On Reinforcement Learning with Adversarial Corruptions and Applications to Block MDP

Tianhao Wu, Yunchang Yang, Simon S. Du, Liwei Wang

ICML 2021

Presenter: Zhengguang Zhang

November 4, 2021

Outline

Outline

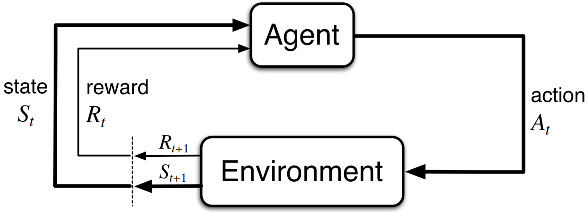
Background

Reinforcement Learning (RL) is ubiquitous for decision-making.

- Agent interacts with the environment based on observations (states, actions, rewards)
- Maximize the cumulative reward through time

Background

Interactions between agent and environment



Example: autonomous driving

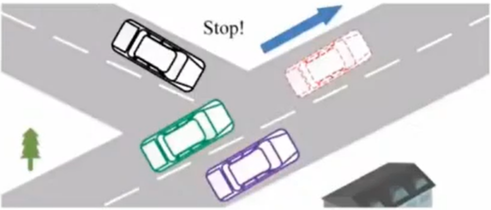
- State: position, velocity, traffic lights, congestion, accidents
- Action: direction, acceleration
- Reward: Energy consumption, safety, comfortability

Motivation

- The truthfulness of the observed state and reward is crucial
- False state and reward observation due to:
  - Non-stationary behaviour
  - Errors in the system
  - Malicious corruption by adversary
- Various threats (efficiency, safety)

Motivation

- The adversary makes corruptions for different purposes:
  - Selfish purpose: e.g. claims false position to clear its lane
  - Malicious purpose: e.g. changes the traffic light to cause congestion or collision



- Question: How to guarantee the safety and robustness of the agent against data corruption?

Related Works

- MAB with Corruption
  - Corrupted rewards, corruption level is unknown, upper bound and lower bound of the regret [Lykouris et al., 2018]
  - Improves above upper bound, and claims an upper bound when corruption level is known [Gupta et al., 2019]
- Episodic RL
  - Bandit feedback and unknown transition, adversarial rewards, upper bound on regret [Jin et al., 2019]
  - Corrupted rewards and transitions of selected episodes, corruption level is unknown [Lykouris et al., 2019]

$$\text{Regret} = \tilde{O}(C\sqrt{SAHK} + CS^2A + C^2SA)$$

Problem: Vacuous when C is large, e.g., when $C = O(\sqrt{K})$, bound grows linearly with respect to K

Stronger Real-world Corruption

- Adversary makes decision of corrupting or not after the agent takes an action at each step (more information $\rightarrow$ stronger)
- Adversary disturbs agent's observation on the state and reward signal, while leaves the underlying state and reward unchanged.
- Example: The robot player receives images with adversarial perturbations, while the true environment remains unchanged

Contributions

- Propose an algorithm that can achieve $\tilde{O}(\sqrt{SAK} + CSA)$ regret[1] when the corruption level C is known
- Prove the lower bound $\Omega(\sqrt{SAK} + CSA)$ with known C, $\Omega(C^{\alpha}K^{\beta})$ with unknown C
- Apply to Block MDP setting and obtain the first algorithm with $\sqrt{K}$-type regret

[1] $\tilde{O}$ hides the logarithmic factor

Contributions

- Propose an algorithm that can achieve $\tilde{O}(\sqrt{SAK} + CSA)$ regret[1] when the corruption level C is known
- Prove the lower bound $\Omega(\sqrt{SAK} + CSA)$ with known C, $\Omega(C^{\alpha}K^{\beta})$ with unknown C
- Apply to Block MDP setting and obtain the first algorithm with $\sqrt{K}$-type regret

[1] $\tilde{O}$ hides the logarithmic factor

Contributions

- Propose an algorithm that can achieve $\tilde{O}(\sqrt{SAK} + CSA)$ regret[1] when the corruption level C is known
- Prove the lower bound $\Omega(\sqrt{SAK} + CSA)$ with known C, $\Omega(C^{\alpha}K^{\beta})$ with unknown C
- Apply to Block MDP setting and obtain the first algorithm with $\sqrt{K}$-type regret

