Corruption-Tolerant Asynchronous Q –Learning with Near-Optimal Rates

Sreejeet Maity Aritra Mitra

Department of Electrical and Computer Engineering

North Carolina State University, Raleigh

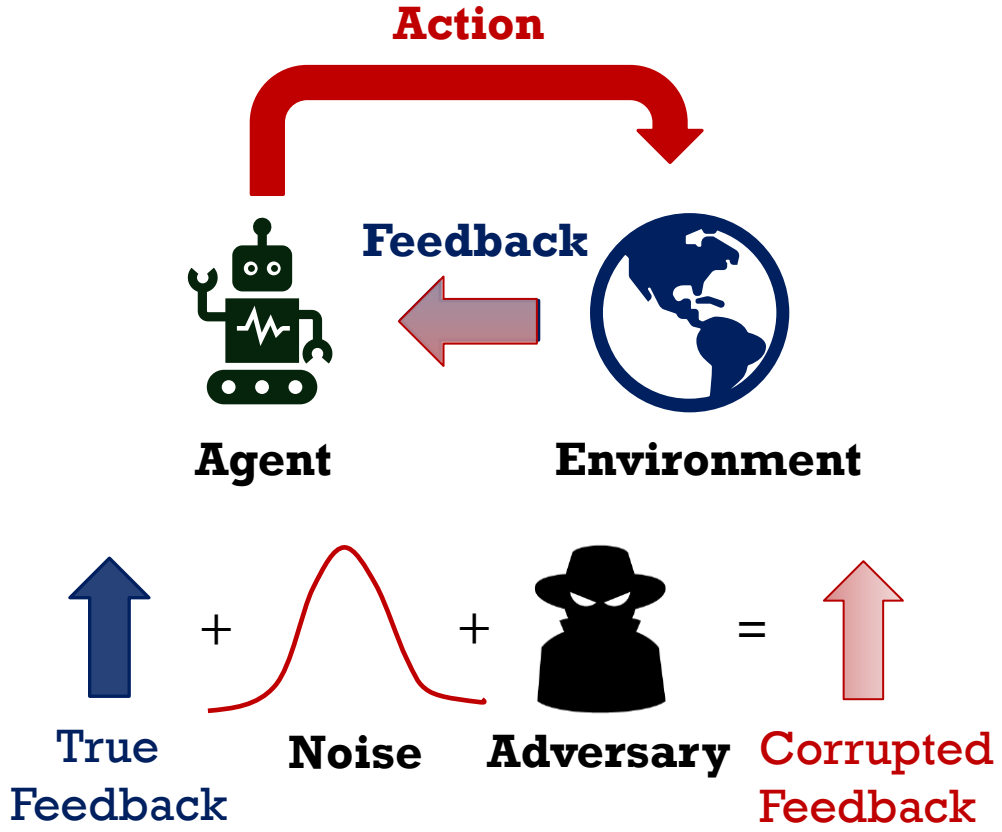
ICML

International Conference
On Machine Learning

NC STATE UNIVERSITY

Are RL algorithms *always* trustworthy ?

- RL algorithms learn from feedback.
- How critical is the assumption of "**perfect**" feedback in practical scenarios?
- In real-world settings, feedback can be **noisy**, contain **outliers**, or even be maliciously **corrupted**.



How "reliable" are RL Algorithms in such real-world settings ?

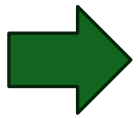
Q – Learning: Estimating Action Values

- We consider an MDP $\mathcal{M} = (S, A, P, R, \gamma)$ with finite state and action spaces.
- $P(s'|s,a)$ is the probability of transitioning from s to s' under action a.
- $R(s,a)$ is the finite, expected reward at state-action pair (s,a).
- Policy $\pi: S \to \Delta(A)$.

