On Pure-Exploration in (Episodic) Markov Decision Processes

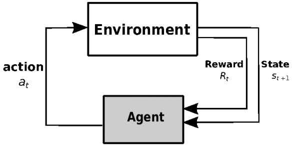
based on joint works with Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent and Michal Valko

RL Theory Seminar @ICML
July 24th, 2021

Rewards maximization versus pure exploration

RL setup: an agent interacts with an environement (MDP)



Several Performance measures:

1. the agent should *adopt* a good behavior
   - ➜ maximize the total rewards (*regret minimization*)
   - ➜ use as much as possible an ε-optimal policy (*PAC-MDP*)
2. the agent should *learn* a good behavior, regarless of rewards gathered during learning
   - ➜ Pure Exploration

Setting: Episodic Markov Decision Process

Episodic MDP: MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, H, s_1)$ for

- (finite) state space $\mathcal{S}$, action space $\mathcal{A}$
- horizon H, initial state s_1
- (inhomogeneous) transition kernel $P = (p_h(s'|s,a))_{(s,a,s',h)\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}\times[H]}$
- (inhomogeneous) reward function $r = (r_h(s,a))_{(s,a,h)\in\mathcal{S}\times\mathcal{A}\times[H]}$

Value of a policy $\pi = (\pi_h)_{h=1}^{H}$, $\pi_h : \mathcal{S} \to \mathcal{A}$:

$$V_h^\pi(s;\mathcal{M}) \triangleq \mathbb{E}\left[\sum_{\ell=h}^{H} r_\ell(s_\ell, \pi_\ell(s_\ell)) \middle| \substack{s_h = s \\ s_{\ell+1}\sim p_\ell(\cdot|s_\ell,\pi_\ell(s_\ell))}\right]$$

Optimal policy: $\pi^\star_{\mathcal{M}}$ such that $V_1^{\pi^\star_{\mathcal{M}}}(s_1;\mathcal{M}) \geq V_1^\pi(s_1;\mathcal{M})$ for any policy π.

Sequential Learning Protocol

Adaptively collect data from the MDP by generating trajectories (episodes) $\neq$ generative model

In each episode $t = 1, 2, \ldots,$ the agent

- selects an exploration policy π^t
- generates an episode under this policy

$$(s_1^t, a_1^t, r_1^t, s_2^t, a_2^t, r_2^t, \ldots, s_H^t, a_H^t, r_H^t)$$

  with $s_1^t = s_1$, $a_h^t = \pi_h^t(s_h^t)$, $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t), r_h^t = r_h(s_h^t, a_h^t)$
- can decide to stop exploration
- if decides to stop, outputs a prediction

General goal: minimize the length of the exploration phase (sample complexity) to reach an accurate prediction (ε, δ-PAC)

What is the prediction?

In each episode $t = 1, 2, \ldots,$ the agent

- selects an exploration policy π^t
- generates an episode under this policy

$$(s_1^t, a_1^t, r_1^t, s_2^t, a_2^t, r_2^t, \ldots, s_H^t, a_H^t, r_H^t)$$

  with $s_1^t = s_1$, $a_h^t = \pi_h^t(s_h^t)$, $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t), r_h^t = r_h(s_h^t, a_h^t)$
- can decide to stop exploration
- if decides to stop, outputs a prediction

Planning in MDPs

Prediction: output the best first action to take ($\simeq \pi_1^\star(s_1)$)

➜ problem-dependent sample complexity [Jonnson et al, 20]

What is the prediction?

In each episode $t = 1, 2, \ldots,$ the agent

- selects an exploration policy π^t
- generates an episode under this policy

$$(s_1^t, a_1^t, r_1^t, s_2^t, a_2^t, r_2^t, \ldots, s_H^t, a_H^t, r_H^t)$$

with $s_1^t = s_1$, $a_h^t = \pi_h^t(s_h^t)$, $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t), r_h^t = r_h(s_h^t, a_h^t)$

- can decide to stop exploration
- if decides to stop, outputs a prediction

Best Policy Identification [Fiechter, 1994], [Jin et al., 18] ...

Prediction: output the best policy $\pi^\star_{\mathcal{M}} = (\pi_1^\star, \ldots, \pi_H^\star)$

➜ UCB-VI [Azar et al., 21] + a data-dependent stopping rule [Ménard et al., 21]

What is the prediction?

In each episode $t = 1, 2, \dots,$ the agent

- selects an exploration policy π^t
- generates an episode under this policy

$$(s_1^t, a_1^t, \cancel{r_1^t}, s_2^t, a_2^t, \cancel{r_2^t}, \dots, s_H^t, a_H^t, \cancel{r_H^t})$$

with $s_1^t = s_1$, $a_h^t = \pi_h^t(s_h^t)$, $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$, $\xcancel{r_h^t = r_h(s_h^t, a_h^t)}$

- can decide to stop exploration
- if decides to stop, outputs a prediction

Reward Free Exploration [Jin et al., 20],[Wang et al. 20]

Given *any* reward function r, output $\pi^\star_{\mathcal{M}}$ for $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$

What is the prediction?

