Pure Exploration for Reinforcement Learning

Tian Xu

xut@lamda.nju.edu.cn

Nanjing University

Mainly based on:
Fast active learning for pure exploration in reinforcement learning

December 8, 2021

Outline

Background

Best Policy Identification

Reward-free Exploration

Performance measure

- In Reinforcement Learning (RL), generally, we may be interested in
  - the performance of the agent during the learning phases.
  - the performance of the final learned policy.

Performance measure

- In the first setting, there are mainly two performance measure: Regret and PAC-MDP.
- High probability regret [Azar et al., 2017a]: There exists a function $F(S, A, H, T, \log(1/\delta))$ such that

$$\Pr\left(\sum_{t=1}^{T}(V^* - V^{\pi_t}) > F(S, A, H, T, \log(1/\delta))\right) \le \delta.$$

- PAC-MDP [Dann and Brunskill, 2015]: There exists a polynomial function $\mathrm{Poly}(S, A, H, 1/\epsilon, \log(1/\delta))$ such that

$$\Pr(N_\epsilon > \mathrm{Poly}(S, A, H, 1/\epsilon, \log(1/\delta))) \le \delta,$$

where $N_\epsilon = \sum_{t=1}^{\infty} \mathbb{I}\left(V^* - V^{\pi_t} \ge \epsilon\right)$.

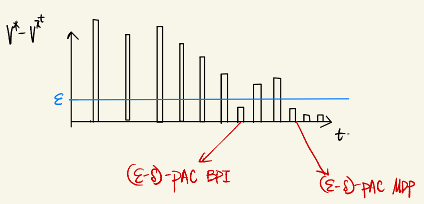
Performance measure

- In this talk, we focus on the second setting (free-exploration).
- Best policy identification (BPI): An algorithm is (ϵ, δ)-PAC for BPI if there exists a $\mathrm{Poly}(S, A, H, 1/\epsilon, \log(1/\delta))$, after $T \geq \mathrm{Poly}(S, A, H, 1/\epsilon, \log(1/\delta))$ episodes, it returns a policy $\hat{\pi}$ satisfies that
$$\Pr\left(V^* - V^{\hat{\pi}} > \epsilon\right) \leq \delta.$$
- Reward-free exploration (RFE): An algorithm is (ϵ, δ)-PAC for RFE if there exists a $\mathrm{Poly}(S, A, H, 1/\epsilon, \log(1/\delta))$, after $T \geq \mathrm{Poly}(S, A, H, 1/\epsilon, \log(1/\delta))$ episodes, it returns a policy $\hat{\pi}$ satisfies that
$$\Pr\left(\text{for any reward function } r, V^*(r) - V^{\hat{\pi}}(r) > \epsilon\right) \leq \delta.$$

Permance measure

- (ϵ, δ)-PAC-MDP v.s. (ϵ, δ)-PAC for BPI
- (ϵ, δ)-PAC-MDP upper bounds the number of time steps in which an algorithm makes ϵ mistakes.
- (ϵ, δ)-PAC for BPI upper bounds the number of time steps before the algorithm outputs an ϵ sub-optimal policy.

Performance measure



- (ϵ, δ)-PAC-MDP is stronger than (ϵ, δ)-PAC for BPI.
- An algorithm which is (ϵ, δ)-PAC for BPI needs a stopping rule to determine when to output the policy.

Background

- The reward-free reinforcement learning can be split into two phases:
  - An exploration phase: The agent interacts with the environment without reward signal and learns an empirical transition model $\hat{p}$.
  - A planning phase: The agent receives a reward function and learns a policy in the constructed model $\hat{p}$.
- Why reward-free reinforcement learning?
  - In some applications, we hope to learn good policies for a wide range of reward functions.
  - We want to explore more efficiently in some environments where the reward signal is sparse (unknown).

Markov Decision Process

- Consider a finite episodic Markov Decision Process $\left(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in[H]}\right)$.
  - $\mathcal{S}$ and $\mathcal{A}$ are the state and action space, respectively.
  - $r_h(s,a) \in [0,1]$ is deterministic reward received after taking the action a in state s at step h.
  - $p_h(s'|s,a)$ specifies the transition probability of s' conditioned on s and a at step h.
  - H is the horizon length.
  - The initial state s_1 is fixed.

