
V^k_t(s) - V^k_{\chi(t)}(s) &\leq \sum_{t'=\chi(t)}^{t-1} Q^k_{t'+1}(s, a^k_t(s)) - Q^k_{t'}(s, a^k_t(s)) \\
&= 2\lambda \sum_{t'=\chi(t)}^{t-1} \left\| \Delta^k_{t'} \right\|_\infty + \gamma\lambda \sum_{t'=\chi(t)}^{t-1} \left(\widetilde{P}^k_{t'}(s, a^k_t(s)) - \bar{P}(s, a^k_t(s)) \right) V^*. \qquad (15)
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
V^k_t(s) - V^k_{\chi(t)}(s) &\geq \sum_{t'=\chi(t)}^{t-1} Q^k_{t'+1}(s, a^k_{\chi(t)}(s)) - Q^k_{t'}(s, a^k_{\chi(t)}(s)) \\
&\geq -2\lambda \sum_{t'=\chi(t)}^{t-1} \left\| \Delta^k_{t'} \right\|_\infty + \gamma\lambda \sum_{t'=\chi(t)}^{t-1} \left(\widetilde{P}^k_{t'}(s, a^k_{\chi(t)}(s)) - \bar{P}(s, a^k_{\chi(t)}(s)) \right) V^*. \qquad (16)
\end{aligned}
$$

Plugging the bounds in (15) and in (16) back into (13), we get

$$
\begin{aligned}
\left\| \frac{1}{K} \sum_{k=1}^{K} \widetilde{P}^k_t (V^* - V^k_t) \right\|_\infty &\leq \left\| \Delta_{\chi(t)} \right\|_\infty + \frac{1}{K} \sum_{k=1}^{K} \left\| V^k_{\chi(t)} - V^k_t \right\|_\infty \\
&\leq \left\| \Delta_{\chi(t)} \right\|_\infty + 2\lambda \frac{1}{K} \sum_{k=1}^{K} \sum_{t'=\chi(t)}^{t-1} \left\| \Delta^k_{t'} \right\|_\infty \\
&\quad + \gamma\lambda \frac{1}{K} \sum_{k=1}^{K} \max_{s,a} \left| \sum_{t'=\chi(t)}^{t-1} \left(\widetilde{P}^k_{t'}(s,a) - \bar{P}(s,a) \right) V^* \right|.
\end{aligned}
$$

D Proof of Lemma 4

When $i \mod E = 0$, then $\Delta^k_i = \Delta_{\chi(i)}$. When $i \mod E \neq 0$, we have

$$
\begin{aligned}
Q^k_i &= (1-\lambda) Q^k_{i-1} + \lambda \left(R + \gamma \widetilde{P}^k_{i-1} V^k_{i-1} \right) \\
&= (1-\lambda) Q^k_{i-1} + \lambda \left(Q^* - R - \gamma \bar{P} V^* + R + \gamma \widetilde{P}^k_{i-1} V^k_{i-1} \right).
\end{aligned}
$$

So,

$$\begin{aligned}\Delta_i^k &= (1-\lambda)\Delta_{i-1}^k + \lambda\gamma\left(\bar{P}V^* - \widetilde{P}_{i-1}^k V_{i-1}^k\right)\\ &= (1-\lambda)\Delta_{i-1}^k + \lambda\gamma(\bar{P} - \widetilde{P}_{i-1}^k)V^* + \lambda\gamma\widetilde{P}_{i-1}^k(V^* - V_{i-1}^k)\\ &\leq (1-\lambda)^{i-\chi(i)}\Delta_{\chi(i)} + \gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}(\bar{P} - \widetilde{P}_j^k)V^*\\ &\quad + \gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}\widetilde{P}_j^k(V^* - V_j^k). \end{aligned} \tag{17}$$

For any state-action pair (s, a),

$$|(1-\lambda)^{i-\chi(i)}\Delta_{\chi(i)}(s,a)| \leq (1-\lambda)^{i-\chi(i)}\left\|\Delta_{\chi(i)}\right\|_\infty. \tag{18}$$

For the second term, we have

$$\left\|\gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}(\bar{P} - \widetilde{P}_j^k)V^*\right\|_\infty \tag{19}$$

$$\leq \left\|\gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}(\bar{P} - P_j^k)V^*\right\|_\infty + \left\|\gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}(P_j^k - \widetilde{P}_j^k)V^*\right\|_\infty \tag{20}$$

$$\leq \frac{\gamma}{1-\gamma}\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-1-j}\kappa + \frac{\gamma}{1-\gamma}\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \tag{21}$$

$$\leq \frac{\gamma}{1-\gamma}\lambda(E-1)\kappa + \frac{\gamma}{1-\gamma}\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}, \tag{22}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}, i \in [T], k \in [K]$. From (20) to (21), since for each timestep, each agent independently samples the next state for all state-action pairs to form the sample transition matrix, we can use Hoeffding's inequality by treating $\lambda(1-\lambda)^{i-j-1}(P_j^k - \widetilde{P}_j^k)V^*$ as the independent random variables with their absolute values bounded by $\lambda(1-\lambda)^{i-j-1}\left\|V^*\right\|_\infty$.

In addition, we have

$$\left\|\gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}\widetilde{P}_j^k(V^* - V_j^k)\right\|_\infty \leq \gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}\left\|\Delta_j^k\right\|_\infty. \tag{23}$$

Combining the bounds in (18), (19), and (23), we get

$$\begin{aligned}\left\|\Delta_i^k\right\|_\infty &\leq (1-\lambda)^{i-\chi(i)}\left\|\Delta_{\chi(i)}\right\|_\infty + \frac{\gamma}{1-\gamma}\lambda(E-1)\kappa + \frac{\gamma}{1-\gamma}\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}\\ &\quad + \gamma\lambda\sum_{j=\chi(i)}^{i-1}(1-\lambda)^{i-j-1}\left\|\Delta_j^k\right\|_\infty\\ &\leq (1-(1-\gamma)\lambda)^{i-\chi(i)}\left\|\Delta_{\chi(i)}\right\|_\infty\\ &\quad + (1+\gamma\lambda)^{i-\chi(i)}\left(\frac{\gamma}{1-\gamma}\lambda(E-1)\kappa + \frac{\gamma}{1-\gamma}\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}\right), \end{aligned} \tag{24}$$

where the last inequality can be shown via inducting on $i - \chi(i) \in \{0, \cdots, E-1\}$. When $\lambda \leq \frac{1}{E}$,

$$(1+\gamma\lambda)^{i-\chi(i)} \leq (1+\lambda)^E \leq (1+1/E)^E \leq e \leq 3.$$

We get

$$\left\|\Delta_i^k\right\|_\infty \leq \left\|\Delta_{\chi(i)}\right\|_\infty + 3\frac{\gamma}{1-\gamma}\lambda(E-1)\kappa + 3\frac{\gamma}{1-\gamma}\sqrt{\lambda \log \frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}.$$

E Proof of Theorem 1

By Lemma 1,

$$\Delta_{t+1} = (1-\lambda)^{t+1}\Delta_0 + \sum_{i=0}^{t}(1-\lambda)^i \frac{\gamma\lambda}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_{t-i}^k)V^* + \sum_{i=0}^{t}(1-\lambda)^i \frac{\gamma\lambda}{K}\sum_{k=1}^{K}\widetilde{P}_{t-i}^k(V^* - V_{t-i}^k).$$

Taking the ℓ_∞ norm on both sides, we get

$$\begin{aligned}\left\|\Delta_{t+1}\right\|_\infty \leq (1-\lambda)^{t+1}\left\|\Delta_0\right\|_\infty &+ \left\|\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\frac{1}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_{t-i}^k)V^*\right\|_\infty \\ &+ \sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{t-i}^k(V^* - V_{t-i}^k)\right\|_\infty.\end{aligned}$$

We bound the three terms in the right-hand-side of the above-displayed equation separately.

