Event-Based Federated Q-Learning

Guener Dilsad ER [1] **Michael Muehlebach** [1]

Abstract

This paper introduces an event-based communication mechanism in federated Q-learning algorithms, enhancing convergence and reducing communication overhead. We present a communication scheme, which leverages event-based communication to update Q-tables between agents and a central server. Through theoretical analysis and empirical evaluation, we demonstrate the convergence properties of event-based QAvg, highlighting its effectiveness in federated reinforcement learning settings.

1. Introduction

Federated reinforcement learning (FedRL) enables multiple agents to collaborate on learning tasks while maintaining data privacy. However, traditional federated algorithms suffer from high communication overhead and slow convergence rates. To address these challenges, we propose an event-based communication mechanism for federated Q-learning algorithms. This mechanism reduces communication frequency by only transmitting critical updates, thus improving the efficiency and speed of the learning process.

2. Related Work

We review existing federated reinforcement learning algorithms and highlight their limitations in terms of communication efficiency and convergence speed. Additionally, we discuss recent advancements in event-based communication strategies and their potential applications in federated learning.

This paper focuses on Q-Learning (Watkins & Dayan, 1992), which aims to learn the optimal Q-function directly without estimating a model of the Markov decision process. Parallel reinforcement learning involves distributing the learning process across multiple agents or processors to speed up learning. Approaches like A3C (Asynchronous Advantage Actor-Critic) (Mnih et al., 2016) enable agents to learn in parallel while updating a global model asynchronously. This method has shown significant improvements in training times and performance.

Federated reinforcement learning extends federated learning methods to the domain of RL, by allowing multiple agents to learn collaboratively without sharing their raw data. Techniques such as FedAvg (McMahan et al., 2017) have been adapted to RL scenarios, focusing on improving the stability and efficiency of learning in heterogeneous environments. Existing federated Q-learning approaches periodically average local Q-estimates (Jin et al., 2022), whereas in our approach, the communication between agents is triggered in an event-based manner.

Event-based methods are widely used for learning dynamical systems (Solowjow & Trimpe, 2020; Umlauft & Hirche, 2019), for Bayesian optimization (Brunzema et al., 2022), and communication efficient distributed optimization (Liu et al., 2019; Singh et al., 2023; Er et al., 2024). In addition, Ornia & Mazo (2022) proposed an event-based approach to selectively share agent experience to a central learner. Inspired by the sent-on-delta concept (Miskowicz, 2006), we reduce the communication load by introducing an event-based communication strategy, such that each agent (or computational node) communicates only if necessary.

3. Problem Formulation

We model the environment heterogeneity among n environment who have the same state action pairs $\mathcal{S} \times \mathcal{A}$, reward function R but different state transitions $\{\mathcal{P}_i\}_{i=1}^n$. Each of the n agents are assumed to be located in different environments. The goal is to learn a single policy that yields high rewards on each environment and that is obtained by sharing local information.

The primary motivation for this paper is to address the communication overhead and slow convergence rates in existing FedRL methods by leveraging an event-based communication mechanism. This approach aims to reduce unnecessary data transmissions by only sharing updates when they are deemed important, thereby improving the overall efficiency

[1]Learning and Dynamical Systems Group, Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany. Correspondence to: Guener Dilsad ER <gder@tue.mpg.de>.

Workshop on Foundations of Reinforcement Learning and Control at the 41st International Conference on Machine Learning, Vienna, Austria.

of the federated learning process.

The goal of FedRL is to enable n agents to jointly learn a policy function or a value function that performs uniformly well across all environments. Due to privacy constraints, the n agents cannot share their previous trajectories. FedRL is formulated as the following optimization problem:

$$\max_{\pi} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{\sum_{t=1}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 \sim \mathcal{D}, \\ a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathcal{P}_i(\cdot \mid s_t, a_t)\} \tag{1}$$

where $\mathcal{D}$, π, and $\gamma \in (0, 1)$ represent the common initial state distribution in these n environments, the policy, and discount factor, respectively.