In each episode $t = 1, 2, \ldots,$ the agent

- selects an exploration policy π^t
- generates an episode under this policy

$$(s_1^t, a_1^t, \not{r}_1^t, s_2^t, a_2^t, \not{r}_2^t, \ldots, s_H^t, a_H^t, \not{r}_H^t)$$

with $s_1^t = s_1$, $a_h^t = \pi_h^t(s_h^t)$, $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$, ~~$r_h^t = r_h(s_h^t, a_h^t)$~~

- can decide to stop exploration
- if decides to stop, outputs a prediction

Reward Free Exploration [Jin et al., 20],[Wang et al. 20]

Given *any* reward function r, output $\pi_r^\star$ for $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r)$

Outline

Reward-Free Exploration (RFE)

RFE algorithm

- exploration policy π^t: may depend on past data $\mathcal{D}_{t-1}$

$$\mathcal{D}_t = \mathcal{D}_{t-1} \cup \left\{ (s_1^t, a_1^t, s_2^t, a_2^t, \ldots, s_H^t, a_H^t) \right\}$$

- stopping rule τ : stopping time w.r.t. $(\mathcal{D}_t)_{t \in \mathbb{N}}$
- prediction $\hat{P} = (\hat{p}_h(\cdot|s,a))_{h,s,a}$: a **transition kernel** that may depend on $\mathcal{D}_\tau$

$\hat{\pi}_r^\star$: optimal policy in the MDP $(\hat{P}, r)$

(ε, δ)-PAC algorithm for Reward-Free Exploration

$$\mathbb{P}\left(\text{for any } r \in \mathcal{B}, V_1^\star(s_1; r) - V_1^{\hat{\pi}_r^\star}(s_1; r) \le \varepsilon \right) \ge 1 - \delta$$

Assumption: uniformly bounded rewards
$\mathcal{B} = \{ r = (r_h(s,a)) \text{ with } r_h(s,a) \in [0,1] \}$

Sample complexity

Lower bounds For any (ε, δ)-PAC algorithm, exists an MDP:

- $\mathbb{E}[\tau] = \Omega\left(\frac{SAH^3}{\varepsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ [Darwiche Domingues et al. 21]
- $\mathbb{E}[\tau] = \Omega\left(\frac{S^2AH^2}{\varepsilon^2}\right)$ [Jin et al. 20] (homogeneous case)

Algorithms

RF-RL-EXPLORE [Jin et al. 20]	$\tau = \tilde{O}\left(\frac{H^5S^2A}{\varepsilon^2}\log\left(\frac{1}{\delta}\right) + \frac{H^7S^2A}{\varepsilon}\log^3\left(\frac{1}{\delta}\right)\right)$
RF-UCRL [Kaufmann et al. 21]	$\tau = \tilde{O}\left(\frac{H^4SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right)$, w.h.p.
RF-Express [Ménard et al. 21]	$\tau = \tilde{O}\left(\frac{H^3SA}{\varepsilon^2}\left(\log\left(\frac{1}{\delta}\right) + S\right)\right)$, w.h.p.

Remark: changing the set $\mathcal{B}$ of candidate reward functions leads to a different scaling of the sample complexity

e.g. finite set $\mathcal{B}$ [Zhang et al. 20], total bounded reward [Zhang et al. 21]

Outline

Estimate of the transition kernel

Number of visits:

$$n_h^t(s,a) = \sum_{k=1}^{t} \mathbb{1}_{\{(s_h^k, a_h^k) = (s,a)\}} \quad n_h^t(s,a,s') = \sum_{k=1}^{t} \mathbb{1}_{\{(s_h^k, a_h^k, s_{h+1}^k) = (s,a,s')\}}$$

Empirical transitions: $\hat{P}^t = \left(\hat{p}_h^t(s'|s,a)\right)_{h,s,a,s'}$

$$\hat{p}_h^t(s'|s,a) = \begin{cases} \frac{n_h^t(s,a,s')}{n_h^t(s,a)} & \text{if } n_h^t(s,a) > 0 \\ \frac{1}{S} & \text{else} \end{cases}$$

RF-Express

RF-Express

- **exploration policy:** π^{t+1} is the greedy policy w.r.t. W_h^t :

$$\forall s \in \mathcal{S}, \forall h \in [H], \quad \pi_h^{t+1}(s) = \arg\max_{a \in \mathcal{A}} W_h^t(s, a)$$