Since $0 \leq Q_0(s,a) \leq \frac{1}{1-\gamma}$, the first term can be bounded as

$$(1-\lambda)^{t+1}\left\|\Delta_0\right\|_\infty \leq (1-\lambda)^{t+1}\frac{1}{1-\gamma}. \tag{25}$$

To bound the second term $\left\|\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\frac{1}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_{t-i}^k)V^*\right\|_\infty$, we have

$$\begin{aligned}\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\frac{1}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_{t-i}^k)V^* &= \sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\frac{1}{K}\sum_{k=1}^{K}(P^k - \widetilde{P}_{t-i}^k)V^* \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma(P^k - \widetilde{P}_{t-i}^k)V^*.\end{aligned}$$

Let $X_{i,k} = \frac{1}{K}\gamma\lambda(1-\lambda)^i(P^k - \widetilde{P}_{t-i}^k)V^*$. It is easy to see that $\mathbb{E}\left[X_{i,k}(s,a)\right] = 0$ for all (s,a). By Lemma 2, we have $|X_{i,k}(s,a)| \leq \frac{2}{K(1-\gamma)}\gamma\lambda(1-\lambda)^i$ for all (s,a). Since the sampling across clients and across iterations are independent, via invoking Hoeffding's inequality, for any given $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left\|\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\frac{1}{K}\sum_{k=1}^{K}(\bar{P} - \widetilde{P}_{t-i}^k)V^*\right\|_\infty \leq \frac{\gamma}{1-\gamma}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}}. \tag{26}$$

To bound the third term $\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{t-i}^k(V^*-V_{t-i}^k)\right\|_\infty$, following the roadmap of Woo et al. (2023), we divide the summation into two parts as follows. For any $\beta E \le t \le T$, we have

$$
\begin{aligned}
&\sum_{i=0}^{t}(1-\lambda)^i\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{t-i}^k(V^*-V_{t-i}^k)\right\|_\infty \\
&=\sum_{i=0}^{t}(1-\lambda)^{t-i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{i}^k(V^*-V_{i}^k)\right\|_\infty \\
&=\sum_{i=0}^{\chi(t)-\beta E}(1-\lambda)^{t-i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{i}^k(V^*-V_{i}^k)\right\|_\infty+\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{i}^k(V^*-V_{i}^k)\right\|_\infty \\
&\le\frac{\gamma}{1-\gamma}(1-\lambda)^{t-\chi(t)+\beta E}+\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{i}^k(V^*-V_{i}^k)\right\|_\infty.
\end{aligned}
$$

By Lemma 3,

$$
\begin{aligned}
&\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_{i}^k(V^*-V_{i}^k)\right\|_\infty \\
&\le\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma\Bigg(\left\|\Delta_{\chi(i)}\right\|_\infty+2\lambda\frac{1}{K}\sum_{k=1}^{K}\sum_{j=\chi(i)}^{i-1}\left\|\Delta_{t'}^k\right\|_\infty \\
&\qquad+\gamma\lambda\frac{1}{K}\sum_{k=1}^{K}\max_{s,a}\left|\sum_{j=\chi(i)}^{i-1}\left(\widetilde{P}_j^k(s,a)-\bar{P}(s,a)\right)V^*\right|\Bigg).
\end{aligned}
$$

Since $\widetilde{P}_j^k(s,a)$'s are independent across time j and across state action pair (s,a), and $|\widetilde{P}_j^k(s,a)-\bar{P}(s,a)V^*|\le\frac{1}{1-\gamma}$ (from Lemma 2), with Hoeffding's inequality and union bound, we get for any $\delta\in(0,1)$, with probability at least $1-\delta$,

$$
\left|\sum_{j=\chi(i)}^{i-1}\left(\widetilde{P}_j^k(s,a)-\bar{P}(s,a)\right)V^*\right|\le(E-1)\frac{1}{1-\gamma}\kappa+\frac{1}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}|\mathcal{A}|KT}{\delta}} \tag{27}
$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, $k\in K$, and i. By Lemma 4, with probability at least $(1-\delta)$, we have

$$
\begin{aligned}
&\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma 2\lambda\frac{1}{K}\sum_{k=1}^{K}\sum_{j=\chi(i)}^{i-1}\left\|\Delta_j^k\right\|_\infty \\
&\le 2\lambda^2\gamma\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\frac{1}{K}\sum_{k=1}^{K}\sum_{j=\chi(i)}^{i-1}\left(\left\|\Delta_{\chi(i)}\right\|_\infty+3\frac{\gamma}{1-\gamma}\lambda(E-1)\kappa+3\frac{\gamma}{1-\gamma}\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}\right) \\
&\le 2\lambda\gamma(E-1)\max_{\chi(t)-\beta E\le i\le t}\left\|\Delta_{\chi(i)}\right\|_\infty+\frac{6\gamma^2\lambda^2}{1-\gamma}(E-1)^2\kappa+\frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}.
\end{aligned}
$$

Thus, by applying the union bound, we get with probability at least $(1-2\delta)$,

$$\begin{aligned}
&\sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_i^k(V^*-V_i^k)\right\|_\infty \\
&\leq \gamma \max_{\chi(t)-\beta E\leq i\leq t}\left\|\Delta_{\chi(i)}\right\|_\infty + 2\lambda\gamma(E-1)\max_{\chi(t)-\beta E\leq i\leq t}\left\|\Delta_{\chi(i)}\right\|_\infty + \frac{6\gamma^2\lambda^2}{1-\gamma}(E-1)^2\kappa \\
&\quad + \frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \\
&\quad + \sum_{i=\chi(t)-\beta E+1}^{t}(1-\lambda)^{t-i}\lambda\gamma\left(\frac{\gamma\lambda}{1-\gamma}(E-1)\kappa + \frac{\gamma\lambda}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}\right) \\
&= \gamma(1+2\lambda(E-1))\max_{\chi(t)-\beta E\leq i\leq t}\left\|\Delta_{\chi(i)}\right\|_\infty + \frac{\gamma^2}{1-\gamma}(6\lambda^2(E-1)^2+\lambda(E-1))\kappa \\
&\quad + \frac{\gamma^2\lambda}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}.
\end{aligned}$$

The third term can be bounded as

$$\begin{aligned}
&\sum_{i=0}^{t}(1-\lambda)^{i}\lambda\gamma\left\|\frac{1}{K}\sum_{k=1}^{K}\widetilde{P}_i^k(V^*-V_i^k)\right\|_\infty \\
&\leq \frac{\gamma}{1-\gamma}(1-\lambda)^{t-\chi(t)+\beta E} + \gamma(1+2\lambda(E-1))\max_{\chi(t)-\beta E\leq i\leq t}\left\|\Delta_{\chi(i)}\right\|_\infty + \frac{\gamma^2}{1-\gamma}(6\lambda^2(E-1)^2+\lambda(E-1))\kappa \\
&+ \frac{\gamma^2\lambda}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}. \qquad (28)
\end{aligned}$$