4. Event-Based Federated Q-Learning Algorithm

We introduce the event-based QAvg (EBQAvg) algorithm, which integrates event-based communication into federated Q-learning. QAvg is proposed by (Jin et al., 2022) which alternates between a local computation and global aggregation. Each agent performs multiple local updates of its value function before the server aggregates the value functions from all n agents. To increase communication efficiency, several local updates are executed between successive aggregations. Following the same local update and aggregation structure, the event-based QAvg algorithm enables agents to communicate Q-table updates to the central server only when significant changes occur, reducing unnecessary communication overhead. We provide a detailed description of the algorithm and its communication protocol.

Our event-based algorithm, stated in Algorithm 1, works as follows. As in classical Q learning, each agent constructs a table of size $|\mathcal{S}| \times |\mathcal{A}|$ by executing the following local update equation:

$$Q^k_{t'+1}(s,a) \leftarrow (1-\lambda_t)\, Q^k_{t'}(s,a) + \lambda_t[R(s,a) \\ + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_k(s'|s,a) \max_{a' \in \mathcal{A}} Q^k_{t'}(s',a')] \tag{2}$$

where k denotes the agent and, t and t' represents the global and local iteration number, respectively.

After several local updates, a communication event is triggered if

$$|Q^k_{t+1} - Q^k_{[t]}| > \delta \tag{3}$$

and the current Q-table is assigned to the new communicated value $Q^k_{[t+1]} \leftarrow Q^k_{t+1}$, where $Q^k_{[t]}$ denotes the value of Q^k that has been last communicated. If the communication event is not triggered by agent k, the last communicated value of its Q^k remains the same $Q^k_{[t+1]} \leftarrow Q^k_{[t]}$.

This procedure ensures that the error $e^k_t := Q^k_t - Q^k_{[t]}$ remains bounded such that

$$|e^k_t|_\infty \leq |e^k_t| \leq \delta, \tag{4}$$

at any time t for all agents $k \in \{1, \ldots, n\}$.

After agents complete their local computations and communications, the server proceeds with the global aggregation:

$$\bar{Q}_{t+1} = \frac{1}{n} \sum_{i=1}^{n} Q^i_{[t+1]}.$$

Following this, the server broadcasts the aggregated value $\bar{Q}_{t+1}$ to every agent. In the next iteration, each agent initiates its local update using the received aggregated value.

Algorithm 1 Event-Based QAvg Algorithm

Require: Number of agents n, number of rounds T, learning rate λ_t, discount factor γ, communication threshold δ, number of local updates E
Initialize Q-tables Q^k for each agent k
for $t = 1$ to T **do**
  for $k = 1$ to n **do**
    Receive of broadcasted $Q^k_t \leftarrow \bar{Q}_t$
    Perform E local updates of Q^k_t according to (2)
    Event-based send of Q^k_{t+1} to the central server $Q^k_{[t+1]} \leftarrow Q^k_{t+1}$
  end for
  Aggregate Q-tables from all agents at central server $\bar{Q}_{t+1} = \frac{1}{n} \sum_{i=1}^{n} Q^i_{[t+1]}$
  Broadcast $\bar{Q}_t$ to all agents
end for

5. Convergence Analysis

We conduct a theoretical analysis to show the effect of the communication threshold on the convergence of the event-based QAvg algorithm. QAvg is proposed as the federated version of Q-Learning. Leveraging mathematical tools such as Markov decision processes and dynamic programming principles, (Jin et al., 2022) demonstrates that QAvg converges to an optimal or near-optimal solution under certain conditions by focusing on the convergence performance of the averaged Q-function, i.e., $\bar{Q}_t$ over several environments. We extend this analysis to the event-based communication case and discuss the impact of event-based communication on convergence speed.

