Robust Federated Q-Learning with Almost No Communication

Sreejeet Maity and Aritra Mitra

Abstract

This work studies *federated* reinforcement learning for discounted tabular Markov decision processes (MDPs) when a fraction of clients can be adversarial. There are M agents that each collect data from the same underlying MDP and communicate with a central server; however, an ε-fraction of agents may be Byzantine and can transmit arbitrary messages. The paper proposes an epoch-based federated Q-learning method, `Robust Fed-Q`, that (i) uses *variance-reduced* local Bellman backups computed from batched samples within each epoch, and (ii) aggregates client messages using a *robust* coordinate-wise median-of-means rule to tolerate Byzantine clients. The main results show that robust learning is possible with *almost no communication* (only one server round per epoch) while retaining a collaborative statistical gain: the clean part of the error scales as $\widetilde{\mathcal{O}}(1/\sqrt{MT})$, and the unavoidable adversarial contribution scales as $\widetilde{\mathcal{O}}(\sqrt{\varepsilon}/\sqrt{T})$. Experiments on a tabular gridworld validate that naive averaging can fail catastrophically under even mild Byzantine behavior, while `Robust Fed-Q` remains stable and improves with the number of honest agents.

Theoretical Results

Using epoch-wise variance-reduced Bellman messages and coordinate-wise MoM aggregation, `Robust Fed-Q` learns $Q^\star$ reliably even when an ε-fraction of clients are adversarial.

- **Collaboration Gain + Diminishing Corruption Penalty.**
  With high probability,

  $$\|Q_K - Q^\star\|_\infty \;\le\; \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{MT}}\right) + \tilde{\mathcal{O}}\left(\frac{\sqrt{\varepsilon}}{\sqrt{T}}\right),$$

  so the clean term enjoys a $1/\sqrt{M}$ federated speedup while the unavoidable corruptive term diminishes with T.

- **$\tilde{\mathcal{O}}(1)$ communication.** Each client uploads one message per epoch and the server broadcasts once per epoch, so the number of communication rounds is

  $$K = \mathcal{O}(\log(MT)),$$

  rather than $\Theta(T)$.

Experiments

Figure 1. $E_K = \|Q_K - Q^*\|_\infty$ for $M = 1000$ and varying ε as a function of the number of epochs K, where the central server simply averages the agent updates.

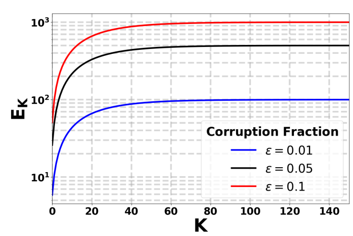
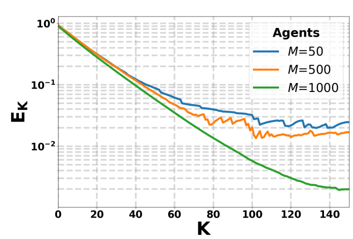


Figure 2. Plots of E_K for `Robust Fed-Q`, with corruption fraction $\varepsilon = 0.1$ and varying agent counts. Establishes **collaboration gain + robustness**.



Overview

Federated reinforcement learning (FRL) aims to exploit many geographically distributed agents (devices, robots, simulators) that interact with *similar* environments, so that each agent can learn faster by sharing information rather than raw trajectories. In practice, large-scale federated deployments are vulnerable to unreliable participants: some clients may be faulty, noisy, or even malicious. This motivates a *Byzantine-robust* FRL formulation in which an adversary can corrupt an ε-fraction of clients and send arbitrary updates to the server.

The paper focuses on a tabular discounted MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ and asks:

- **Learnability under Byzantine clients:** Can one still obtain a vanishing $\|Q - Q^\star\|_\infty$ error as the number of samples grows?
- **Collaboration gain:** If learning is possible, can one achieve the same improvement as in clean federated learning, namely a $1/\sqrt{M}$ reduction in statistical error due to M parallel data sources?
- **Communication efficiency:** Can this be done with *few* server rounds, instead of communicating after every environment step?

A key difficulty is that standard federated averaging is extremely fragile: a single adversarial client can inject large-magnitude messages that dominate the mean and derail learning. The central idea of this paper is to combine **epoching** (to form concentrated, low-variance client messages) with **robust aggregation** (to prevent Byzantine outliers from influencing the update).

Main Results

■ **Setting.** There are M agents connected to a central server. Time is partitioned into K epochs of length H, so each agent collects $T = KH$ samples. The goal is to return an estimate Q_K of the optimal action-value function $Q^\star$, measured in $\|\cdot\|_\infty$:

$$e_K := \|Q_K - Q^\star\|_\infty.$$

An ε-fraction of clients are Byzantine: they can observe the protocol and transmit arbitrary messages to the server in every epoch.

■ **Algorithm: `Robust Fed-Q`.** Each epoch consists of one *local computation + upload* step and one *server aggregation + broadcast* step.

(1) Local variance-reduced Bellman message. Within epoch k, each honest client i uses its H samples to form an empirical estimate of the Bellman backup at the current global iterate Q_k. Concretely, the client computes a message $d_{i,k}$ intended to approximate $(\mathcal{T}Q_k)(s,a)$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$, where

$$(\mathcal{T}Q)(s,a) \;=\; \mathcal{R}(s,a) \;+\; \gamma\, \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[\max_{a'} Q(s', a') \right].$$

Using a batch of size H reduces the stochastic noise in $d_{i,k}$ compared to a single-sample temporal-difference target.

(2) Robust server aggregation across clients. For each coordinate $(s, a) \in \mathcal{S} \times \mathcal{A}$, the server aggregates $\{d_{i,k}(s, a)\}_{i=1}^{M}$ using a coordinate-wise median-of-means rule. MoM remains accurate even if an ε-fraction of inputs are arbitrarily corrupted, provided ε is below a constant and the honest inputs are sufficiently concentrated.

(3) Global update and broadcast. The server performs a convex combination update

$$Q_{k+1}(s, a) \;=\; (1-\alpha)\, Q_k(s, a) \;+\; \alpha\, \widetilde{d}_k(s, a),$$

where $\widetilde{d}_k$ is the robustly aggregated message, and broadcasts Q_{k+1} to all clients. Thus, the number of communication rounds is K (one per epoch), rather than T (one per interaction step).

■ **High-probability Error Bound.** With suitable choices of step size α and epoch count K (typically logarithmic in T up to $(1-\gamma)$ factors), the paper proves that with high probability,

$$\|Q_K - Q^\star\|_\infty \;\le\; \underbrace{\widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{MT}}\right)}_{\text{Clean statistical term scaling with M}} \;+\; \underbrace{\widetilde{\mathcal{O}}\left(\frac{\sqrt{\varepsilon}}{\sqrt{T}}\right)}_{\text{Byzantine penalty}} .$$

The two terms have complementary meanings:

- **Collaboration gain persists:** in the absence of Byzantine clients ($\varepsilon = 0$), the error scales as $\widetilde{\mathcal{O}}(1/\sqrt{MT})$, matching the intuition that M agents provide M times more data.
- **Robustness penalty is unavoidable:** when $\varepsilon > 0$, the additional $\widetilde{\mathcal{O}}(\sqrt{\varepsilon}/\sqrt{T})$ term captures the irreducible effect of adversarial participation. Importantly, this term does not vanish with M at the same rate as the clean term, reflecting that more agents also create more potential adversarial channels.
- **Communication is epoch-level:** the protocol needs only K server rounds, which can be far smaller than T in long-horizon learning.