[1] $\tilde{O}$ hides the logarithmic factor

Outline

Episodic MDP

- Finite-horizon MDP: $M = (\mathcal{S}, \mathcal{A}, H, P, R)$
- Known state space $\mathcal{S}$, action space $\mathcal{A}$
- Unknown distribution for transition P and reward R
- K episodes, each of H steps
- At episode $k = 1, \ldots, K$:
  - Initial state i.i.d from fixed distribution, i.e., $s_1^k \sim \mu$
  - Agent commits to policy: $\pi^k = \{\pi_h^k \mid \pi_h^k : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$
  - At step $h = 1, \ldots, H$:
    - Agent takes action $a_h^k \sim \pi^k(s_h^k)$
    - Agent receives reward $r_h^k \sim R(s_h^k, a_h^k)$
    - Transits to state $s_{h+1}^k \sim P(\cdot \mid s_h^k, a_h^k)$
  - Observes state-action-reward trajectory $(s_h^k, a_h^k, r_h^k)_{h=1}^H$

Episodic MDP with Corruption

- Finite-horizon MDP: $M = (\mathcal{S}, \mathcal{A}, H, P, R)$
- Known state space $\mathcal{S}$, action space $\mathcal{A}$
- Unknown distribution for transition P and reward R
- K episodes, each of H steps
- At episode $k = 1, \ldots, K$:
    - Initial state i.i.d from fixed distribution, i.e., $s_1^k \sim \mu$
    - Agent commits to policy: $\pi^k = \{\pi_h^k \mid \pi_h^k : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$
    - At step $h = 1, \ldots, H$:
        - Agent takes action $a_h^k \sim \pi^k(s_h^k)$
        - Adversary decides whether to corrupt
        - if yes: corrupts current reward r_h^k with arbitrary $(r_h^k)'$, generates arbitrary next state $(s_{h+1}^k)'$ and corresponding reward function $\tilde{r}((s_{h+1}^k)', \cdot) \in R^S$
        - Otherwise, normal episodic MDP

Zoom in one Episode

Remove Dependency on k

1. **Time step h:** The agent takes the action a_h at state s_h
2. The adversary decides whether to corrupt current reward r_h and next state s_{h+1}.
3. If the adversary decides to corrupt, it generates arbitrary reward r'_h, next state s'_{h+1}, and corresponding reward function $\tilde{r}(s'_{h+1}, \cdot) \in R^S$
4. **Time step $h+1$:**The agent observes the corrupted state s'_{h+1}, it takes action a_{h+1} and observes $\tilde{r}(s'_{h+1}, a_{h+1})$ instead of $r(s_{h+1}, a_{h+1})$.

Setting of This Paper

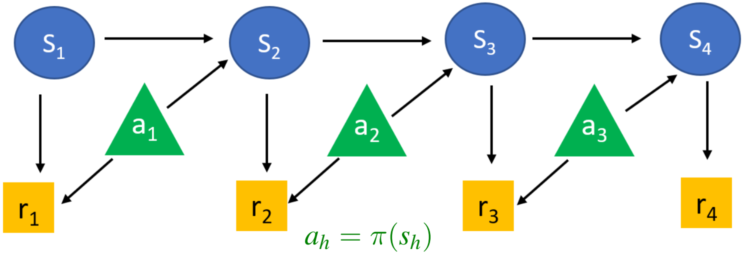
Denote state $\tilde{s}_h$ and reward $\tilde{r}_h$ observed by the agent

- $\tilde{s}_h = s'_h$ when corrupted, and $\tilde{s}_h = s_h$ when no corruption
- $\tilde{r}_h = r'_h$ when corrupted, and $\tilde{r}_h = r_h$ when no corruption
- The underlying state and reward are always s_h and r_h
- Corruption level C: The number of time steps that is corrupted in each episode.

Assumption 1 (Bounded Total Reward).

The reward r_h satisfied that $r_h \geq 0$ for all $h \in [H]$. Moreover, for all policy π, $\sum_{h=1}^{H} r_h \leq 1$ almost surely.