Markov Decision Process

- A deterministic policy is a collection of functions $\pi_h : \mathcal{S} \to \mathcal{A}$ for all $h \in [H]$.
- The value function and Q-value function of π:

$$V_h^{\pi}(s) \triangleq \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \mid s_h = s\right]$$

$$Q_h^{\pi}(s, a) \triangleq \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}\left(s_{h'}, a_{h'}\right) \mid s_h = s, a_h = a\right]$$

Markov Decision Process

- The expectation operator regarding p: $pf(s,a) \triangleq \mathbb{E}_{s' \sim p(\cdot|s,a)} [f(s')]$
- The composition with the policy π: $(\pi g)(s) \triangleq \pi g(s) \triangleq g(s, \pi(s))$.
- The variance operator regarding p: $\mathrm{Var}_p(f)(s,a) = \mathbb{E}_{s' \sim p(\cdot|s,a)}[(f(s') - pf(s,a))^2]$
- The Bellman and Bellman optimality equations:

$$V_h^\pi(s) = \pi_h Q_h^\pi(s), \text{ with } Q_h^\pi(s,a) \triangleq r_h(s,a) + p_h V_{h+1}^\pi(s,a)$$

$$V_h^\star(s) = \max_a Q_h^\star(s,a), \text{ with } \quad Q_h^\star(s,a) \triangleq r_h(s,a) + p_h V_{h+1}^\star(s,a)$$

Preliminaries

- Let $\left(s_h^i, a_h^i, s_{h+1}^i\right)$ be the state, the action, and the next state observed at step h of episode i.
- Let $n_h^t(s,a) \triangleq \sum_{i=1}^{t} \mathbb{I}\left\{\left(s_h^i, a_h^i\right) = (s,a)\right\}$ be the number of times the state-action pair (s,a) was visited in step h in the first t episodes.
- Let $n_h^t(s,a) = \mathbb{E}[n_h^t(s,a)] = \sum_{t'=1}^{t} p_h^{t'}(s,a)$ be the pseudo-counts, where $p_h^{t'}(s,a)$ is the probability of visiting (s,a) at h when executing $\pi^{t'}$.

Preliminaries

- The empirical transitions:

$$\widehat{p}_h^t\left(s' \mid s, a\right) \triangleq \frac{n_h^t\left(s, a, s'\right)}{n_h^t(s, a)} \text{ if } n_h^t(s, a) > 0 \text{ and } \widehat{p}_h^t\left(s' \mid s, a\right) \triangleq \frac{1}{S} \text{ otherwise .}$$

- Let $\widehat{V}_h^{t,\pi}(s)$ and $\widehat{Q}_h^{t,\pi}(s, a)$ be the value and Q-value function with respect to the transition model $\hat{p}^t$.

Concentration Events

$$\mathcal{E} \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \mathrm{KL}\left(\widehat{p}_h^t(\cdot \mid s,a), p_h(\cdot \mid s,a)\right) \leq \frac{\beta\left(n_h^t(s,a), \delta\right)}{n_h^t(s,a)}\right\}$$

$$\mathcal{E}^{\mathrm{cnt}} \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : n_h^t(s,a) \geq \frac{1}{2}\bar{n}_h^t(s,a) - \beta^{\mathrm{cnt}}(\delta)\right\},$$ and

$$\mathcal{E}^{\star} \triangleq \left\{\forall t \in \mathbb{N}, \forall h \in [H], \forall (s,a) \in \mathcal{S} \times \mathcal{A} : \left|\left(\widehat{p}_h^t - p_h\right) V_{h+1}^{\star}(s,a)\right| \leq \min\left(H, \sqrt{2\,\mathrm{Var}_{p_h}\left(V_{h+1}^{\star}\right)(s,a)\frac{\beta^{\star}\left(n_h^t(s,a), \delta\right)}{n_h^t(s,a)}} + 3H\frac{\beta^{\star}\left(n_h^t(s,a), \delta\right)}{n_h^t(s,a)}\right)\right\}$$

- For all $\delta \in (0,1)$, $\Pr\left(\mathcal{E} \cap \mathcal{E}^{\mathrm{cnt}} \cap \mathcal{E}^{\star}\right) \geq 1 - \delta$.

Outline

Background

Best Policy Identification

Reward-free Exploration

From Regret to BPI

- The main difficulty for converting a regret-minimization method to BPI lies in high-probability prediction of an ϵ-optimal policy.
- For UCB-VI [Azar et al., 2017b], with probability at least $1-\delta'$, $\sum_{t=1}^{T} V^{\star}(s_1) - V_1^{\pi^t}(s_1) \leq \sqrt{H^3 SA \log(1/\delta') T}$.

From Regret to BPI

- If we choose $\hat{\pi}$ uniformly sampled from $\left(\pi^t\right)_{t\in[T]}$, by Markov's inequality, we have

$$\Pr\left(V_1^\star\left(s_1\right) - V_1^{\hat{\pi}}\left(s_1\right) \geq \varepsilon\right) \leq \frac{1}{\varepsilon}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} V^\star\left(s_1\right) - V_1^{\pi^t}\left(s_1\right)\right] \leq \frac{1}{\varepsilon}\left(C\sqrt{\frac{H^3SA}{T}\log\left(1/\delta'\right)} + \delta' H\right)$$

- Let the first term in RHS be $\frac{\delta}{2}$, we have

$$T \triangleq \frac{2H^3SA}{\varepsilon^2\delta^2}\log\left(\frac{2H}{\varepsilon\delta}\right)$$

- The sample complexity scales with $1/\delta^2$ whereas we expect $\log(1/\delta)$.