We know that the error between the estimate of the average Q-table $\bar{Q}_t$ and the average of Q-tables;

$$\left| \bar{Q}_t - \frac{1}{n} \sum_{k=1}^{n} Q^k_t \right| = \left| \frac{1}{n} \sum_{k=1}^{n} \left(Q^k_{[t]} - Q^k_t \right) \right|$$

is bounded by the event rule in (3). We therefore extend the analysis of QAvg for event-based communication of agents by using the estimate of the average Q-table $\bar{Q}_t$ as

$$\bar{Q}_t = \frac{1}{n}\sum_{k=1}^{n} Q_{[t]}^k = \frac{1}{n}\sum_{k=1}^{n} Q_t^k + e_t^k,$$

where e_t^k represents the error due to event-based communication of agent k at iteration t.

Theorem 5.1 states the main theoretical result that represents the trade-off between the communication threshold δ and the convergence rate of the Q-function $\bar{Q}_t$.

Theorem 5.1. *Let the constant step-size for Algorithm 1 be $\lambda_t = \alpha$ and the discount factor be γ. Let Q_* be the Q-function of the optimal policy π^* (see (1)). If the number of local updates E is chosen as $E \geq \frac{\log 2}{\alpha(1-\gamma)}$, the following inequality holds,*

$$\left|\bar{Q}_t - Q_*\right|_\infty \leq \left(\frac{1}{2}\right)^t \left|\bar{Q}_0 - Q_*\right|_\infty + 2\delta + 3\epsilon,$$

where $\bar{Q}_t$ is the average of the distributed Q functions $Q_{[t]}^k$, which are last communicated by agents $\{1,\ldots,n\}$ at iteration t, i.e., $\bar{Q}_t = \frac{1}{n}\sum_{k=1}^{n} Q_{[t]}^k$. Furthermore, δ represents the communication threshold and ϵ bounds the difference between Q_* and the locally optimal Q-functions, i.e. $|Q_*^k - Q_*| \leq \epsilon$.

Proof. See Appendix A. □

6. Empirical Evaluation

We evaluate the performance of EBQAvg in the following environments: Windy Cliff and Cart Pole. Specifically, we investigate the effect of the communication threshold on the performance and the communication load.

Windy Cliff: The Windy Cliff environment is a variation of the classic cliff walking problem (Sutton & Barto, 1998), where an agent is required to navigate from a start position to a goal while avoiding cliffs. Our experimental setup is composed of multiple Windy Cliff environments, each a modified version of the classic cliff environment by the addition of random noise for wind intensity θ blowing from the north, uniformly sampled from $[0, 1]$ (Paul et al., 2019; Jin et al., 2022). This means the agent could unintentionally move south, with a probability of $\frac{\theta}{3}$. Our experiments use a 4×4 grid map, with rewards set at 100 for reaching the goal and -100 for falling off the cliff. For the federated learning task, we set $n = 10$ agents and trained on different Windy Cliff environments. After sampling different state transitions $\{\mathcal{P}'_k\}_{k=0}^N$, we create different environments by setting $\{\mathcal{P}_k = \kappa_k \mathcal{P}'_k + \mathcal{P}'_0\}_{k=1}^N$. Therefore, environments with $\{\mathcal{P}_k\}_{k=1}^N$ will have a heterogeneity controlled by κ_k. In our experiments, $\kappa_k = 0.4$ for $k = \{1, 2\}$ and $\kappa_k = 0.8$ for $k = \{3, \ldots, 10\}$. Throughout the experiment, the discount factor is set $\gamma = 0.95$.

Figure 1 shows the evolution of objective function over episodes in the case of different communication thresholds. This visualization provides insight into how different thresholds affect the convergence behavior of the algorithm over time. In addition, Figure 2 indicates the tradeoff between objective and communication load. These figures collectively provide an understanding of the trade-off between performance and communication efficiency in the Windy Cliff environment.