Combing the bounds for terms (25), (26), and (28), we get the following recursion holds for all rounds T with probability at least $(1-3\delta)$:

$$\begin{aligned}
\left\|\Delta_{t+1}\right\|_\infty &\leq (1-\lambda)^{t+1}\frac{1}{1-\gamma} + \frac{\gamma}{1-\gamma}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} + \frac{\gamma}{1-\gamma}(1-\lambda)^{t-\chi(t)+\beta E} \\
&\quad + \gamma(1+2\lambda(E-1))\max_{\chi(t)-\beta E\leq i\leq t}\left\|\Delta_{\chi(i)}\right\|_\infty + \frac{\gamma^2}{1-\gamma}(6\lambda^2(E-1)^2+\lambda(E-1))\kappa \\
&\quad + \frac{\gamma^2\lambda}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \\
&\leq \gamma(1+2\lambda(E-1))\max_{\chi(t)-\beta E\leq i\leq t}\left\|\Delta_{\chi(i)}\right\|_\infty + \frac{2}{1-\gamma}(1-\lambda)^{\beta E} + \frac{\gamma^2}{1-\gamma}(6\lambda^2(E-1)^2+\lambda(E-1))\kappa \\
&\quad + \frac{\gamma^2\lambda}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \\
&\quad + \frac{\gamma}{1-\gamma}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}}.
\end{aligned}$$

Let

$$\begin{aligned}
\rho :=& \frac{2}{1-\gamma}(1-\lambda)^{\beta E} + \frac{\gamma^2}{1-\gamma}(6\lambda^2(E-1)^2+\lambda(E-1))\kappa \\
&+ \frac{\gamma^2\lambda}{1-\gamma}\sqrt{(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{6\gamma^2\lambda}{1-\gamma}(E-1)\sqrt{\lambda\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} \\
&+ \frac{\gamma}{1-\gamma}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}}. \qquad (29)
\end{aligned}$$

With the assumption that $\lambda \leq \frac{1-\gamma}{4\gamma(E-1)}$, the above recursion can be written as

$$\|\Delta_{t+1}\|_\infty \leq \frac{1+\gamma}{2} \max_{\chi(t)-\beta E \leq i \leq t} \left\|\Delta_{\chi(i)}\right\|_\infty + \rho.$$

Unrolling the above recursion L times where $L\beta E \leq t \leq T$, we obtain that

$$\begin{aligned}
\|\Delta_{t+1}\|_\infty &\leq (\frac{1+\gamma}{2})^L \max_{\chi(t)-L\beta E \leq i \leq t} \left\|\Delta_{\chi(i)}\right\|_\infty + \sum_{i=0}^{L-1} (\frac{1+\gamma}{2})^i \rho \\
&\leq (\frac{1+\gamma}{2})^L \frac{1}{1-\gamma} + \frac{2}{1-\gamma}\rho.
\end{aligned}$$

Choosing $\beta = \left\lfloor \frac{1}{E}\sqrt{\frac{(1-\gamma)T}{2\lambda}} \right\rfloor$, $L = \left\lceil \sqrt{\frac{\lambda T}{1-\gamma}} \right\rceil$, $t+1 = T$, we get

$$\begin{aligned}
\|\Delta_T\|_\infty \leq{}& \frac{1}{1-\gamma}(\frac{1+\gamma}{2})^{\sqrt{\frac{\lambda T}{1-\gamma}}} + \frac{2}{1-\gamma}\left(\frac{2}{1-\gamma}(1-\lambda)^{\beta E} + \frac{\gamma^2}{1-\gamma}(6\lambda^2(E-1)^2 + \lambda(E-1))\kappa \right. \\
&\left. + \left(\frac{6\gamma^2\lambda}{1-\gamma}\sqrt{E-1} + \frac{\gamma^2\sqrt{\lambda}}{1-\gamma}\right)\sqrt{\lambda(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{\gamma}{1-\gamma}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}}\right) \\
\leq{}& \frac{1}{1-\gamma}(\frac{1+\gamma}{2})^{\sqrt{\frac{\lambda T}{1-\gamma}}} + \frac{4}{(1-\gamma)^2}(1-\lambda)^{\sqrt{\frac{(1-\gamma)T}{2\lambda}}} + \frac{2\gamma^2}{(1-\gamma)^2}(6\lambda^2(E-1)^2 + \lambda(E-1))\kappa \\
&+ \left(\frac{12\gamma^2\lambda}{(1-\gamma)^2}\sqrt{E-1} + \frac{2\gamma^2\sqrt{\lambda}}{(1-\gamma)^2}\right)\sqrt{\lambda(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \\
\leq{}& \frac{1}{1-\gamma}\exp\left\{-\frac{1}{2}\sqrt{(1-\gamma)\lambda T}\right\} + \frac{4}{(1-\gamma)^2}\exp\left\{-\sqrt{(1-\gamma)\lambda T}\right\} \\
&+ \frac{2\gamma^2}{(1-\gamma)^2}(6\lambda^2(E-1)^2 + \lambda(E-1))\kappa \\
&+ \left(\frac{12\gamma^2\lambda}{(1-\gamma)^2}\sqrt{E-1} + \frac{2\gamma^2\sqrt{\lambda}}{(1-\gamma)^2}\right)\sqrt{\lambda(E-1)\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}} \\
\leq{}& \frac{4}{(1-\gamma)^2}\exp\left\{-\frac{1}{2}\sqrt{(1-\gamma)\lambda T}\right\} + \frac{2\gamma^2}{(1-\gamma)^2}(6\lambda^2(E-1)^2 + \lambda(E-1))\kappa \\
&+ \left(\frac{14\gamma^2\lambda}{(1-\gamma)^2}\sqrt{E-1}\right)\sqrt{\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}} + \frac{2\gamma}{(1-\gamma)^2}\sqrt{\frac{1}{K}\lambda\log\frac{|\mathcal{S}||\mathcal{A}|TK}{\delta}}.
\end{aligned}$$

By the assumption that $(E-1) \leq \frac{1}{K\lambda}$, the above can be further simplified as

$$\|\Delta_T\|_\infty \leq \frac{4}{(1-\gamma)^2}\exp\left\{-\frac{1}{2}\sqrt{(1-\gamma)\lambda T}\right\} + \frac{14\gamma^2}{(1-\gamma)^2}\lambda(E-1)\kappa + \frac{16}{(1-\gamma)^2}\sqrt{\frac{\lambda}{K}\log\frac{|\mathcal{S}||\mathcal{A}|KT}{\delta}}.$$

F Proof of Theorem 2

Let $|\mathcal{A}| = 1$, in which case Q-function coincides with the V-function. According to Algorithm 1, when $(t+1) \bmod E \neq 0$, we have

$$Q_{t+1}^k = \left((1-\lambda)I + \lambda\gamma P^k\right) Q_t^k + \lambda R.$$

Define $A^k \triangleq (1-\lambda)I + \lambda\gamma P^k$. We obtain the following recursion between two synchronization rounds:

$$Q_{(z+1)E}^k = (A^k)^E Q_{zE}^k + \left((A^k)^0 + \dots (A^k)^{E-1}\right) \lambda R.$$