Value Functions without Corruptions

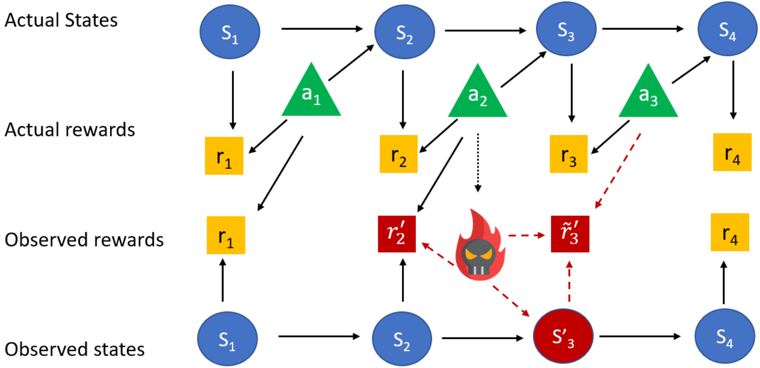


$$V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{i=h}^{H} r(s_i, \pi(s_i)) \mid s_h = s\right] \xrightarrow{\pi^*} V^*(s)$$

$$Q_h^\pi(s,a) = \mathbb{E}_\pi\left[r(s,a) + \sum_{i=h+1}^{H} r(s_i, \pi(s_i)) \mid s_h = s, a_h = a\right]$$

MDP with Corruption

$$a_3 = \pi(s_3') \Rightarrow \begin{cases} r_3 = r(s_3, \pi(s_3')) & \text{actual reward agent received} \\ \tilde{r}_3' = \tilde{r}(s_3', \pi(s_3')) & \text{reward agent observed} \end{cases}$$



Value Functions with Corruptions

Let $\tilde{Q}, \tilde{V}$ be the rewards the agent actually receives under corruption. Rewards are calculated by environment based on underlying state and agent's action under corruption.

$$\tilde{V}_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{i=h}^{H} r(s_i, \pi(\tilde{s}_i)) \mid s_h = s\right]$$

$$\tilde{Q}_h^\pi(s, a) = \mathbb{E}_\pi\left[r(s, a) + \sum_{i=h+1}^{H} r(s_i, \pi(\tilde{s}_i)) \mid s_h = s, a_h = a\right]$$

$$\mathrm{Regret}(K) = \sum_{k=1}^{K} V_1^*(s_1^k) - \tilde{V}_1^{\pi^k}(s_1^k)$$

Outline

Unbiased Empirical Estimator

Number of visits (omit k for simplicity):

- $\hat{N}^k(s,a,s') \to \hat{N}(s,a,s')$
- $\hat{N}^k(s,a) \to \hat{N}(s,a)$

Transition dynamics:

$$\hat{P}_{s,a}(s') = \hat{P}(s' \mid s,a) = \frac{\hat{N}(s,a,s')}{\hat{N}(s,a)}$$

$$\hat{Q}_h(\hat{N},\hat{P})(s,a) = \hat{r}(s,a) + \hat{P}_{s,a}V_{h+1} + \hat{b}_h(s,a)$$

Biased Empirical Estimator Due to Corruption

Number of visits (omit k for simplicity):

- $\tilde{N}^k(s,a,s') \to \tilde{N}(s,a,s')$
- $\tilde{N}^k(s,a) \to \tilde{N}(s,a)$

$$|\hat{N}(s,a,s') - \tilde{N}(s,a,s')| \le C$$

$$|\hat{N}(s,a) - \tilde{N}(s,a)| \le C$$

Transition dynamics:

$$\tilde{P}_{s,a}(s') = \hat{P}(s' \mid s,a) = \frac{\tilde{N}(s,a,s')}{\tilde{N}(s,a)}$$

Logic Behind CR-MVP

Optimism in the face of uncertainty

- Typical approach: maintains an optimistic estimation of Q-function by adding a bonus term to the empirical Bellman Equation

$$\hat{Q}_h(s,a) = \hat{r}(s,a) + \hat{P}_{s,a}V_{h+1} + \hat{b}_h(s,a)$$

- Problem: Relies on the access to the unbiased estimators $\hat{N}$ and $\hat{P}$, which are unavailable in the corrupted setting.
- Solution:

$$\begin{aligned} Q_h(\tilde{N},\tilde{P})(s,a) &= \tilde{r}(s,a) + \tilde{P}_{s,a}V_{h+1} + \tilde{b}_h(s,a) \\ \geq \hat{Q}_h(\hat{N},\hat{P})(s,a) &= \hat{r}(s,a) + \hat{P}_{s,a}V_{h+1} + \hat{b}_h(s,a) \end{aligned}$$