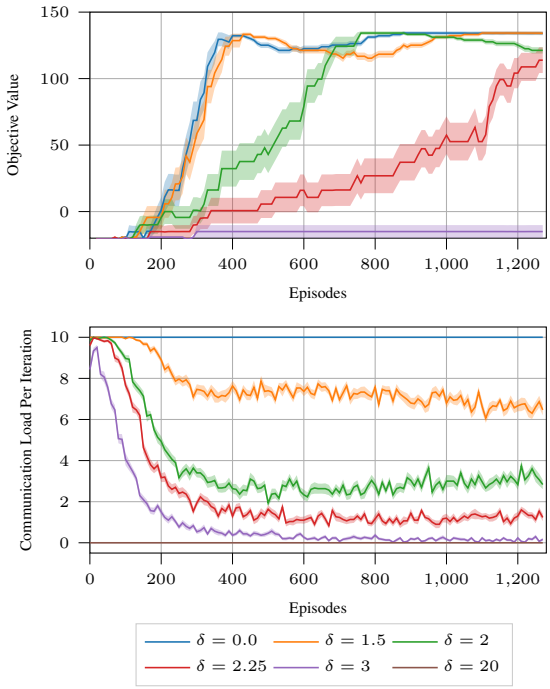

Figure 1: The top panel shows the evolution of the objective function value over episodes, for different values of communication thresholds selected in the Windy Cliff environment experiments, whereas the bottom panel shows the corresponding communication loads.

Cart Pole: The Cart Pole environment is constructed from variations of the classic Cart Pole task, which involves balancing a pole on a moving cart. The agent must apply forces to the cart to keep the pole upright while moving the cart within the bounds of the environment. In our experimental setup, we set different pole lengths for each agent, which leads to varying state transitions among agents. The pole length follows a uniform distribution $[0.5, 0.7]$ among the agents, creating a range of dynamics that the agent must learn to handle. We choose $n = 5$ agents and environments for the federated learning experiments, with a discount factor of $\gamma = 0.99$. In this case, averaged Q-functions are not Q-tables but Q-networks (Mnih et al., 2015) with two linear

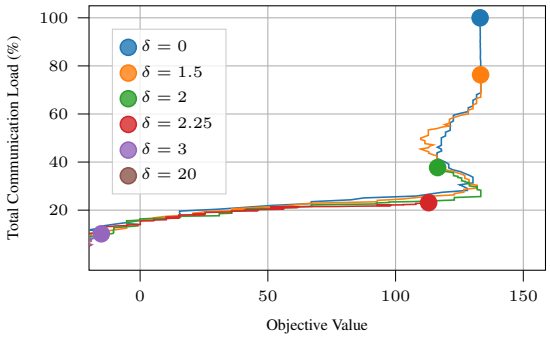

Figure 2: The figure presents the trajectory of total communication load versus objective function value for the different communication thresholds selected in the Windy Cliff environment experiments. A 100% total communication load indicates that all agents communicate at every episode.

layers with ReLU activation in between.

In the Cart Pole environment, we demonstrate, again, the performance and communication load for EBQAvg across various communication thresholds. Figure 3 presents the objective function's evolution over episodes, providing insight into the agent's learning progress in balancing the pole. Additionally, we compare our event-based approach to a random selection approach, where communicated agents are chosen randomly with a rate of ρ. Figure 4 shows the trade-off between convergence performance and total communication load in the Cart Pole environment for both the event-based and random communication schemes. The results indicate a significant reduction in communication cost with EBQAvg without compromising convergence speed.

In both Windy Cliff Walking and Cart Pole environments, our results show that EBQAvg requires fewer communication rounds to reach the same objective value. This demonstrates the efficiency of our event-based approach in environments with dynamic conditions.

7. Discussion and Future Work

In this paper, we present an event-based federated Q-learning algorithm, EBQAvg, and provide an analysis of its convergence properties. Through empirical evaluation, we demonstrate the effectiveness of EBQAvg in reducing communication overhead in federated reinforcement learning settings.

Our results highlight the potential of event-based communication in federated reinforcement learning. By strategically transmitting updates, EBQAvg effectively reduces communication overhead while maintaining performance. This has significant implications for the scalability and efficiency of federated learning systems, particularly in environments with diverse and dynamic conditions.

Future research directions include extending the event-based communication strategy to other RL algorithms and environments to validate the practicality and effectiveness of EBQAvg in large-scale federated learning scenarios. Additionally, investigating adaptive thresholds for event-triggered communication could further enhance efficiency.