Define

$$\bar{A}^{(\ell)} \triangleq \frac{1}{K}\sum_{k=1}^{K}(A^k)^\ell. \tag{30}$$

Note that Q^* is the fixed point under the transition kernel $\bar{P}$, we have $\lambda R = \lambda(I-\gamma\bar{P})Q^* = (I-\bar{A}^{(1)})Q^*$ since $\bar{A}^{(1)} = I - \lambda(I-\gamma\bar{P})$. Furthermore, since $Q_{tE}^1, \dots, Q_{tE}^K$ are identical due to synchronization, we get

$$\bar{Q}_{(z+1)E} = \bar{A}^{(E)}\bar{Q}_{zE} + \left(I + \bar{A}^{(1)} + \dots \bar{A}^{(E-1)}\right)\left(I - \bar{A}^{(1)}\right) Q^*.$$

Consequently,

$$\begin{aligned}\Delta_{(z+1)E} &= Q^* - \bar{Q}_{(z+1)E} \\ &= \bar{A}^{(E)}\Delta_{zE} + \left(\left(I - \bar{A}^{(E)}\right) - \left(I + \bar{A}^{(1)} + \dots \bar{A}^{(E-1)}\right)\left(I - \bar{A}^{(1)}\right)\right) Q^*. \end{aligned} \tag{31}$$

Next, consider $|\mathcal{S}| = 2$ and even K with

$$P^{2k-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P^{2k} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \text{for } k \in \mathbb{N}.$$

Then $\bar{P} = \frac{1}{2}\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1}$ denotes the all ones vector. For the above transition kernels, we have

$$\frac{1}{K}\sum_{k=1}^{K}(P^k)^\ell = \begin{cases} I, & \ell \text{ even}, \\ \bar{P}, & \ell \text{ odd}. \end{cases}$$

Applying the definition of $\bar{A}^{(\ell)}$ in (30) yields that

$$\begin{aligned}\bar{A}^{(\ell)} &= \frac{1}{K}\sum_{k=1}^{K}(A^k)^\ell \\ &= \frac{1}{K}\sum_{k=1}^{K}((1-\lambda)I + \lambda\gamma P^k)^\ell \\ &= \frac{1}{K}\sum_{k=1}^{K}\sum_{j=0}^{\ell}\binom{\ell}{j}(\lambda\gamma P^k)^j((1-\lambda)I)^{\ell-j} \\ &= \sum_{j \text{ even}}\binom{\ell}{j}(1-\lambda)^{\ell-j}(\lambda\gamma)^j(I - \bar{P} + \bar{P}) + \sum_{j \text{ odd}}\binom{\ell}{j}(1-\lambda)^{\ell-j}(\lambda\gamma)^j\bar{P} \\ &= \underbrace{\frac{1}{2}((1-\lambda-\lambda\gamma)^\ell + (1-\lambda+\lambda\gamma)^\ell)}_{\triangleq\alpha_\ell}(I - \bar{P}) + \underbrace{(1-\lambda+\lambda\gamma)^\ell}_{\triangleq\beta_\ell}\bar{P} \\ &= \alpha_\ell(I - \bar{P}) + \beta_\ell\bar{P}, \end{aligned}$$

which is the eigen-decomposition of $\bar{A}^{(\ell)}$. Let

$$\lambda_1 \triangleq (1+\gamma)\lambda, \lambda_2 \triangleq (1-\gamma)\lambda, \quad \nu_1 = 1-\lambda_1, \nu_2 = 1-\lambda_2.$$

Then

$$\alpha_\ell = \frac{1}{2}(\nu_1^\ell + \nu_2^\ell), \quad \beta_\ell = \nu_2^\ell. \tag{32}$$

Note that $0 \le \alpha \le \beta \le 1$ and $I - \bar{P}$ and $\bar{P}$ are orthogonal projection matrices satisfying $(I - \bar{P})\bar{P} = 0$. The matrices for the second term of the error on the right-hand side of 31 reduce to

$$\begin{aligned}
&\left(I + \bar{A}^{(1)} + \dots \bar{A}^{(E-1)}\right)\left(I - \bar{A}^{(1)}\right) \\
&= \left(\sum_{\ell=0}^{E-1} \alpha_\ell (I - \bar{P}) + \sum_{\ell=0}^{E-1} \beta_\ell \bar{P}\right)\left((\alpha_0 - \alpha_1)(I - \bar{P}) + (\beta_0 - \beta_1)\bar{P}\right) \\
&= \left((1 - \alpha_1)\sum_{\ell=0}^{E-1} \alpha_\ell (I - \bar{P})^2 + (1 - \beta_1)\sum_{\ell=0}^{E-1} \beta_\ell \bar{P}^2\right) \text{ since } \alpha_0 = \beta_0 = 1 \\
&= \left((1 - \alpha_1)\sum_{\ell=0}^{E-1} \alpha_\ell (I - \bar{P}) + (1 - \beta_1)\sum_{\ell=0}^{E-1} \beta_\ell \bar{P}\right) \text{ since } (I - \bar{P}) \text{ and } \bar{P} \text{ are idempotent.}
\end{aligned}$$

It follow that

$$\begin{aligned}
&\left(I - \bar{A}^{(E)}\right) - \left(I + \bar{A}^{(1)} + \dots \bar{A}^{(E-1)}\right)\left(I - \bar{A}^{(1)}\right) \\
&= \underbrace{\left((1 - \alpha_E) - (1 - \alpha_1)\left(\sum_{i=0}^{E-1} \alpha_i\right)\right)}_{\triangleq \kappa_E}(I - \bar{P}) + \underbrace{\left((1 - \beta_E) - (1 - \beta_1)\left(\sum_{i=0}^{E-1} \beta_i\right)\right)}_{=0}\bar{P}.
\end{aligned}$$

Applying (32) yields that

$$\kappa_E = -\frac{\gamma}{2}\left(\frac{1 - \nu_2^E}{1 - \gamma} - \frac{1 - \nu_1^E}{1 + \gamma}\right). \tag{33}$$

It follows from (31) that the error evolves as

$$\Delta_{(z+1)E} = \left(\alpha_E(I - \bar{P}) + \beta_E \bar{P}\right)\Delta_{zE} + \kappa_E (I - \bar{P})Q^*,$$

which further yields the following full recursion of the error:

$$\begin{aligned}
\Delta_{zE} &= \left(\alpha_E(I - \bar{P}) + \beta_E \bar{P}\right)^z \Delta_0 + \sum_{\ell=0}^{z-1}\left(\alpha_E(I - \bar{P}) + \beta_E \bar{P}\right)^\ell \kappa_E (I - \bar{P})Q^* \\
&= \left(\alpha_E^z(I - \bar{P}) + \beta_E^z \bar{P}\right)\Delta_0 + \sum_{\ell=0}^{z-1}\left(\alpha_E^\ell(I - \bar{P}) + \beta_E^\ell \bar{P}\right)\kappa_E (I - \bar{P})Q^* \\
&\quad \text{since } \left(\alpha_E(I - \bar{P}) + \beta_E \bar{P}\right)^\ell = \alpha_E^\ell(I - \bar{P}) + \beta_E^\ell \bar{P}, \forall \ell \in \mathbb{N} \\
&= \left(\alpha_E^z(I - \bar{P}) + \beta_E^z \bar{P}\right)\Delta_0 + \frac{1 - \alpha_E^z}{1 - \alpha_E}\kappa_E (I - \bar{P})Q^* \\
&= \left(\alpha_E^z + \frac{1 - \alpha_E^z}{1 - \alpha_E}\kappa_E\right)(I - \bar{P})Q^* + \beta_E^z \bar{P}Q^*,
\end{aligned}$$

where the last equality applied the zero initialization condition.

