Online Sparse Reinforcement Learning

Botao Hao, Tor Lattimore, Csaba Szepesvári, Mengdi Wang

Contribution

- Online sparse RL in the **high-dimensional regime**.
- **Linear regret** unavoidable even there exists a policy that collects well-conditioned data.
- With an **oracle access** to a policy that collects well-conditioned data, a **sub-linear regret** is possible.

HD Statistics v.s. Sparse Bandits v.s. Sparse RL

- **High-Dimensional Statistics.**
  "Best of both worlds": high representation power with many features while sparsity leads to efficient estimation.
- **Sparse Linear Bandits.**
  Existence of an exploratory policy $\Rightarrow$ **dimension-free** $\Theta(n^{2/3})$ regret bound (Hao et al. NeurIPS 2020).
- **Online Sparse RL.**
  Even though there exists an exploratory policy, finding the exploratory policy is also hard!

Problem Setting

- **Episodic Markov decision process:** $\mathcal{M} = (\mathcal{X}, \mathcal{A}, H, P, r)$ with $\mathcal{X}$ the state-space, $\mathcal{A}$ the action space, H the episode length, $P : \mathcal{X} \times \mathcal{A} \to \Delta_{\mathcal{X}}$ the transition kernel and $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$ the reward function.
- **Value function:**

$$V_h^{\pi}(x) := \mathbb{E}^{\pi}\left[\sum_{h'=h}^{H} r(x_{h'}, a_{h'}) \Big| x_h = x\right] .$$

- **Cumulative regret:**

$$R_N = \sum_{n=1}^{N} \left(V_1^{*}(x_1^n) - V_1^{\pi_n}(x_1^n)\right) .$$

  The **high-dimensional regime** is referred to $N \leq d$.
- **Sparse linear MDP:** Fix a feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and assume the episodic MDP $\mathcal{M}$ is linear in ϕ. We say $\mathcal{M}$ is (s, ϕ)-sparse if there exists an active set $\mathcal{K} \subseteq [d]$ with $|\mathcal{K}| \leq s$ and some functions $\psi(\cdot) = (\psi_k(\cdot))_{k \in \mathcal{K}}$ such that for all pairs of (x, a): $P(x'|x, a) = \sum_{k\in\mathcal{K}} \phi_k(x, a)\psi_k(x')$.

Hardness Of Online Sparse RL

Definition (Exploratory policy). Let Σ^{π} be the expected uncentered covariance matrix induced by policy π and feature map ϕ, given by

$$\Sigma^{\pi} := \mathbb{E}^{\pi}\left[\frac{1}{H}\sum_{h=1}^{H} \phi(x_h, a_h)\phi(x_h, a_h)^{\top}\right] ,$$

where $x_1 \sim \xi_0, a_h \sim \pi(\cdot|x_h), x_{h+1} \sim P(\cdot|x_h, a_h)$. We call **a policy π exploratory** if $\sigma_{\min}(\Sigma^{\pi}) > 0$.

Theorem (Minimax Lower Bound). For any algorithm π, there exists **a sparse linear MDP** $\mathcal{M}$ and associated exploratory policy π_e for which $\sigma_{\min}(\Sigma^{\pi_e})$ is a strictly positive universal constant independent of N and d, such that for any $N \leq d$,

$$R_N \geq \frac{1}{128} H d .$$

Remark.

- Even if the MDP transition kernel can be exactly represented by a sparse linear model and there exists an exploratory policy, the learner could still suffer linear regret in the high-dimensional regime.
- This is in stark contrast to linear bandits, where the existence of an exploratory policy is sufficient for dimension-free regret. The problem in RL is that **finding** the exploratory policy can be very hard.

Hard-to-learn MDP Instance

- The intuition is to construct an **informative state** with only one of a large set of actions leading to the informative state **deterministically**.
- The exploratory policy has to visit that informative state to produce well-conditioned data. In order to find this informative state, the learner should take a large number of trials that will suffer high regret.