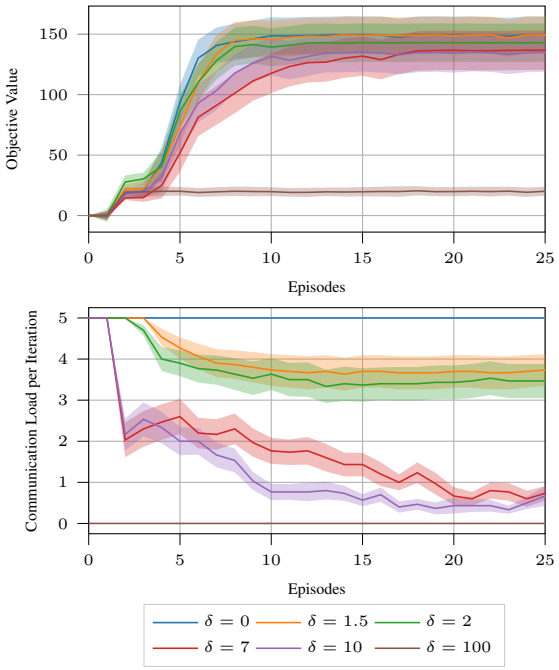

Figure 3: The top panel shows the evolution of the objective function value over episodes, for different values of communication thresholds selected in the Cart Pole environment experiments, whereas the bottom panel shows the corresponding communication loads.

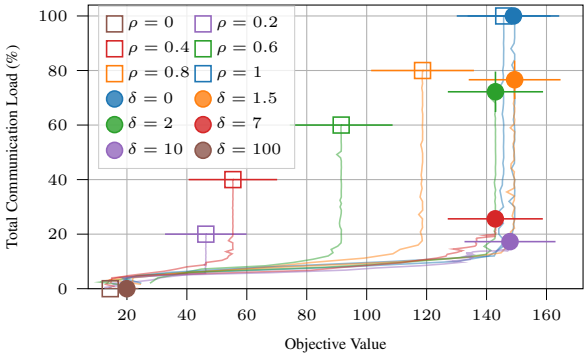

Figure 4: The figure compares different communication methods with respect to the resulting trade-off between total communication load and the objective function. Results of event-based communication are represented by circles, and random selection by squares.

References

Brunzema, P., von Rohr, A., Solowjow, F., and Trimpe, S. Event-triggered time-varying Bayesian optimization. *arXiv:2208.10790*, 2022.

Er, G. D., Trimpe, S., and Muehlebach, M. Distributed event-based learning via ADMM. *arxiv:2405.10618*, 2024.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. Federated reinforcement learning with environment heterogeneity. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 151, 2022.

Liu, Y., Xu, W., Wu, G., Tian, Z., and Ling, Q. Communication-censored ADMM for decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 67(10):2565–2579, 2019.

McMahan, H. B., Moore, E., Ramage, D., and Hampson, S. Communication-efficient learning of deep networks from decentralized data. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 54: 1273–1282, 2017.

Miskowicz, M. Send-on-delta concept: an event–based data reporting strategy. *Sensors*, 6(1):49–63, 2006.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., and et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *Proceedings of the International Conference on Machine Learning*, 48:1928–1937, 2016.

Ornia, D. J. and Mazo, M. Event-based communication in distributed Q-learning. *IEEE Conference on Decision and Control*, pp. 2379–2386, 2022.

Paul, S., Osborne, M. A., and Whiteson, S. Fingerprint policy optimisation for robust reinforcement learning. *Proceedings of the International Conference on Machine Learning*, 97:5082–5091, 2019.

Singh, N., Data, D., George, J., and Diggavi, S. SPARQ-SGD: Event-triggered and compressed communication in decentralized optimization. *IEEE Transactions on Automatic Control*, 68(2):721–736, 2023.

Solowjow, F. and Trimpe, S. Event-triggered learning. *Automatica*, 117:109009, 2020.

Sutton, R. S. and Barto, A. G. *Introduction to reinforcement learning*. MIT press Cambridge, 1998.