From Regret to BPI

- The additional dependence on $\frac{1}{\delta^2}$ comes from the randomness of $\hat{\pi}$. If we can deterministically output a policy $\hat{\pi}$ with $V_1^{\hat{\pi}}(s_1) = \frac{1}{T}\sum_{t=1}^{T} V_1^{\pi^t}(s_1)$, this issue is solved.
- Note that $V_1^{\pi} = \sum_{h=1}^{H} \mathbb{E}_{(s,a)\sim p_h^{\pi}}[r(s,a)]$, we construct $\hat{\pi}$ such that $p_h^{\hat{\pi}}(s,a) = \bar{p}_h(s,a) = \frac{1}{T}\sum_{t=1}^{T} p_h^t(s,a)$.

$$\bar{\pi}_h(a \mid s) \triangleq \begin{cases} \frac{\bar{p}_h(s,a)}{\sum_{b\in\mathcal{A}} \bar{p}_h(s,b)} & \text{if } \sum_{b\in\mathcal{A}} \bar{p}_h(s,b) > 0, \text{ and} \\ 1/A & \text{otherwise.} \end{cases}$$

From Regret to BPI

- For $\hat{\pi}$, with probability at least $1 - \delta'$, we have

$$V_1^{\star}(s_1) - V_1^{\hat{\pi}}(s_1) = \frac{1}{T}\sum_{t=1}^{T} V^{\star}(s_1) - V_1^{\pi^t}(s_1) \leq \sqrt{\frac{H^3 SA}{T}\log(1/\delta')}$$

- Choosing $\delta' \triangleq \delta$ and $T \triangleq H^3 SA \log(1/\delta)/\varepsilon^2$ would lead to an (ϵ, δ)-PAC algorithm for BPI with a minimax optimal sample complexity.
- However, we can not compute $p_h^t(s, a)$ without the knowledge of transition probability.

BPI-UCBVI

- BPI-UCBVI is a model-based UCB method.
- In the iteration $t+1$, it estimates an empirical transition model $\hat{p}^t$ and computes $\tilde{Q}_h^t(s,a)$ based on $\hat{p}^t$. In each t, $\tilde{Q}_h^t(s,a)$ is a UCB of $Q_h^*(s,a)$ for all (s,a,h).
- The sampling policy π^{t+1} is the greedy policy with respect to $\tilde{Q}_h^t(s,a)$.
- For the stopping rule, BPI-UCBVI establishes an upper bound of $V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1)$

Upper Confidence Bound

$$\tilde{Q}_h^t(s,a) \triangleq \min\left(H, r_h(s,a) + \widehat{p}_h^t \tilde{V}_{h+1}^t(s,a) + b_h^t(s,a)\right)$$

$$b_h^t(s,a) = 3\sqrt{\operatorname{Var}_{\widehat{p}_h^t}\left(\tilde{V}_{h+1}^t\right)(s,a)\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + 14H^2\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \frac{1}{H}\hat{p}_h^t\left(\tilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t\right)(s,a)$$

$$\tilde{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}} \tilde{Q}_h^t(s,a)$$

$$\tilde{V}_{H+1}^t(s) \triangleq 0$$

- where $\underset{\sim}{V}_h^t$ is the lower confidence bound (LCB) of V_h^*.

Lower Confidence Bound

$$\underset{\sim}{Q}_h^t(s,a) \triangleq \min\left(H, r_h(s,a) + \widehat{p}_h^t \underset{\sim}{V}_{h+1}^t(s,a) - b_h^t(s,a)\right)$$

$$b_h^t(s,a) = 3\sqrt{\operatorname{Var}_{\widehat{p}_h^t}\left(\tilde{V}_{h+1}^t\right)(s,a)\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + 14H^2\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \frac{1}{H}\hat{p}_h^t\left(\tilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t\right)(s,a)$$

$$\underset{\sim}{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}} \underset{\sim}{Q}_h^t(s,a)$$

$$\underset{\sim}{V}_{H+1}^t(s) \triangleq 0$$

Upper Confidence Bound

Lemma

We have that for all t, all $h \in [H]$, and all (s, a),

$$\underset{\sim}{Q}_h^t(s,a) \leq Q_h^\star(s,a) \leq \tilde{Q}_h^t(s,a) \quad \textit{and}$$
$$\underset{\sim}{V}_h^t(s) \leq V_h^\star(s) \leq \tilde{V}_h^t(s)$$

Proof Sketch

- The proof is based on backward induction.
- For $h = H + 1$, this result is true. Assume the inequalities hold for $h' > h$.
- We will show that $\tilde{Q}_h(s,a) - Q_h^\star(s,a) \geq 0$.