Note that $(I - \bar{P})Q^*$ and $\bar{P}Q^*$ are orthogonal vectors. Since $|\mathcal{S}| = 2$, we have

$$\|\Delta_{zE}\|_\infty \ge \frac{1}{\sqrt{2}}\|\Delta_{zE}\|_2 \ge \frac{\min\{\|(I - \bar{P})Q^*\|_2, \|\bar{P}Q^*\|_2\}}{\sqrt{2}} \cdot \max\left\{|\alpha_E^z + \frac{1 - \alpha_E^z}{1 - \alpha_E}\kappa_E|, \beta_E^z\right\}.$$

Let $R = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$, since $Q^* = (I - \gamma\bar{P})^{-1}R = (I - \bar{P})R + \frac{1}{1-\gamma}\bar{P}R$, we obtain that

$$(I - \bar{P})Q^* = (I - \bar{P})R = \frac{1}{2}\begin{bmatrix} r_1 - r_2 \\ r_2 - r_1 \end{bmatrix}, \qquad \bar{P}Q^* = \frac{1}{1-\gamma}\bar{P}R = \frac{1}{2(1-\gamma)}\begin{bmatrix} r_1 + r_2 \\ r_1 + r_2 \end{bmatrix}.$$

$$\|(I - \bar{P})Q^*\|_2 = \frac{\sqrt{2}}{2}\left|r_1 - r_2\right|, \qquad \|\bar{P}Q^*\|_2 = \frac{\sqrt{2}}{2(1-\gamma)}\left|r_1 + r_2\right|.$$

When $r_1 = r_2$, the error Δ_{zE} reduces to $\beta_E^z \bar{P}Q^*$, and $\|\Delta_{zE}\|_\infty = \frac{1}{2(1-\gamma)}\left|r_1 + r_2\right|\left|\beta_E^z\right|$; otherwise, $\min\{\|(I - \bar{P})Q^*\|_2, \|\bar{P}Q^*\|_2\} = \frac{\sqrt{2}}{2}\min\{\left|r_1 - r_2\right|, \frac{1}{1-\gamma}\left|r_1 + r_2\right|\}$. It remains to analyze the coefficients as functions of λ. To this end, we introduce the following lemma:

Lemma 5. *The following properties hold:*

1. *Negativity:* $\kappa_E < 0$*;*
2. *Monotonicity:* $\frac{\kappa_E}{1-\alpha_E}$ *is monotonically decreasing for* $\lambda \in (0, \frac{1}{1+\gamma})$*;*
3. *Upper bound:* $|\frac{\kappa_E}{1-\alpha_E}| \leq \frac{\gamma^2}{1-\gamma^2}$ *for* $\lambda \in (0, \frac{1}{1+\gamma})$*;*
4. *Lower bound: if* $(1+\gamma)\lambda \leq \frac{1}{2E}$*, then* $|\frac{\kappa_E}{1-\alpha_E}| \geq \frac{\lambda\gamma^2(E-1)}{4}$*.*

Proof. We prove the properties separately.

1. Note that $\nu_1 < \nu_2$, $1 - \nu_1 = (1+\gamma)\lambda$, and $1 - \nu_2 = (1-\gamma)\lambda$. Then it follows from (33) that

$$\kappa_E = -\frac{\lambda\gamma}{2}\sum_{i=1}^{E-1}(\nu_2^i - \nu_1^i) < 0.$$

2. For the monotonicity, it suffices to show that $\frac{d}{d\lambda}\frac{\kappa_E}{1-\alpha_E} \leq 0$. We calculate the derivative as

$$\frac{d}{d\lambda}\frac{\kappa_E}{1-\alpha_E} = \frac{\gamma E(1-\nu_1^E)(1-\nu_2^E)}{2(1-\gamma^2)(1-\alpha_E)^2}\left(\frac{(1+\gamma)\nu_1^{E-1}}{1-\nu_1^E} - \frac{(1-\gamma)\nu_2^{E-1}}{1-\nu_2^E}\right).$$

   Note that

$$\frac{(1+\gamma)\nu_1^{E-1}}{1-\nu_1^E} - \frac{(1-\gamma)\nu_2^{E-1}}{1-\nu_2^E} = \frac{1}{\lambda}\left(\frac{\nu_1^{E-1}}{1+\nu_1+\cdots+\nu_1^{E-1}} - \frac{\nu_2^{E-1}}{1+\nu_2+\cdots+\nu_2^{E-1}}\right) \leq 0.$$

3. For the upper bound, it suffices to show the result at $\lambda = \frac{1}{1+\gamma}$ due to the negativity and monotonicity. At $\lambda = \frac{1}{1+\gamma}$, we have

$$\left|\frac{\kappa_E}{1-\alpha_E}\right| = \frac{\gamma}{1-\gamma^2}\left(\gamma - \frac{(\frac{2\gamma}{1+\gamma})^E}{2-(\frac{2\gamma}{1+\gamma})^E}\right) \leq \frac{\gamma^2}{1-\gamma^2}.$$

4. For the lower bound, the case $E = 1$ trivially holds. Next, consider $E \geq 2$. We have

$$\begin{aligned}\frac{\kappa_E}{1-\alpha_E} &= -\frac{\gamma}{1-\gamma^2}\frac{(1+\gamma)(1-\nu_2^E) - (1-\gamma)(1-\nu_1^E)}{(1-\nu_1^E) + (1-\nu_2^E)} \\ &= -\lambda\gamma\frac{\sum_{\ell=1}^{E-1}(\nu_2^\ell - \nu_1^\ell)}{(1-\nu_1^E) + (1-\nu_2^E)}.\end{aligned}$$

Note that $1 - nx \le (1-x)^n \le 1 - \frac{1}{2}nx$ for $n \ge 1$ and $0 \le x \le \frac{1}{n}$. Then, for $(1+\gamma)\lambda \le \frac{1}{2E}$, we have

$$\nu_1^E = (1-(1+\gamma)\lambda)^E \ge 1 - (1+\gamma)\lambda E \ge \frac{1}{2},$$
$$\nu_2^E = (1-(1-\gamma)\lambda)^E \ge 1 - (1-\gamma)\lambda E.$$

Moreover, for all $x \in [\nu_1, \nu_2] \subseteq [0,1]$ and $\ell - 1 \le E$, we have

$$x^{\ell-1} \ge x^E \ge \nu_1^E \ge \frac{1}{2}.$$

We obtain that

$$\frac{\sum_{\ell=1}^{E-1}(\nu_2^\ell - \nu_1^\ell)}{(1-\nu_1^E) + (1-\nu_2^E)} \ge \frac{\sum_{\ell=1}^{E-1}\int_{\nu_1}^{\nu_2} \ell \cdot x^{\ell-1} dx}{2\lambda E} \ge \frac{\sum_{\ell=1}^{E-1} \ell \frac{1}{2}(\nu_2 - \nu_1)}{2\lambda E} = \frac{1}{4}\gamma(E-1).$$