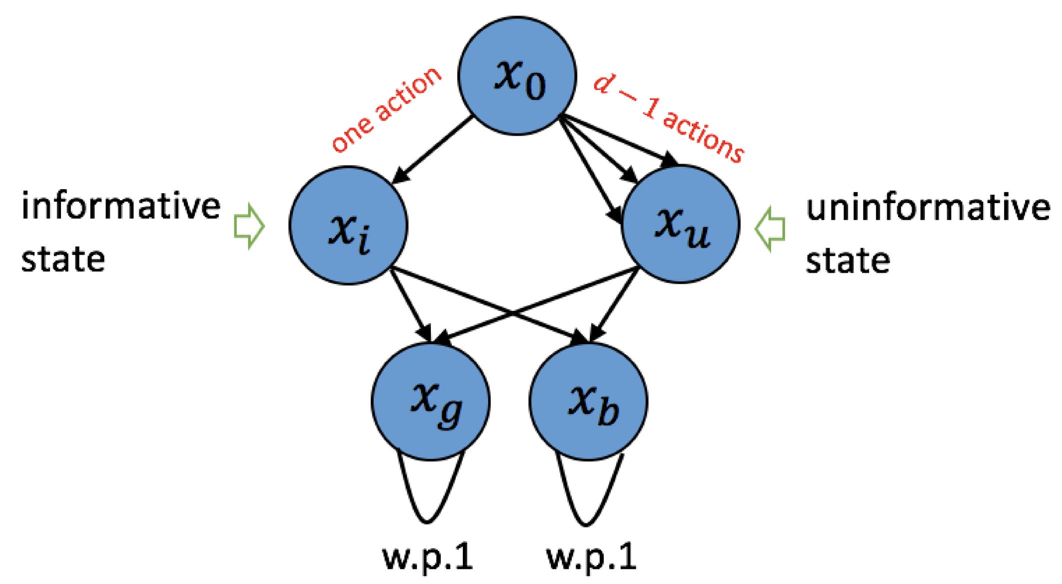

Figure 2: A hard-to-learn MDP instance that includes an informative state and an uninformative state.

Online Lasso-FQI

Assume **an oracle access** to the exploratory policy. The algorithm uses the explore-then-commit template:

- **Exploration phase.** The exploration phase includes N_1 episodes where N_1 will be chosen later based on regret bound and can be factorized as $N_1 = RH$, where $R > 1$ is an integer. At the beginning of each episode, the agent follows **the exploratory policy π_e**.
- **Learning phase.** Based on the exploratory dataset $\mathcal{D}$, the agent executes an extension of FQI combining with Lasso for feature selection. To define the algorithm, it is useful to introduce $Q_w(x, a) = \phi(x, a)^{\top} w$. At each step $h \in [H]$, we fit $\hat{w}_h$ through Lasso:

$$\hat{w}_h = \operatorname*{argmin}_{w} \frac{1}{|\mathcal{D}|} \sum_{(x_i, a_i, x_i')} \left(\max_{a\in\mathcal{A}} Q_{\hat{w}_{h+1}}(x_i', a) - \phi(x_i, a_i)^{\top} w\right)^2 + \lambda_1 \|w\|_1 .$$

- **Exploitation phase.** For the rest $N - N_1$ episodes, the agent commits to the greedy policy with respect to the estimated Q-value $\{Q_{\hat{w}_h}\}_{h=1}^{H}$.

Theorem (regret bound of online lasso-FQI). The cumulative regret of online Lasso-FQI satisfies $R_N \lesssim H^{\frac{4}{3}} s^{\frac{2}{3}} N^{\frac{2}{3}}$.

Remark.

- Without oracle access $\Rightarrow$ linear regret lower bound.
- With oracle access $\Rightarrow$ sublinear regret upper bound.

Conclusion

- **Summary.** We emphasize that in high-dimensional regime, exploiting the sparsity to reduce the regret needs an exploratory policy but finding the exploratory policy is as hard as solving the MDP itself - an irresolvable "chicken and egg" problem.