$$\tilde{Q}_h(s,a) - Q_h^\star(s,a) \geq \widehat{p}_h^t \left(\tilde{V}_{h+1}^t - V_{h+1}^\star \right)(s,a) + \left(\widehat{p}_h^t - p_h \right) V_{h+1}^\star(s,a) + b_h^t(s,a)$$

Proof Sketch

$$\left|\left(\widehat{p}_h^t - p_h\right) V_{h+1}^\star(s,a)\right| \le \sqrt{2\,\mathrm{Var}_{p_h}\left(V_{h+1}^\star\right)(s,a)\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + 3H\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}$$

$$\mathrm{Var}_{p_h}\left(V_{h+1}^\star\right)(s,a) \le 4\,\mathrm{Var}_{\widehat{p}_h^t}\left(\tilde{V}_{h+1}^t\right)(s,a) + 4H\widehat{p}_h^t\left(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t\right)(s,a) + 4H^2\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}$$

$$\left|\left(\widehat{p}_h^t - p_h\right) V_{h+1}^\star(s,a)\right| \le 3\sqrt{\mathrm{Var}_{\widehat{p}_h^t}\left(\widetilde{V}_{h+1}^t\right)(s,a)\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + 14H^2\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}$$
$$+\frac{1}{H}\widehat{p}_h^t\left(\widetilde{V}_{h+1}^t - \underset{\sim}{V}_{h+1}^t\right)(s,a) = b_h^t(s,a)$$

Stopping Rule

- We need to build an UCB on the policy value gap $V_1^{\star}(s_1) - V_1^{\pi^{t+1}}(s_1)$.

$$G_h^t(s,a) \triangleq \min\left(H, 6\sqrt{\operatorname{Var}_{\widehat{p}_k^t}\left(\widetilde{V}_{h+1}^t\right)(s,a)\frac{\beta^{\star}\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + 36H^2\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}\right.$$

$$\left. + \left(1+\frac{3}{H}\right)\widehat{p}_h^t G_{h+1}^t(s)\right)$$

$$G_{h+1}^t(s) = G_{h+1}^t\left(s, \pi_{h+1}^{t+1}(s)\right)$$

$$G_{H+1}^t(s,a) \triangleq 0$$

Stopping Rule

Lemma

For all t, $V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1) \le \pi_1^{t+1} G_1^t(s_1)$

BPI-UCBVI

Theorem

For $\delta \in (0,1)$, $\epsilon \in (0, 1/S^2]$, BPI-UCBVI is (ϵ, δ)-PAC for BPI. Moreover, w.p. $1-\delta$,

$$\tau \leq \frac{H^3SA}{\varepsilon^2}(\log(3SAH/\delta)+1)C_1 + 1,$$

where $C_1 \triangleq 5904e^{26}\log\left(e^{30}(\log(3SAH/\delta)+S)H^3SA/\varepsilon\right)^2$.

- The rate of BPI-UCBVI is of order $\widetilde{\mathcal{O}}\left(H^3SA\log(1/\delta)/\varepsilon^2\right)$ when ϵ is small enough and matches the lower bounds of $\Omega\left(H^3SA\log(1/\delta)/\varepsilon^2\right)$ by [Domingues et al., 2020] up tp poly-log terms.

Proof Sketch

- If BPI-UCBVI stops at time τ, then we have $V_1^{\widehat{\pi}}(s_1) = V_1^{\pi^{\tau+1}}(s_1) \geq V_1^{\star}(s_1) - \pi_1^{\tau+1} G_1^{\tau}(s_1) \geq V_1^{\star}(s_1) - \varepsilon$.
- For all $t < \tau$, by the stopping rule, we have $\varepsilon \leq \pi_1^{t+1} G_1^{t}(s_1)$. Then we have $\tau\epsilon \leq \sum_{t=0}^{\tau-1} \pi_1^{t+1} G_1^{t}(s_1)$
- For all $t < \tau$, we upper bound $\pi_1^{t+1} G_1^{t}(s_1)$ and build a formula regarding τ.
- Solving the established formula results in the upper bound of τ.