The proof is completed. □

We consider two regimes of the stepsize separated by $\lambda_0 \triangleq \frac{\log z}{(1-\gamma)zE} < \frac{1}{1+\gamma}$, where the dominating error is due to the small stepsize and the environment heterogeneity, respectively:

Slow rate due to small stepsize when $\lambda \le \lambda_0$. Since β_E^z monotonically decreases as λ increases,

$$\beta_E^z = (1-(1-\gamma)\lambda)^{zE} \ge (1-(1-\gamma)\lambda_0)^{zE} = \left(1 - \frac{\log z}{zE}\right)^{zE}.$$

Note that $\frac{\log z}{zE} \in (0, \frac{1}{2})$, applying the fact $\log(1-x) + x \ge -x^2$ for $x \in [0, \frac{1}{2}]$ yields that

$$\log\left(1 - \frac{\log z}{zE}\right) + \frac{\log r}{zE} \ge -\left(\frac{\log z}{zE}\right)^2 \ge -\frac{1}{zE}.$$

Then we get

$$\beta_E^z \ge \left(1 - \frac{\log z}{zE}\right)^{zE} \ge \frac{1}{ez}.$$

Slow rate due to environment heterogeneity when $\lambda \ge \lambda_0$. Recall that $\lambda < \frac{1}{1+\gamma}$. Applying the triangle inequality yields that

$$\left|\alpha_E^z + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E\right| \ge \left|\frac{\kappa_E}{1-\alpha_E}\right| - \left(1 + \left|\frac{\kappa_E}{1-\alpha_E}\right|\right)\alpha_E^z.$$

For the first term, by the negativity and monotonicity in Lemma 5, it suffices to show the lower bound at $\lambda = \lambda_0$. Since $\lambda < \frac{1}{1+\gamma}$, then $\alpha_E = \frac{1}{2}\left((1-(1-\gamma)\lambda)^E + (1-(1+\gamma)\lambda)^E\right)$ decreases as λ increases. For $z \ge \exp\left\{-W_{-1}\left(-\frac{1-\gamma}{2(1+\gamma)}\right)\right\}$, where W_{-1} is the Lambert W function, such that $(1+\gamma)\lambda_0 \le \frac{1}{2E}$, we apply the lower bound in Lemma 5 and obtain that

$$\left|\frac{\kappa_E}{1-\alpha_E}\right| \ge \frac{\lambda_0\gamma^2(E-1)}{4} \ge \frac{\frac{\log z}{(1-\gamma)zE}\gamma^2(E-1)}{4} \ge \frac{(E-1)}{4E}\gamma^2\frac{\log z}{(1-\gamma)z}.$$

Additionally, applying the upper bound in Lemma 5 yields

$$\left(1 + \left|\frac{\kappa_E}{1-\alpha_E}\right|\right)\alpha_E^z \le \frac{\nu_2^{zE}}{1-\gamma^2} = \frac{(1-(1-\gamma)\lambda)^{zE}}{1-\gamma^2} \le \frac{1}{(1-\gamma^2)z}.$$

Therefore,

$$\begin{aligned}\left|\alpha_E^z + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E\right| &\geq \left|\frac{\kappa_E}{1-\alpha_E}\right| - \left(1 + \left|\frac{\kappa_E}{1-\alpha_E}\right|\right)\alpha_E^z \\ &\geq \frac{(E-1)}{4E}\gamma^2\frac{\log z}{(1-\gamma)z} - \frac{1}{(1-\gamma^2)z} \\ &= \frac{1}{(1-\gamma^2)z}\left((1+\gamma)\gamma^2\log(z)(E-1)/(4E) - 1\right) \\ &= \frac{1}{(1-\gamma)z}\left(\frac{\gamma^2\log(z)(E-1) - 4E/(1+\gamma)}{4E}\right).\end{aligned}$$

When $r_1 = r_2$,

$$\begin{aligned}\|\Delta_{zE}\|_\infty &= \frac{|r_1+r_2|}{2(1-\gamma)}\left|\beta_E^z\right| \\ &\geq \frac{|r_1+r_2|}{2(1-\gamma)}\frac{E}{eT};\end{aligned}$$

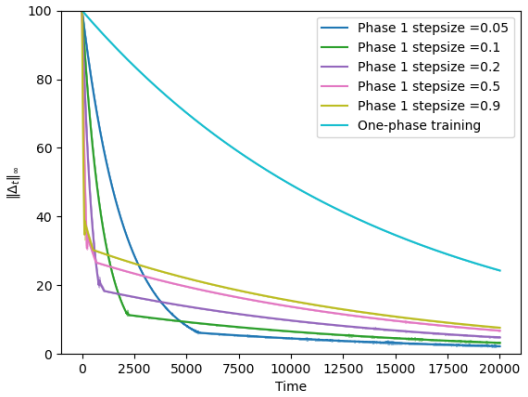
otherwise,

$$\begin{aligned}\|\Delta_{zE}\|_\infty &\geq \frac{\min\{\|(I-\bar{P})Q^*\|_2, \|\bar{P}Q^*\|_2\}}{\sqrt{2}} \cdot \max\left\{|\alpha_E^z + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E|, \beta_E^z\right\} \\ &\geq \frac{1}{2}\min\left\{|r_1-r_2|, \frac{1}{1-\gamma}|r_1+r_2|\right\}\max\left\{|\alpha_E^z + \frac{1-\alpha_E^z}{1-\alpha_E}\kappa_E|, \beta_E^z\right\} \\ &\geq \frac{1}{2}\min\left\{|r_1-r_2|, \frac{1}{1-\gamma}|r_1+r_2|\right\}\max\left\{\frac{1}{(1-\gamma)z}\left(\frac{\gamma^2\log(z)(E-1) - 4E/(1+\gamma)}{4E}\right), \frac{1}{ez}\right\} \\ &= \frac{1}{2}\min\left\{|r_1-r_2|, \frac{1}{1-\gamma}|r_1+r_2|\right\}\max\left\{\frac{E}{(1-\gamma)T}\left(\frac{\gamma^2\log(z)(E-1) - 4E/(1+\gamma)}{4E}\right), \frac{E}{eT}\right\}.\end{aligned}$$

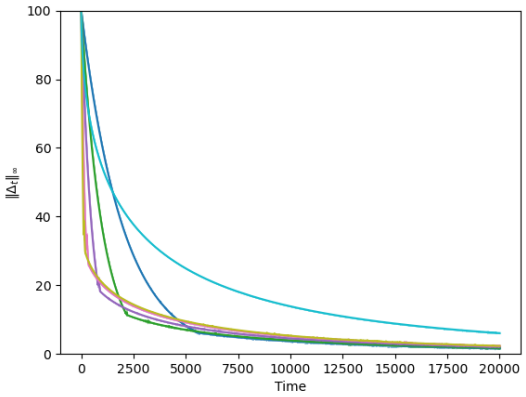
We can choose $\log(z) \geq \frac{4E(\gamma+2)}{(1+\gamma)\gamma^2(E-1)}, E \geq 2$ so that $\left(\frac{\gamma^2\log(z)(E-1)-4E/(1+\gamma)}{4E}\right) \geq 1$. Then the first term inside the max operator is bigger. Then,

$$\|\Delta_{zE}\|_\infty \geq \frac{1}{2}\min\left\{|r_1-r_2|, \frac{1}{1-\gamma}|r_1+r_2|\right\}\frac{E}{(1-\gamma)T}.$$

