On Reinforcement Learning with Adversarial Corruptions

and applications to block MDP

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

June 20, 2021

Overview

Our contrubutions:

- We propose an algorithm that can achieve $\tilde{O}(\sqrt{SAK} + CSA)$ regret when we know the corruption level C
- 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

Episodic MDP

- Finite-horizon MDP: $M = (\mathcal{S}, \mathcal{A}, H, \mathcal{P}, R)$
- Unknown dynamic and reward: $s_{h+1} \sim P(\cdot|s_h, a_h)$
- Policy: $\pi = \{\pi_h|\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^{H}$
- Value: $V_h^\pi(s) = E_\pi[\sum_{i=h}^{H} r_i|s_h = s, a_h = \pi(s)]$
- Q-function: $Q_h^\pi(s, a) = E_\pi[\sum_{i=h}^{H} r_i|s_h = s, a_h = a]$
- Number of episodes: K

MDP with corruptions

Corruption: The adversary replace the state s and r with arbitrary $\tilde{s}$ and $\tilde{r}$.

- strong adversary: after the agent plays an action a_{t-1}, the adversary decide whether to corrupt the value r_{t-1} and the next time step.
- If so, generate arbitrary r'_{t-1}, s'_t and $\tilde{r}(s'_t, \cdot)$.
- Corruption level C: The number of time steps that is corrupted.

Algorithm 1 Corruption Robust Monotonic Value Propagation

Input: C is the corruption level.
for $k = 1, 2, \ldots, K$ **do**
  for $h = 1, 2, \ldots, H$ **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 $h = H, H-1, \ldots, 1$ **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

Main Result

By setting $\tilde{b}_h =$

$$2\min\{\frac{2C}{|n-C|},1\}+c_1\min\{\sqrt{\frac{\mathbb{V}(\tilde{P},V_{h+1})\iota}{|n-C|}}+\frac{\sqrt{C\iota}}{|n-C|},1\}+c_2\min\{\sqrt{\frac{\tilde{r}\iota}{|n-C|}}+\frac{\sqrt{C\iota}}{|n-C|},1\}+c_3\min\{\frac{\iota}{|n-C|},1\},$$

Theorem

(Regret upper bound of CR-MVP) With probability at least $1-\delta$, the regret of `CR-MVP` *satisfies:*

$$\text{Regret}(K) \le \tilde{O}(\sqrt{SAK} + S^2A + CSA),$$

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

Lower Bound

Theorem

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

- 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.

Theorem

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

Application to Episodic Block MDP

- $M = (\mathcal{S}, \mathcal{X}, \mathcal{A}, H, P, r, q)$
- $\mathcal{S}$ is the hidden state space that the agent cannot observe, finite
- $\mathcal{X}$ is the context space that the is observable, possible infinite
- P is the transition over $\mathcal{S}$, $P(s'|s, a), (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$
- q is the context emission function: $q : \mathcal{S} \to \Delta(\mathcal{X})$

Every step h, the agent first observe x_h and execute a_h, recieve a reward $r(s_h, a_h)$,transition to the hidden state $s_{h+1} \sim P(\cdot|s_h, a_h)$. The evironment generate the context $x_{h+1} \sim q(s_{h+1})$, the agent observe x_{h+1} and so on.
And here the q function satisfies the block structure assumption: the support of $q(s)$ and $q(s')$ doesn't overlap for $\forall s \neq s'$

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.

Thank you !