For a fixed policy π, we define $Q_\pi: S \times A \mapsto \mathbb{R}$:

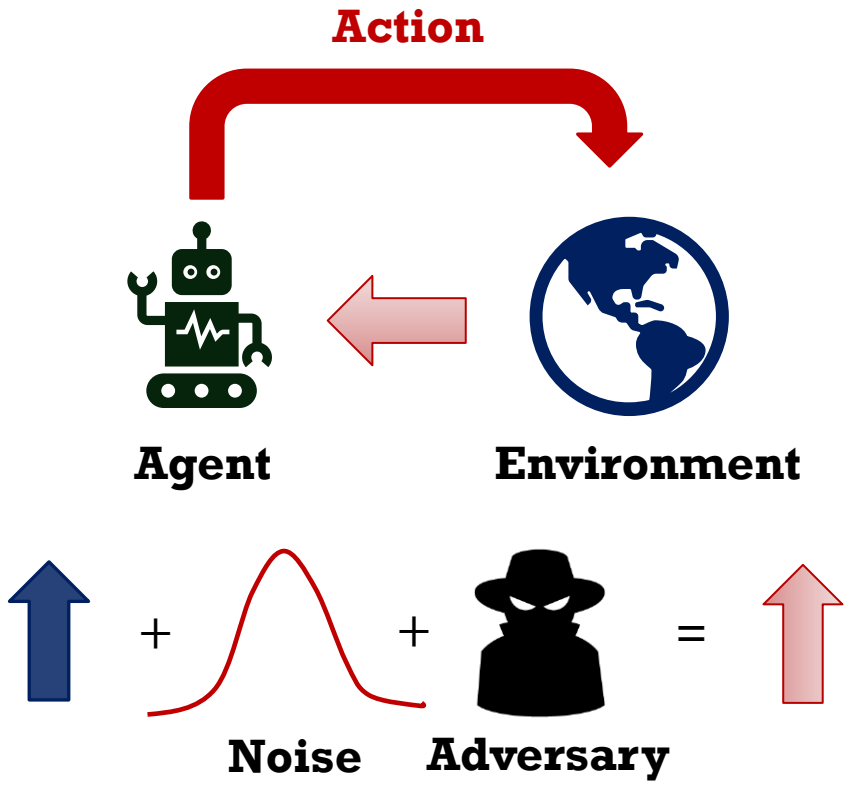
$$Q_\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s\,, a_0 = a, \pi\right].$$

$$Q^*(s,a) = \max_\pi Q_\pi(s,a);\ \pi^*(s) = \arg\max_{a \in A} Q^*(s,a)\,.$$



How to find Q^* ? (Watkins et al., 1988)

Q – Learning: Estimating Action Values



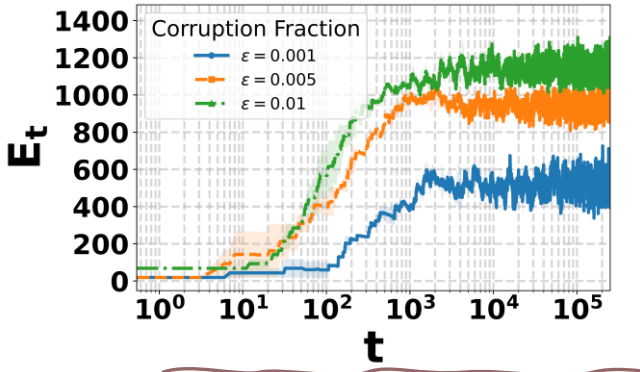
Vanilla Q –Learning with ***corrupted feedback***

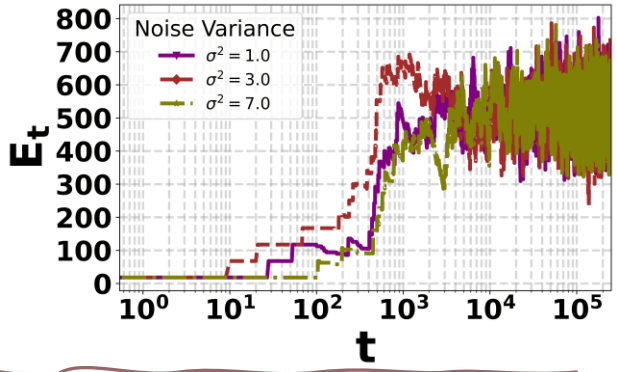
Corruption Model: At each time-step t, if learner is in state s_t and plays action a_t, it observes:

$$y_t(s_t, a_t) = \begin{cases} R(s_t, a_t) + n_t & \text{with probability } (1-\varepsilon) \\ (*) & \text{with probability } \varepsilon \end{cases}$$

n_t: Heavy-tailed noise, finite variance; $(*)$: Arbitrary.