Proof Sketch

- Upper bound on $G_h^t(s,a)$:

$$G_h^t(s,a) \le 6\sqrt{\mathrm{Var}_{\widehat{p}_h^t}\left(\widetilde{V}_{h+1}^t\right)(s,a)\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + 36H^2\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \left(1+\frac{3}{H}\right)\widehat{p}_h^t\pi_{h+1}^{t+1}G_{h+1}^t(s,a)$$

- Replace $\widehat{p}_h^t, \widetilde{V}_{h+1}^t$ with $p_h, V_{h+1}^{\pi^{t+1}}$:

$$G_h^t(s,a) \le 12\sqrt{\mathrm{Var}_{p_h}\left(V_{h+1}^{\pi^{t+1}}\right)(s,a)\left(\frac{\beta^\star\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} \wedge 1\right)} + 84H^2\left(\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} \wedge 1\right) + \left(1+\frac{13}{H}\right)p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a)$$

Proof Sketch

- Unfold the above equation and replace the counts by the pseudo-counts.

$$\pi_1^{t+1} G_1^t (s_1) \leq 12e^{13} \sum_{h=1}^{H} \sum_{s,a} p_h^{t+1}(s,a) \sqrt{\operatorname{Var}_{p_h}\left(V_{h+1}^{\pi+1}\right)(s,a) \left(\frac{\beta\left(\bar{n}_h^t(s,a),\delta\right)}{\bar{n}_h^t(s,a) \vee 1} \right)}$$
$$+336e^{13} H^2 \sum_{h=1}^{H} \sum_{s,a} p_h^{t+1}(s,a) \left(\frac{\beta\left(\bar{n}_h^t(s,a),\delta\right)}{\bar{n}_h^t(s,a) \vee 1} \right)$$

- The law of total variance:

$$H^2 \geq \mathbb{E}_\pi \left[\left(\sum_{h=1}^{H} r_h(s_h,a_h) - V_1^\pi(s_1) \right)^2 \right] = \sum_{h=1}^{H} \sum_{s,a} p_h^\pi(s,a) \operatorname{Var}_{p_h}\left(V_{h+1}^\pi\right)(s,a)$$

Proof Sketch

- We build the upper bound of $\pi_1^{t+1} G_1^t\left(s_1\right)$

$$
\begin{aligned}
\pi_1^{t+1} G_1^t\left(s_1\right) \leq & 24 e^{13} H \sqrt{\sum_{h=1}^H \sum_{s, a} p_h^{t+1}(s, a) \frac{\beta^{\star}\left(\bar{n}_h^t(s, a), \delta\right)}{\bar{n}_h^t(s, a) \vee 1}} \\
& +336 e^{13} H^2 \sum_{h=1}^H \sum_{s, a} p_h^{t+1}(s, a) \frac{\beta\left(\bar{n}_h^t(s, a), \delta\right)}{\bar{n}_h^t(s, a) \vee 1}
\end{aligned}
$$

- Builds the formula regarding τ via $\tau \epsilon \leq \sum_{t=0}^{\tau-1} \pi_1^{t+1} G_1^t\left(s_1\right)$:

$$
\varepsilon \tau \leq 48 e^{13} \sqrt{\tau H^3 S A \beta^{\star}(\tau-1, \delta) \log (\tau+1)}+1344 e^{13} H^3 S A \beta(\tau-1, \delta) \log (\tau+1)
$$

Proof Sketch

- Solving the formula results in the sample complexity.

$$\tau \leq \frac{H^3 SA}{\varepsilon^2}(\log(3SAH/\delta) + S)C_1 + 1.$$

Outline

Background

Best Policy Identification

Reward-free Exploration

Reward-free Exploration (RFE)

- One approach to RFE relies on known cumulative-regret minimization methods.
  - RF-RL-Explore [Jin et al., 2020] runs EULER algorithm for each (s, h) with a reward function encouraging the visit of state s at step h.
- Another methods [Kaufmann et al., 2020, Ménard et al., 2020] build the upper bound of the estimation error $|Q_h^\pi(s,a;r) - \hat{Q}_h^\pi(s,a;r)|$ of any policy and any reward function, and the agent acts greedily with respect to the upper bounds to minimize the estimation error.

$$V_1^\star(s_1; r) - V_1^{\widehat{\pi}^\star, r}(s_1; r) \leq 2 \max_a |Q_1^\pi(s,a;r) - \hat{Q}_1^\pi(s,a;r)|$$

Statistical Goal

Definition

An algorithm is (ϵ, δ)-PAC for reward-free exploration if

$$\mathbb{P}\left(\text{ for any reward function } r, V_1^{\star}\left(s_1; r\right) - V_1^{\widehat{\pi}_r^{\star}}\left(s_1; r\right) \leq \varepsilon \right) \geq 1 - \delta,$$

where $\widehat{\pi}_r^{\star}$ is the optimal policy in the empirical MDP with $\hat{p}$ and r.

- The number of episodes required to achieve (ϵ, δ)-PAC.

Summary of RFE Algorithms

Algorithms	Upper bound (non-stationary setting)	Lower bound (stationar
RF-RL-Explore[Jin et al., 2020]	$\frac{H^7 S^2 A}{\varepsilon} \log^3\left(\frac{1}{\delta}\right) + \frac{H^5 S^2 A}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)$	$\frac{H^2 SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)$
RF-UCRL [Kaufmann et al., 2020]	$\frac{H^4 SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)$	
RF-Express [Ménard et al., 2020]	$\frac{H^3 SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)$	

- RF-Express are sub-optimal only by a factor of H.
- Note that lower bound is proved in the stationary setting and RF-Express may be minimax-optimal in the non-stationary setting.