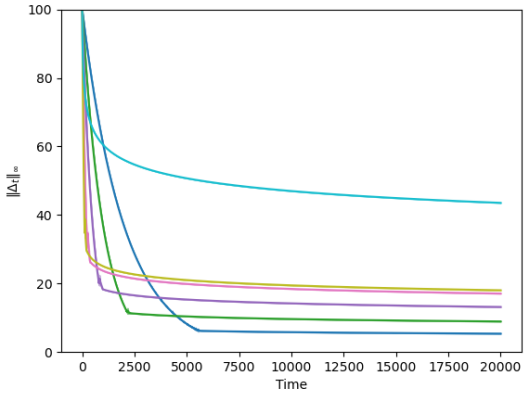
G Additional experiments



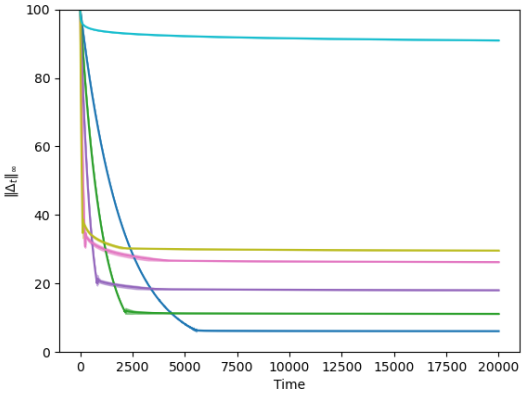
(a) Phase 2 stepsize $\lambda = \frac{1}{\sqrt{T}}$



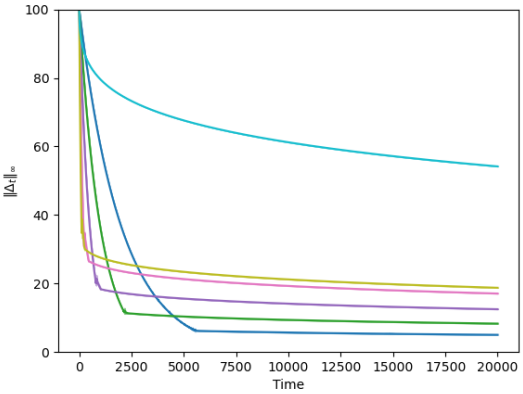
(b) Phase 2 stepsize $\lambda_t = \frac{1}{\sqrt{t+1}}$



(c) Phase 2 stepsize $\lambda_t = \frac{c+1}{t+c}$, where $c = 10$



(d) Phase 2 stepsize $\lambda_t = \frac{1}{t+1}$

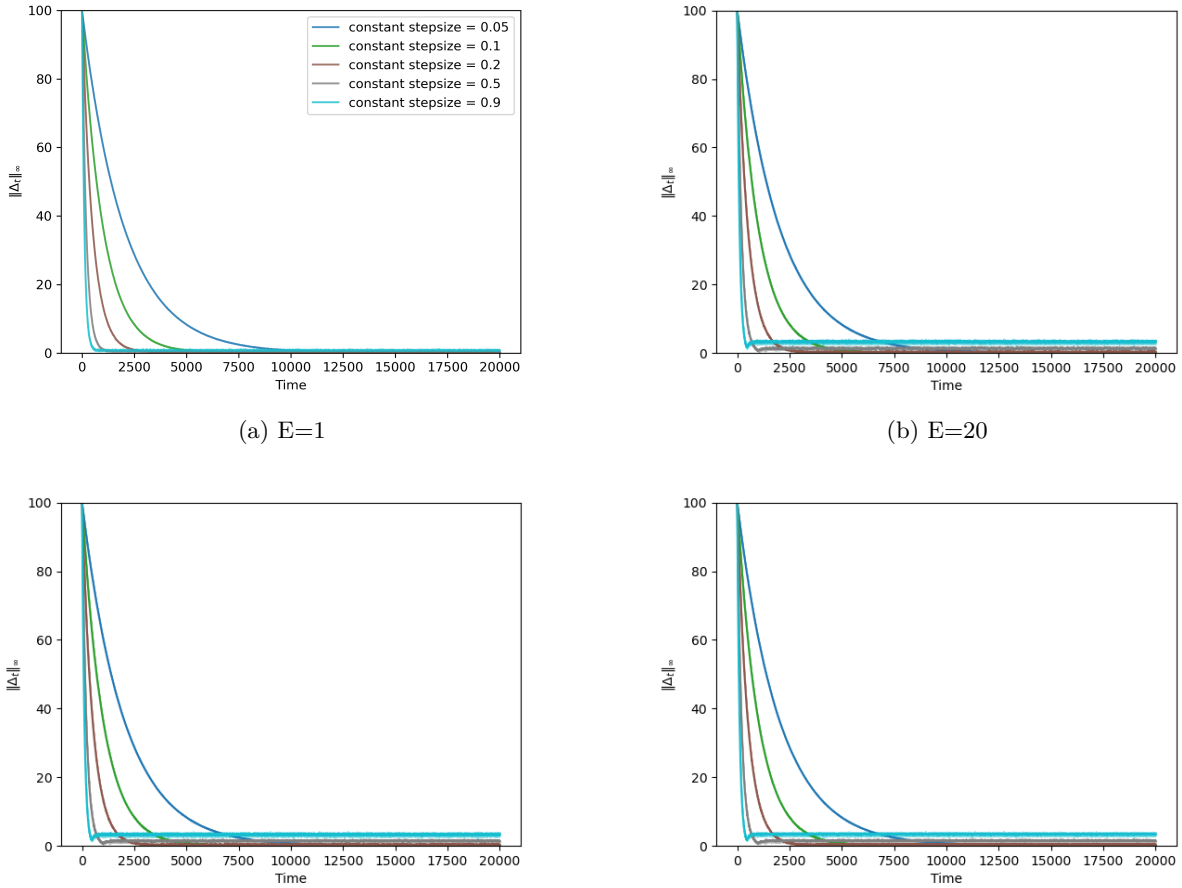


(e) Phase 2 stepsize $\lambda_t = \frac{1}{(t+1)^{0.7}}$

Figure 4: Choosing different stepsizes for phases 1 and 2 leads to faster overall convergence. $E = 10$.

G.1 Impacts of E on homogeneous settings.

For the homogeneous settings, in addition to $E = 10$, we also consider $E = \{1, 20, 40, \infty\}$, where $E = \infty$ means no communication among the agents throughout the entire learning process. Similar to Figure 2b, there is no obvious two-phase phenomenon even in the extreme case when $E = \infty$. Also, though there is indeed performance degradation caused by larger E, the overall performance degradation is nearly negligible compared with the heterogeneous settings shown in Figures 2a and 3.



(a) E=1

(b) E=20

(c) E=40

(d) E=∞

Figure 5: Homogeneous federated Q-learning with varying E.