Design of Bonus Term

Lemma 1

Suppose $c_1, c_2, c_3 \geq 0$, let $\tilde{b}_h = \tilde{b}_{h,con} + \tilde{b}_{h,bia}$, then $Q_h \geq \hat{Q}_h$

$$\begin{aligned}\tilde{b}_{h,bia} =& 2\min\left\{\frac{2C}{|\tilde{N}-C|}, 1\right\} \\ &+ (c_1 + c_2)\min\left\{\frac{\sqrt{C\iota}}{|\tilde{N}-C|}, 1\right\}.\end{aligned}$$

$$\begin{aligned}\tilde{b}_{h,con} =& c_1 \min\left\{\sqrt{\frac{\mathbb{V}(\tilde{P}, V_{h+1})\iota}{|\tilde{N}-C|}}, 1\right\} \\ &+ c_2 \min\left\{\sqrt{\frac{\tilde{r}\iota}{|\tilde{N}-C|}}, 1\right\} + c_3 \min\left\{\frac{\iota}{|\tilde{N}-C|}, 1\right\},\end{aligned}$$

CR-MVP Algorithm

Algorithm 1 Corruption Robust Monotonic Value Propagation

Input: C is the corruption level.
for $k = 1, 2, ..., K$ **do**
  for $h = 1, 2, ..., H$ **do**
    Observe s_h^k, take action $a_h^k = \arg\max_a Q_h(s_h^k, a)$;
    Receive reward r_h^k and next state s_{h+1}^k.
    Update empirical estimate $\tilde{P}_{s,a,\cdot} \leftarrow \tilde{N}_{s,a,\cdot}/\tilde{N}(s,a)$, and $\tilde{r}(s,a)$.
    for $h = H, H-1, ..., 1$ **do**
      for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
        Set confidence bonus term $\tilde{b}_h$.
        $Q_h(s,a) \leftarrow \min\{\tilde{r}(s,a) + \tilde{P}_{s,a} V_{h+1} + \tilde{b}_h(s,a), 1\}$,
        $V_h(s) \leftarrow \max_a Q_h(s,a)$.
      end for
    end for
  end for
end for

Regret Upper Bound of CR-MVP

By setting $\tilde{b}_h$ as in Lemma 1:

Theorem 1

With probability at least $1-\delta$ the regret of CR-MVP satisfies:

$$\mathrm{Regret}(K) \leq O(\sqrt{SAK} + S^2A + CSA),$$

where K is the total number of episodes. In other words, the regret caused by the corruptions only scales linearly with regard to C.

Proof Sketch

Lemma 4. *For any vector* $V \in R^S$, $V(s) \in [0,1]$ *for any* $s \in S$, *it holds that*

$$||\tilde{P}_{s,a} - \hat{P}_{s,a}||_1 \leq 2\min\{\frac{C}{|\tilde{n}(s,a) - C|}, 1\},$$

$$|\mathbb{V}(\tilde{P}_{s,a}, V) - \mathbb{V}(\hat{P}_{s,a}, V)| \leq 6\min\{\frac{C}{|\tilde{n}(s,a) - C|}, 1\},$$

$$|\tilde{r} - \hat{r}| \leq \min\{\frac{C}{|\tilde{n} - C|}, 1\}.$$

Proof Sketch

Bounding Bellman Error

Lemma 5. *With probability* $1 - 3S^2AH(\log_2(KH)+1)\delta$, *for any* $1 \le k \le K$, $1 \le h \le H$ *and* (s,a), *it holds that*

$$Q_h^k(s,a) - r(s,a) - P_{s,a}V_{h+1}^k$$
$$\le \min\{2\tilde{b}_h^k(s,a) + \frac{2C}{\tilde{n}(s,a)+C} + \sqrt{\frac{2\mathbb{V}(P_{s,a},V_{h+1}^*)\iota}{\hat{n}^k(s,a)}} + \sqrt{\frac{2S\mathbb{V}(P_{s,a},V_{h+1}^k - V_{h+1}^*)\iota}{\hat{n}^k(s,a)}} + \frac{2S\iota}{3\hat{n}^k(s,a)}, 1\}.$$