- **New observation sequence:** $\mathcal{D}_t = \{s_t, a_t, y_t(s_t, a_t), s_{t+1}\}$.





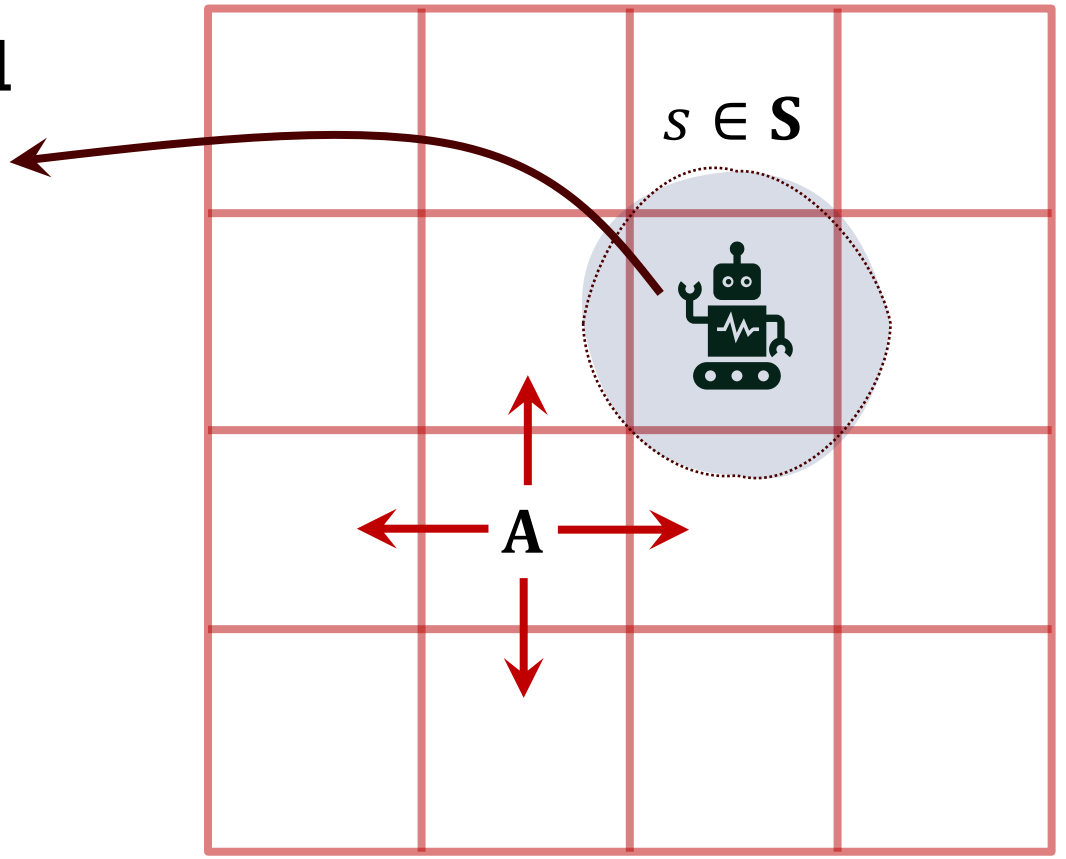
Vanilla Q –Learning [1] fails **catastrophically!**

[1] C. J. Watkins and P. Dayan, "Q-learning," Machine learning, 1992.

Towards Robust Q – Learning

Idea 1: Use historical data to build robust estimates of reward means. What about rare events?

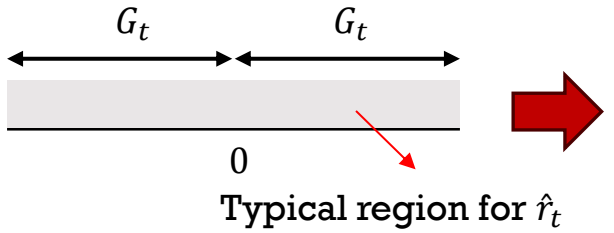
Idea 2: Reject estimates that deviate "too much" from expected bounds. How to design rejection threshold?