Umlauft, J. and Hirche, S. Feedback linearization based on Gaussian processes with event-triggered online learning. *IEEE Transactions on Automatic Control*, 65(10):4154–4169, 2019.

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

A. Proof of Theorem 5.1

We start with some preliminary definitions and lemmas.

Definition A.1. We denote the Bellman Operator in the k-th environment as:

$$\mathcal{T}_k Q(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_k\left(s' \mid s,a\right) \max_{a' \in \mathcal{A}} Q\left(s',a'\right)$$

where $R(s,a)$ is the reward received when taking action a in state s, γ is the discount factor, $\mathcal{P}_k(s' \mid s,a)$ is the transition probability of moving to state s' from state s after taking action a in the k-th environment, and $Q(s,a)$ is the action-value function.

Theorem A.2. *The Bellman operator in the k-th environment, $\mathcal{T}_k$, is a γ-contractor. For any Q_1 and Q_2, it satisfies:*

$$\left|\mathcal{T}_k Q_1 - \mathcal{T}_k Q_2\right|_\infty \leq \gamma \left|Q_1 - Q_2\right|_\infty .$$

Proof. Consider two arbitrary action-value functions Q_1 and Q_2. Let $\mathcal{T}_k$ be the Bellman operator for the k-th environment.

By definition of the Bellman operator, we have

$$\mathcal{T}_k Q_1(s,a) - \mathcal{T}_k Q_2(s,a) = \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_k(s' \mid s,a) \left(\max_{a' \in \mathcal{A}} Q_1(s',a') - \max_{a' \in \mathcal{A}} Q_2(s',a') \right).$$

The maximum operator is non-expansive and therefore we get

$$|\mathcal{T}_k Q_1(s,a) - \mathcal{T}_k Q_2(s,a)|_\infty \leq \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_k(s' \mid s,a) \max_{a' \in \mathcal{A}} |Q_1(s',a') - Q_2(s',a')|_\infty.$$

Now, using the fact that the transition probabilities form a probability distribution ($\sum_{s' \in \mathcal{S}} \mathcal{P}_k(s' \mid s,a) = 1$), and taking the supremum over all states and actions, we get:

$$|\mathcal{T}_k Q_1 - \mathcal{T}_k Q_2|_\infty \leq \gamma |Q_1 - Q_2|_\infty.$$

Thus, we have shown that for arbitrary action-value functions Q_1 and Q_2, the difference in their images under the Bellman operator is bounded by γ times the difference between the functions themselves. Therefore, $\mathcal{T}_k$ is a γ-contraction. □

The average Q-table is defined as the average of last communicated Q-tables of each agent, which is

$$\bar{Q}_t = \frac{1}{n} \sum_{k=1}^{n} Q^k_{[t]} = \frac{1}{n} \sum_{k=1}^{n} Q^k_t + e^k_t. \tag{5}$$

The local update rule in Algorithm 1 can be rewritten as

$$Q^k_{t'+1} = (1 - \lambda_t)\, Q^k_{t'} + \lambda_t \mathcal{T}_k Q^k_{t'}, \tag{6}$$

where t and t' represents the global and local iteration number, respectively. The local iterations are initialized with $Q^k_{t'}\big|_{t'=0} = \bar{Q}_t$.

Due to the environment heterogeneity, the optimal Q-functions for the individual environments are different from the optimal Q-function in (1). The following assumption quantifies this difference.

Assumption A.3. Let the optimal Q-function corresponding to (1) be denoted by Q_* and let the optimal Q-function of environment k be denoted by Q^k_*. The difference between Q^k_* and Q_* is bounded by

$$|Q^k_* - Q_*| \leq \epsilon. \tag{7}$$

Lemma A.4. *Let Assumption A.3 be satisfied and let λ_t be the step-size, γ be the discount factor for the local update step, and E be the number of local updates at each global iteration in Algorithm 1. Then the following inequality holds,*

$$|Q_{t+1}^k - Q_*|_\infty \leq e^{-\lambda_t(1-\gamma)E}|\bar{Q}_t - Q_*|_\infty + (1 + e^{-\lambda_t(1-\gamma)E})\epsilon, \tag{8}$$

for all environments $k \in \{1, \dots, n\}$, where $\bar{Q}_t$ is the estimated average broadcasted by the server after the global aggregation step at (global) iteration t and Q_* is the optimal Q-function corresponding to (1).