Proof Sketch

Regret Analysis

$$
\begin{aligned}
\mathrm{Regret}(K) &:= \sum_{k=1}^{K}(V_1^*(s_1^k) - \tilde{V}_1^{\pi^k}(s_1^k)) \\
&\le \sum_{k=1}^{K}(V_1^k(s_1^k) - \tilde{V}_1^{\pi^k}(s_1^k)) \\
&= \sum_{k=1}^{K}(\bar{V}_1^k(s_1^k) - \tilde{V}_1^{\pi^k}(s_1^k)) \\
&= \sum_{k=1}^{K}(\bar{V}_1^k(s_1^k) - \sum_{h=1}^{H}\bar{r}_h^k) + \sum_{k=1}^{K}(\sum_{h=1}^{H}\bar{r}_h^k - \tilde{V}_1^{\pi^k}(s_1^k)) \\
&= \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k}\bar{V}_{h+1}^k - \bar{V}_{h+1}^k(s_{h+1}^k)) + \sum_{k=1}^{K}\sum_{h=1}^{H}(\bar{V}_h^k(s_h^k) - \bar{r}_h^k - P_{s_h^k,a_h^k}\bar{V}_{h+1}^k) + \sum_{k=1}^{K}(\sum_{h=1}^{H}\bar{r}_h^k - \tilde{V}_1^{\pi^k}(s_1^k)) \\
&\le \sum_{k=1}^{K}\sum_{h=1}^{H}(P_{s_h^k,a_h^k}\bar{V}_{h+1}^k - \bar{V}_{h+1}^k(s_{h+1}^k)) + \sum_{k=1}^{K}\sum_{h=1}^{H}\bar{\beta}_h^k(s_h^k,a_h^k) + \sum_{k=1}^{K}(\sum_{h=1}^{H}\bar{r}_h^k - \tilde{V}_1^{\pi^k}(s_1^k)) + |\mathcal{K}^C|.
\end{aligned}
$$

Lower Bound

Theorem 2

For any fixed C, A, and any algorithm , there exists an episodic MDP, such that the regret incurred after K episodes is at least $\Omega(CSA)$, where K satisfies $K \geq 2CSA$.

Special case: MAB

$S = 1, H = 1, C$ is the number of episodes being corrupted

- If an algorithm visit all arms for at least C times, then directly lead to a $\Omega(CA)$ regret.
- If the number of visit of arm i is less than C times, directly lead to a $\Omega(K)$ regret.

Proposition 1

In an MAB instance with adversarial corruptions, assume that the corruption level C is unknown. If there exists an algorithm that can achieve a high probability regret upper bound $\tilde{O}(\sqrt{K} + C^{\alpha}K^{\beta})$ for any C and K, then $\alpha + \beta/2 \geq 1$.

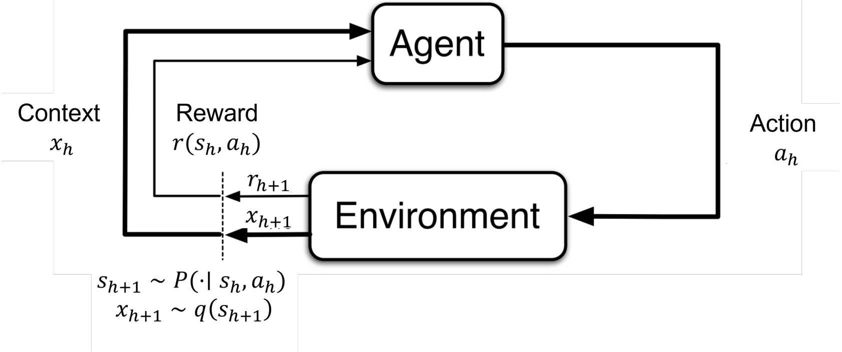
Outline

BMDP

- $M = (\mathcal{S}, \mathcal{X}, \mathcal{A}, H, P, r, q)$
- $\mathcal{S}$ is finite hidden state space that the agent can't observe
- $\mathcal{X}$ is the observable context space, maybe infinite
- P is the transition dynamics $P(\cdot \mid s, a)$
- q is the context emission function: $q : \mathcal{S} \rightarrow \Delta(\mathcal{X})$

$$\forall s \neq s', q(s) \neq q(s')$$

Agent-environment Interactions in BMDP



BMDP with a Decoding Function

Decoding function: f

$$f : \mathcal{X} \to \mathcal{S}$$

We say the decoding function is an ϵ-error decoding if $P_{x \sim q(s)}(f(x) = s) \geq 1 - \epsilon$ holds for all s. The block assumption ensures a 0-error decoding.

- Under some assumptions, the PCID can output a ϵ-error decoding function within $O(poly(H, S, A)/\epsilon)$ time steps
- BMDP with a ϵ-error decoding function can be seen as a MDP with adversarial corruptions and $C = \epsilon HK\iota$. (if $\alpha f(x) = s' \neq s$, it is equivalent to a adversary that substitute s with s')

So combine PCID and CR-MVP, we have regret $O(poly(H, S, A)/\epsilon + \epsilon SAHK + \sqrt{SAK})$, set ϵ properly we have $O(\sqrt{K})$ regret.