Reward-free Exploration

Algorithm: Reward-free Exploration algorithm

1: **for** $t = 1, 2, \cdots$ **do**

2: Interact with environment without reward via sampling policy π^t and obtain a reward-free episode $z^t \triangleq \left(s_1^t, a_1^t, s_2^t, a_2^t, \ldots, s_H^t, a_H^t\right)$.

3: Update the dataset $\mathcal{D}_t \triangleq \mathcal{D}_{t-1} \cup \{z_t\}$

4: Stop or continue according to a stopping time τ

5: **end for**

6: **Output:** The empirical transition model $\hat{p}$ built on $\mathcal{D}_\tau$.

- Two key parts: sampling policy and stopping rule.

The Upper Bound

- We build the upper bound of the estimation error and the sampling policy is greedy with respect to the upper bound.
- After episode t, we define the estimation error $\widehat{e}_h^{t,\pi}(s,a;r) \triangleq \left|\widehat{Q}_h^{t,\pi}(s,a;r) - Q_h^{\pi}(s,a;r)\right|$.
- The functions $W_h^t(s,a)$ are defined inductively:

$$W_{H+1}^t(s,a) \triangleq 0 \ \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

$$W_h^t(s,a) \triangleq \min\left(H, \underbrace{15H^2 \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}}_{\text{bonus term}} + \left(1 + \frac{1}{H}\right) \sum_{s'} \widehat{p}_h^t\left(s' \mid s,a\right) \max_{a'} W_{h+1}^t\left(s',a'\right) \right)$$

where $\beta(n,\delta) \triangleq \log(3SAH/\delta) + S\log(8e(n+1))$.

The Upper Bound

Lemma

With probability at least $1-\delta$, for any episode t, policy π, and reward function r,

$$\widehat{e}_1^{t,\pi}\left(s_1,\pi_1\left(s_1\right);r\right) \leq 3e\sqrt{\max_{a\in\mathcal{A}} W_1^t\left(s_1,a\right)} + \max_{a\in\mathcal{A}} W_1^t\left(s_1,a\right)$$

- The sampling rule: the policy π^{t+1} is the greedy policy with respect to W_h^t: $\forall s\in\mathcal{S}, \forall h\in[H], \quad \pi_h^{t+1}(s) = \arg\max_{a\in\mathcal{A}} W_h^t(s,a)$.
- The stopping rule: $\tau = \inf\left\{t\in\mathbb{N} : 3e\sqrt{\pi_1^{t+1}W_1^t\left(s_1\right)} + \pi_1^{t+1}W_1^t\left(s_1\right) \leq \varepsilon/2\right\}$.

The Upper Bound

- $W_h^t(s,a) \triangleq \min\left(H, \underbrace{15H^2 \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}}_{\text{bonus term}} + \left(1+\frac{1}{H}\right) \sum_{s'} \widehat{p}_h^t\left(s' \mid s,a\right) \max_{a'} W_{h+1}^t\left(s',a'\right) \right)$
- The bonus term scale with $\frac{1}{N}$ rather than $\frac{1}{\sqrt{N}}$, suggesting that RL-Express is more exploratory.

The Upper Bound

- Fixing a policy π, let P^{π} be the probability distribution of a trajectory in the true MDP and $\hat{P}^{t,\pi}$ be the counterpart in the empirical MDP $\hat{p}^t$. $\mathrm{KL}\left(\widehat{P}^{t,\pi}, P^{\pi}\right) = \sum_{h=1}^{H} \sum_{s,a} \widehat{p}_h^{t,\pi}(s,a) \mathrm{KL}\left(\widehat{p}_h^{t,\pi}(\cdot|s,a), p_h(\cdot|s,a)\right) \leq \sum_{h=1}^{H} \sum_{s,a} \widehat{p}_h^{t,\pi}(s,a) \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$.
- $\pi_1^{t+1} W_1^t(s_1) = 15H^2 \sum_{h=1}^{H} (1+\frac{1}{H})^h \sum_{s,a} \hat{p}_h^{t,\pi}(s,a) \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$
- $\max_{\pi} \mathrm{KL}\left(\widehat{P}^{t,\pi}, P^{\pi}\right) \lesssim \frac{\pi_1^{t+1} W_1^t(s_1)}{H^2}$
- Therefore, RF-Express can be interpreted as an algorithm minimizing an upper-confidence bound on $\max_{\pi} \mathrm{KL}\left(\widehat{P}^{t,\pi}, P^{\pi}\right)$.