Proof. By subtracting Q_*^k from both sides of (6) and using the fact $\mathcal{T}_k Q_*^k = Q_*^k$, we obtain

$$Q_{t'+1}^k - Q_*^k = (1 - \lambda_t)\,(Q_{t'}^k - Q_*^k) + \lambda_t(\mathcal{T}_k Q_{t'}^k - \mathcal{T}_k Q_*^k).$$

This further implies

$$|Q_{t'+1}^k - Q_*^k|_\infty \leq (1 - \lambda_t)\,|Q_{t'}^k - Q_*^k|_\infty + \lambda_t|\mathcal{T}_k Q_{t'}^k - \mathcal{T}_k Q_*^k|_\infty.$$

The Bellman operator is a γ contractor, and therefore the following holds

$$|Q_{t'+1}^k - Q_*^k|_\infty \leq (1 - \lambda_t)\,|Q_{t'}^k - Q_*^k|_\infty + \lambda_t\gamma|Q_{t'}^k - Q_*^k|_\infty = (1 - \lambda_t(1-\gamma))|Q_{t'}^k - Q_*^k|_\infty.$$

Given the initialization of the local update with $Q_{t'}^k\big|_{t'=0} = \bar{Q}_t$, the value after E local steps, $Q_{t'}^k\big|_{t'=E} = Q_{t+1}^k$, satisfies the following inequality

$$|Q_{t+1}^k - Q_*^k|_\infty \leq (1 - \lambda_t(1-\gamma))^E|\bar{Q}_t - Q_*^k|_\infty.$$

Then, using the fact that $1 + x \leq e^x$ for any $x \in \mathbb{R}$, we get

$$|Q_{t+1}^k - Q_*^k|_\infty \leq e^{-\lambda_t(1-\gamma)E}|\bar{Q}_t - Q_*^k|_\infty.$$

We add and subtract Q_* to both sides of the equation. In addition, we apply the triangle inequality and Assumption A.3, which yields the following

$$|Q_{t+1}^k - Q_*|_\infty \leq e^{-\lambda_t(1-\gamma)E}|\bar{Q}_t - Q_*|_\infty + (1 + e^{-\lambda_t(1-\gamma)E})\epsilon.$$

□

Lemma A.5. *Let Assumption A.3 be satisfied and let λ_t be the step-size of the local update, γ be the discount factor and δ be the communication threshold for the event-based communication in Algorithm 1. Then, the following inequality holds:*

$$|\bar{Q}_{t+1} - Q_*|_\infty \leq e^{-\lambda_t(1-\gamma)E}|\bar{Q}_t - Q_*|_\infty + (1 + e^{-\lambda_t(1-\gamma)E})\epsilon + \delta.$$

where $\bar{Q}_t$ is the estimated average broadcasted by the server after the global aggregation step at (global) iteration t and Q_* is the optimal Q-function corresponding to (1).

Proof. Using (5) and (6), we have:

$$\begin{aligned}
\left|\bar{Q}_{t+1} - Q_*\right|_\infty &= \left|\frac{1}{n}\sum_{k=1}^n (Q_{t+1}^k + e_{t+1}^k) - Q_*\right|_\infty = \left|\frac{1}{n}\sum_{k=1}^n (Q_{t+1}^k - Q_*) + \frac{1}{n}\sum_{k=1}^n e_{t+1}^k\right|_\infty \\
&\leq \left|\frac{1}{n}\sum_{k=1}^n (Q_{t+1}^k - Q_*)\right|_\infty + \left|\frac{1}{n}\sum_{k=1}^n e_{t+1}^k\right|_\infty \\
&\leq \frac{1}{n}\sum_{k=1}^n \left|Q_{t+1}^k - Q_*\right|_\infty + \frac{1}{n}\sum_{k=1}^n \left|e_{t+1}^k\right|_\infty.
\end{aligned}$$