Components of Robust Q – Learning

- Q_0
- Step-size α
- Failure prob. δ
- Corruption frac. ε
- Total iterations T

1. Observe $y_t(s_t, a_t)$.
2. Append $y_t(s_t, a_t)$ to buffer $\mathcal{F}_t(s_t, a_t)$.
3. Compute $\hat{r}_t$ = TRIM $[\mathcal{F}_t(s_t, a_t), \delta^*, \varepsilon]$.



If $|\hat{r}_t| > G_t := c\bar{\sigma}\left(\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\lambda t}} + \sqrt{\varepsilon}\right) + \bar{\sigma}$, set $\hat{r}_t = 0$.

$$|\hat{r}_t - R(s_t, a_t)| \leq c\bar{\sigma}\left(\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{\lambda t}} + \sqrt{\varepsilon}\right)$$

Corruption fraction

$\bar{\sigma} = \max(\sigma, \bar{R})$, where σ^2 is noise variance and $|R(s,a)| \leq \bar{R}, \forall s, a$.

Minimum visitation probability

Robust Q-Learning Update

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t\left(\tilde{r}_t + \gamma \max_{a' \in A} Q_t(s_{t+1}, a')\right)$$

Convergence Guarantee of Robust Q – Learning

Theorem: Given any $\delta \in (0,1)$, Robust Q guarantees the following with probability at least $1-\delta$:

$$\|Q_T - Q^*\|_\infty \leq \widetilde{\mathcal{O}}\underbrace{\left(\frac{\bar{\sigma}}{(1-\gamma)^{2.5}\lambda^{1.5}\sqrt{T}}\right)}_{(*)} + \underbrace{c \cdot \left(\frac{\bar{\sigma}\sqrt{\varepsilon}}{\lambda(1-\gamma)}\right)}_{(**)}$$

- ❑ The term $(*)$ matches known rates for vanilla Q-learning in the absence of corruption (Wainwright 2018, Qu, 2019, Li et al., 2024).

- ❑ The term $(**)$ captures the effect of data-corruption.

Key point: It only depends on corruption fraction, not corruption magnitude.

Information-Theoretic Fundamental Limits

MDP with **one state** and **one action.** → **R**obust **M**ean **E**stimation in Rewards.

$$|Q_1 - Q_2| \geq \frac{C\bar{R}\sqrt{\varepsilon}}{2(1-\gamma)}$$

Theorem: (Fundamental Limit) There exist two different instances of our problem such that on at least one instance, the estimation error of **any** algorithm, given **any** number of samples, will be at least on the order of $\left(\frac{\sigma\sqrt{\varepsilon}}{(1-\gamma)}\right)$ with probability $\frac{1}{4}$.

Extensions

➢ Extension 1 (Theorem 4)

- ❑ Can we get similar guarantees, when we have no prior knowledge about the reward statistics? **Yes**!
- ❑ Use $m(t) = t^p$, as a proxy for $\overline{\sigma}$ while designing the threshold function G_t.

Key Tool: A variant of Azuma-Hoeffding Inequality for martingale differences that satisfy a coarse bound deterministically, and a finer bound w.h.p.

➢ Extension 2 (Theorem 5)

- ❑ Extending our analysis to the general **Markovian** sampling.
- ❑ Use "**Blocking Technique**": subsample trajectory, and use coupling arguments.
- ❑ The i.i.d analysis follows with the sub-sampled trajectory, and gives a similar bound with an inflation parameter τ, commensurate with the mixing-time of the underlying Markov chain.

Summary and Related Works

- Vanilla Q-Learning can fail catastrophically even under a small fraction of corrupted rewards.
- Our robust Q-Learning algorithm filters corrupted feedback and achieves near-optimal convergence rates.
- The final error depends on the **corruption fraction** ε, not on the magnitude of corrupted rewards.
- We prove a nearly matching lower bound, showing that the $\sqrt{\varepsilon}$ corruption dependence is unavoidable.

First tight finite-sample analysis of Q-Learning under adversarial reward feedback, with matching upper bound and lower bound.

Related Works:

[1] *Robust Q − Learning under Corrupted Rewards*, S. Maity and A. Mitra, **IEEE CDC** 2024.

[2] *Adversarially-Robust TD Learning with Markovian Data*, S. Maity and A. Mitra, **AISTATS** 2025.

[3] Robust Federated Q −Learning with Almost No Communication, S. Maity and A. Mitra, **IEEE ACC** 2026.