Proof Sketch

- Error decomposition:
$$\widehat{e}_h^{t,\pi}(s,a;r) \leq \left|\widehat{Q}_h^{t,\pi}(s,a;r) - Q_h^{\pi}(s,a;r)\right| \leq \left|\left(\widehat{p}_h^t - p_h\right) V_{h+1}^{\pi}(s,a;r)\right| + \widehat{p}_h^t \left|\widehat{V}_{h+1}^{t,\pi} - V_{h+1}^{\pi}\right|(s,a;r)$$
$$= \left|\left(\widehat{p}_h^t - p_h\right) V_{h+1}^{\pi}(s,a;r)\right| + \hat{p}_h^t \pi_{h+1}^t \widehat{e}_{h+1}^{t,\pi}(s,a;r)$$
- By Bernstein inequality,
$$\left|\left(\widehat{p}_h^t - p_h\right) V_{h+1}^{\pi}(s,a)\right| \leq \sqrt{2\mathrm{Var}_{p_h}\left(V_{h+1}^{\pi}\right)(s,a;r)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2}{3}H\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$
- $\mathrm{Var}_{p_h}\left(V_{h+1}^{\pi}\right)(s,a;r) \leq 4\mathrm{Var}_{\widehat{p}_h}\left(\widehat{V}_{h+1}^{t,\pi}\right)(s,a;r) + 4H\hat{p}_h^t\left|V_{h+1}^{\pi} - \widehat{V}_{h+1}^{t,\pi}\right|(s,a) + 4H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$
- $\widehat{e}_h^{t,\pi}(s,a;r) \leq$
$$3\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h^t}(\widehat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right)} + 15H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right)\hat{p}_h^t\pi_{h+1}\widehat{e}_{h+1}^{t,\pi}(s,a;r)$$

Proof Sketch

- $\pi_1 \widehat{e}_1^{t,\pi}(s_1;r) \leq \pi_1 Y_1^{t,\pi}(s_1;r) + \pi_1 W_1^{t,\pi}(s_1)$

  $Y_h^{t,\pi}(s,a;r) \triangleq 3\sqrt{\frac{\mathrm{Var}_{\widehat{p}_h{}^t}\left(\widehat{V}_{h+1}^{t,\pi}\right)(s,a;r)}{H^2}\left(\frac{H^2\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} \wedge 1\right)} + \left(1+\frac{1}{H}\right)\hat{p}_h^t \pi_{h+1} Y_{h+1}^{t,\pi}(s,a;r)$

  $W_h^{t,\pi}(s,a) \triangleq 15H^2 \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right)\hat{p}_h^t \pi_{h+1} W_{h+1}^{t,\pi}(s,a)$

- Law of total variance:

  $\sum_{h=1}^H \sum_{s,a} p_h^\pi(s,a)\,\mathrm{Var}_{p_h}\left(V_{h+1}^\pi\right)(s,a) = \mathbb{E}_\pi\left[\left(\sum_{h=1}^H r_h(s_h,a_h) - V_1^\pi(s_1)\right)^2\right] \leq H^2.$

- $\pi Y_1^{t,\pi}(s_1;r) \leq 3e\sqrt{\sum_{s,a}\sum_{h=1}^H \hat{p}_h^{t,\pi}(s,a)\left(\frac{H^2\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} \wedge 1\right)} \leq 3e\sqrt{\pi_1 W_1^{t,\pi}(s_1)}.$

- $\pi_1 \hat{e}_1^{t,\pi}(s_1;r) \leq 3e\sqrt{\pi_1 W_1^{t,\pi}(s_1)} + \pi_1 W_1^{t,\pi}(s_1).$

The Sample Complexity

Theorem

For $\delta \in (0,1), \varepsilon \in (0,1]$, RF-Express is (ϵ, δ)-PAC for reward-free exploration, Moreover, RF-Express stops after τ episodes where, with probability at least $1-\delta$,

$$\tau \leq \frac{H^3SA}{\varepsilon^2}(\log(3SAH/\delta) + S)C_1 + 1$$

and where $C_1 \triangleq 5587e^6 \log\left(e^{18}(\log(3SAH/\delta) + S)H^3SA/\varepsilon\right)^2$.

- The sample complexity of RF-Express matches the lower bound of $\Omega\left(H^2S^2A/\varepsilon^2\right)$ [Jin et al., 2020] up to a factor of H.
- Up to a factor H, the result also matches the lower bound of $\Omega\left(H^2SA\log(1/\delta)/\varepsilon^2\right)$ [Dann and Brunskill, 2015] which is informative in the regime where ϵ is fixed and δ tends to 0.