G.2 Impacts of E on time-decaying stepsize

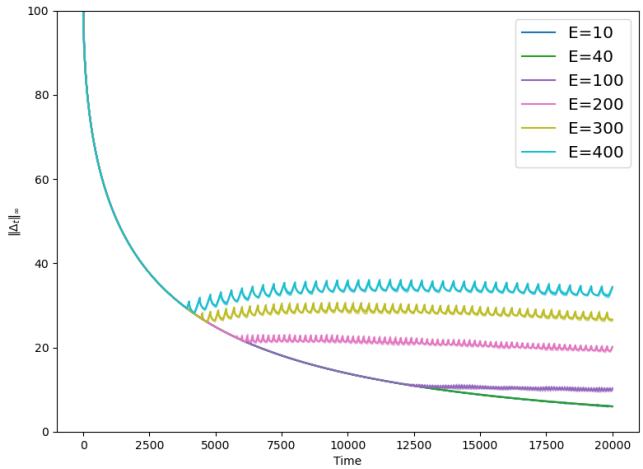


Figure 6: Using time-decaying stepsize $\lambda_t = \frac{1}{\sqrt{t+1}}$, the overall convergence becomes worse as E increases

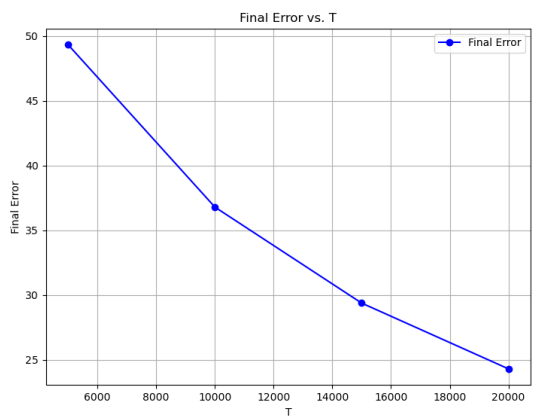
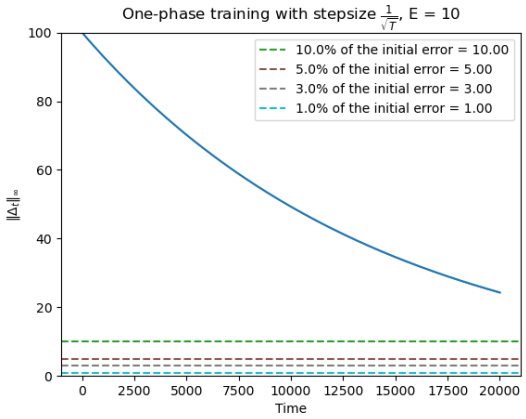


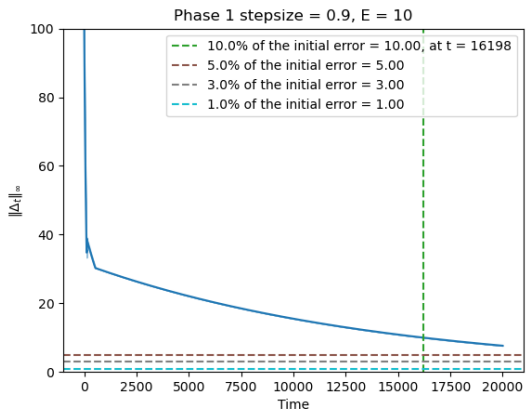
Figure 7: Final error versus T. It is clear that when choosing $\lambda = \frac{1}{\sqrt{T}}$, the final error decays as T increases.

G.3 Different target error levels.

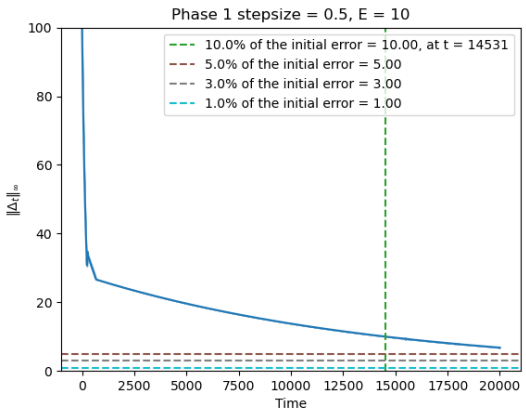
In Figure 8, we show the error levels that these training strategies can achieve within a time horizon $T = 20,000$. The tolerance levels are $10\%, 5\%, 3\%,$ and 1% of the initial error $\|\Delta_0\|_\infty$, respectively. At a high level, choosing different stepsizes for phases 1 and 2 can speed up convergence.



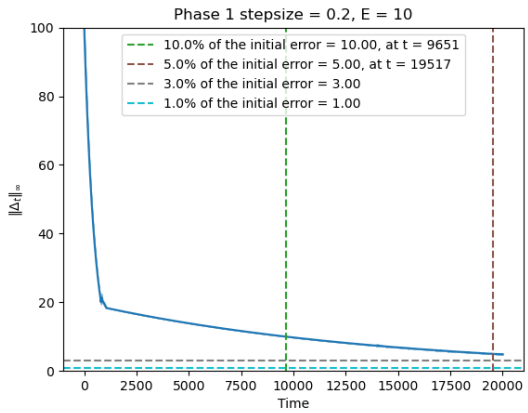
(a) One common $\lambda = \frac{1}{\sqrt{T}}$ throughout. $\|\Delta_t\|_\infty$ does meet any of the tolerance levels within 20000 iterations



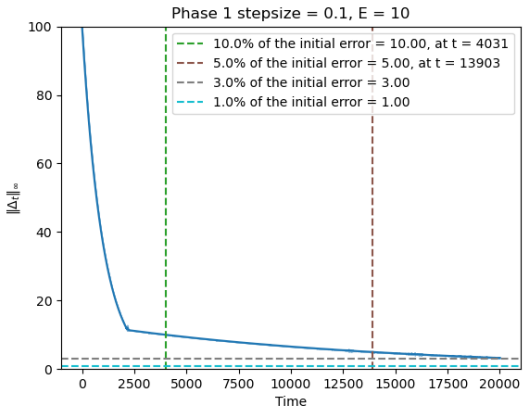
(b) With a phase 1 stepsize of 0.9, it meets the 10% tolerance level at iteration 16198.



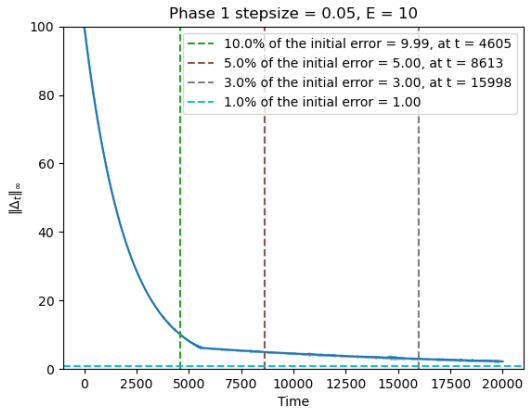
(c) With a phase 1 stepsize of 0.5, it meets the 10% tolerance level at iteration 14531.



(d) With a phase 1 stepsize of 0.2, it meets the 10% and 5% tolerance level at iterations 9651 and 19517, respectively.



(e) With a phase 1 stepsize of 0.1, it meets the 10% and 5% tolerance level at iterations 4031 and 13903, respectively.



(f) With a phase 1 stepsize of 0.05, it meets the 10%, 5%, and 3% tolerance levels at iterations 4605, 8613, and 15998, respectively.

Figure 8: Convergence performance of different tolerance levels of different stepsize choices. The horizontal dashed lines represent the tolerance levels not met, while the vertical dashed lines indicate the iterations at which the training processes meet the corresponding tolerance levels.