By applying Lemma A.4, substituting the bound on the error (4) and by Assumption A.3, we get:

$$|\bar{Q}_{t+1} - Q_*|_\infty \leq e^{-\lambda_t(1-\gamma)E}|\bar{Q}_t - Q_*|_\infty + (1 + e^{-\lambda_t(1-\gamma)E})\epsilon + \delta.$$

□

Lemma A.6. *Let the sequence* $|\bar{Q}_t - Q_*|_\infty \geq 0$ *satisfy*

$$|\bar{Q}_{t+1} - Q_*|_\infty \leq |\bar{Q}_t - Q_*|_\infty(1-\tilde{\alpha}) + \tilde{\beta}\tilde{\alpha}, \tag{9}$$

for all $t \geq 0$*, where the parameters* $\tilde{\alpha}$*,* $\tilde{\beta}$ *satisfy* $0 < \tilde{\alpha} < 1$ *and* $0 \leq \tilde{\beta}$*. Then, the following holds for all* $t \geq 0$*:*

$$|\bar{Q}_t - Q_*|_\infty \leq |\bar{Q}_0 - Q_*|_\infty(1-\tilde{\alpha})^t + \tilde{\beta}. \tag{10}$$

Proof. We prove the lemma by induction.

The claim holds for $t = 0$. We therefore assume that the claim holds for t and show that, as a result, the claim holds for $t+1$. More precisely,

$$\begin{aligned} |\bar{Q}_{t+1} - Q_*|_\infty &\leq |\bar{Q}_t - Q_*|_\infty(1-\tilde{\alpha}) + \tilde{\beta}\tilde{\alpha} \\ &\leq |\bar{Q}_0 - Q_*|_\infty(1-\tilde{\alpha})^{t+1} + (1-\tilde{\alpha})\tilde{\beta} + \tilde{\beta}\tilde{\alpha} \\ &\leq |\bar{Q}_0 - Q_*|_\infty(1-\tilde{\alpha})^{t+1} + \tilde{\beta}, \end{aligned} \tag{11}$$

which completes the induction argument. □

We are now, ready to prove Theorem 5.1, which we restate for the convenience of the reader.

Theorem. *Let the constant step-size for Algorithm 1 be* $\lambda_t = \alpha$ *and the discount factor be* γ*. Assume* Q_* *is the Q-function of the optimal policy* π^**. If the number of local updates* E *is chosen as* $E \geq \frac{\log 2}{\alpha(1-\gamma)}$*, the following inequality holds,*

$$\left|\bar{Q}_t - Q_*\right|_\infty \leq \left(\frac{1}{2}\right)^t \left|\bar{Q}_0 - Q_*\right|_\infty + 2\delta + 3\epsilon,$$

where $\bar{Q}_t$ *is the average of the distributed Q functions* $Q^k_{[t]}$*, which are last communicated by agents* $\{1, \ldots, n\}$ *at iteration* t*, i.e.,* $\bar{Q}_t = \frac{1}{n}\sum_{k=1}^n Q^k_{[t]}$*. Furthermore,* δ *represents the communication threshold and* ϵ *bounds the difference between* Q_* *and the optimal Q-functions.*

Proof. For a fixed step-size $\lambda_t = \alpha$, we apply Lemma A.5 to Lemma A.6 and conclude

$$|\bar{Q}_{t+1} - Q_*|_\infty \leq e^{-\alpha(1-\gamma)Et}|\bar{Q}_t - Q_*|_\infty + \frac{\delta + (1 + e^{-\alpha(1-\gamma)E})\epsilon}{1 - e^{-\alpha(1-\gamma)E}}.$$

If the number of local updates E is chosen as $E \geq \frac{\log 2}{\alpha(1-\gamma)}$, then the following inequality holds

$$\left|\bar{Q}_{t+1} - Q_*\right|_\infty \leq \left(\frac{1}{2}\right)^t \left|\bar{Q}_0 - Q_*\right|_\infty + 2\delta + 3\epsilon.$$

□