Proof Sketch

- Upper bound on W_1^t:

$$W_h^t(s,a) \leq 15H^2 \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right) \sum_{s'} \widehat{p}_h^t\left(s' \mid s,a\right) \max_{a'} W_{h+1}^t\left(s',a'\right)$$

$$= 15H^2 \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} + \left(1+\frac{1}{H}\right)\left(\hat{p}_h^t - p_h\right) \pi_{h+1}^{t+1} W_{h+1}^t(s,a) + \left(1+\frac{1}{H}\right) p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a)$$

- By Bernstein inequality,

$$\left(\widehat{p}_h^t - p_h\right) \pi_{h+1}^{t+1} W_{h+1}^t(s,a) \leq \sqrt{2 \operatorname{Var}_{p_h}\left(\pi_{h+1}^{t+1} W_{h+1}^t\right)(s,a) \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}} + \frac{2}{3} H \frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)}$$

- With $\operatorname{Var}_{p_h}\left(\pi_{h+1}^{t+1} W_{h+1}^t\right)(s,a) \leq H p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a)$, we have

$$W_h^t(s,a) \leq 21H^2 \left(\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} \wedge 1\right) + \left(1+\frac{3}{H}\right) p_h \pi_{h+1}^{t+1} W_{h+1}^t(s,a)$$

- Unfolding the above inequality obtains that

$$\pi_1^{t+1} W_1^t\left(s_1\right) \leq 21 e^3 H^2 \sum_{h=1}^H \sum_{s,a} p_h^{t+1}(s,a) \left(\frac{\beta\left(n_h^t(s,a),\delta\right)}{n_h^t(s,a)} \wedge 1\right)$$

Proof Sketch

- Define the pseudo-counts: $\bar{n}_h^t(s,a) \triangleq \sum_{\ell=1}^t p_h^\ell(s,a)$ and we can replace the counts with pseudo-counts: $\pi_1^{t+1} W_1^t(s_1) \leq 84e^3 H^2 \sum_{h=1}^H \sum_{s,a} p_h^{t+1}(s,a) \frac{\beta(\bar{n}_h^t(s,a),\delta)}{\bar{n}_h^t(s,a) \vee 1}$
- For $t \leq T < \tau$, $\varepsilon \leq 3e\sqrt{\pi_1^{t+1} W_1^t(s_1)} + \pi_1^{t+1} W_1^t(s_1)$ due to the stopping rule.
- Take summation over $0 \leq t \leq T$ and apply Cauchy-Schwarz inequality, we have $(T+1)\epsilon \leq 3e\sqrt{(T+1)\sum_{t=0}^T \pi_1^{t+1} W_1^t(s_1)} + \sum_{t=0}^T \pi_1^{t+1} W_1^t(s_1)$
- $\sum_{t=0}^T \pi_1^{t+1} W_1^t(s_1) \leq 336e^3 H^3 SA \log(T+2)\beta(T,\delta)$
- Thus we obtain the inequality on τ: $\varepsilon\tau \leq 55e^3\sqrt{\tau H^3 SA \log(\tau+1)\beta(\tau-1,\delta)} + 336e^3 H^3 SA \log(\tau+1)\beta(\tau-1,\delta)$.
- Solving the above inequality obtains the final result: $\tau \leq \frac{H^3 SA}{\varepsilon^2}(\log(3SAH/\delta) + S)C_1 + 1$.

Bibliography

[Azar et al., 2017a] Azar, M. G., Osband, I., and Munos, R. (2017a).

Minimax regret bounds for reinforcement learning.

In Precup, D. and Teh, Y. W., editors, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 263–272. PMLR.

[Azar et al., 2017b] Azar, M. G., Osband, I., and Munos, R. (2017b).

Minimax regret bounds for reinforcement learning.

In Precup, D. and Teh, Y. W., editors, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017, volume 70 of Proceedings of Machine Learning Research, pages 263–272. PMLR.

Bibliography (cont.)

[Dann and Brunskill, 2015] Dann, C. and Brunskill, E. (2015).

Sample complexity of episodic fixed-horizon reinforcement learning.

In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 2818–2826.

[Domingues et al., 2020] Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2020).

Episodic reinforcement learning in finite mdps: Minimax lower bounds revisited.

CoRR, abs/2010.03531.

Bibliography (cont.)

[Jin et al., 2020] Jin, C., Krishnamurthy, A., Simchowitz, M., and Yu, T. (2020).

Reward-free exploration for reinforcement learning.

CoRR, abs/2002.02794.

[Kaufmann et al., 2020] Kaufmann, E., Ménard, P., Domingues, O. D., Jonsson, A., Leurent, E., and Valko, M. (2020).

Adaptive reward-free exploration.

CoRR, abs/2006.06294.

Bibliography (cont.)

[Ménard et al., 2020] Ménard, P., Domingues, O. D., Jonsson, A., Kaufmann, E., Leurent, E., and Valko, M. (2020).

Fast active learning for pure exploration in reinforcement learning.

CoRR, abs/2007.13442.

Thank you!

Feel free to contact me for more discussions!

xut@lamda.nju.edu.cn