- **stopping rule:**

$$\tau = \inf\left\{ t \in \mathbb{N} : 3e\sqrt{\max_a W_1^t(s_1, a)} + \max_a W_1^t(s_1, a) \leq \varepsilon/2 \right\}$$

- **prediction:** output the empirical transition kernel $\hat{P} = \hat{P}^{\tau}$

where

$$W_h^t(s, a) = \min\left[H, 15H^2 \frac{\beta(n_h^t(s, a), \delta)}{n_h^t(s, a)} + \left(1 + \frac{1}{H}\right) \sum_{s'} \hat{p}_h^t(s'|s, a) \max_{a'} W_{h+1}^t(s', a') \right]$$

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

Theoretical guarantees

Theorem

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

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

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

Outline

Notation

Q-Values

$$Q_h^\pi(s,a) := r_h(s,a) + p_h V_{h+1}^\pi(s,a)$$
$$Q_h^\star(s,a) := r_h(s,a) + p_h V_{h+1}^\star(s,a)$$

with the notation $p_h f(s,a) = \mathbb{E}_{s' \sim p_h(\cdot|s,a)} E[f(s')]$

The Bellman equations can be expressed as

$$V_h^\pi(s) = Q_h^\pi(s,\pi(s)) \quad \text{and} \quad V_h^\star(s) = \max_{a \in \mathcal{A}} Q_h^\star(s,a)$$

Empirical values:

- $\hat{V}_h^{t,\pi}(s;r)$ values in the empirical MDP $(\mathcal{S},\mathcal{A},\hat{P}^t,r)$
- $\hat{Q}_h^{t,\pi}(s;r)$ Q-values in the empirical MDP $(\mathcal{S},\mathcal{A},\hat{P}^t,r)$

Upper bounding the error

Sufficient condition

A sufficient condition to be (ε, δ)-PAC for RFE is to have accurate estimates of the value function for all π and r:

$$\mathbb{P}\left(\forall \pi, \forall r,\ |\hat{V}_1^{\tau,\pi}(s_1; r) - V_1^{\pi}(s_1; r)| \leq \varepsilon/2\right) \geq 1 - \delta.$$

Proof.

$$\begin{aligned} V_1^{\star}(s_1; r) - V_1^{\hat{\pi}_r^{\star}}(s_1; r) \quad = \quad & V_1^{\pi_r^{\star}}(s_1; r) - \hat{V}_1^{t,\pi_r^{\star}}(s_1; r) + \underbrace{\hat{V}_1^{t,\pi_r^{\star}}(s_1; r) - \hat{V}_1^{t,\hat{\pi}_r^{\star}}(s_1; r)}_{\leq 0} \\ & + \hat{V}_1^{t,\hat{\pi}_r^{\star}}(s_1; r) - V_1^{\hat{\pi}_r^{\star}}(s_1; r). \end{aligned}$$

Rationale for the stopping rule: Introducing the estimation error

$$\hat{e}_h^{t,\pi}(s, a; r) := \left|\hat{Q}_h^{t,\pi}(s, a; r) - Q_h^{\pi}(s, a; r)\right|,$$

we want to stop when $\max_{\pi,r} \hat{e}_1^{t,\pi}(s, \pi(s_1); r) \leq \varepsilon/2$.

Upper bounding the error

Lemma

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

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

- ➜ data-dependent upper bound, independent of π and r
- ➜ justifies the stopping rule

$$\tau = \inf\left\{t \in \mathbb{N} : 3e\sqrt{\max_a W_1^t(s_1, a)} + \max_a W_1^t(s_1, a) \leq \varepsilon/2\right\}$$

- ➜ note that the bound on $\hat{e}_h^{t,\pi}(s, a; r)$ is only valid for $h = 1$

Proof sketch

$$\hat{e}_h^{t,\pi}(s,a;r) := \left|\hat{Q}_h^{t,\pi}(s,a;r) - Q_h^\pi(s,a;r)\right|$$

Writing the Bellman equations

$$\begin{aligned} \hat{Q}_h^{t,\pi}(s,a;r) &= r_h(s,a) + \hat{p}_h^t \hat{V}_{h+1}^{t,\pi}(s,a) \\ \text{and} \quad Q_h^\pi(s,a;r) &= r_h(s,a) + p_h V_{h+1}^\pi(s,a)\,. \end{aligned}$$

the reward cancel and one obtains

$$\begin{aligned} \hat{e}_h^{t,\pi}(s,a;r) &\le \left|(\hat{p}_h^t - p_h)V_{h+1}^\pi(s,a)\right| + \hat{p}_h^t\left|\hat{V}_{h+1}^{t,\pi} - V_{h+1}^\pi\right|(s,a) \\ &= \underbrace{\left|(\hat{p}_h^t - p_h)V_{h+1}^\pi(s,a)\right|}_{\substack{\text{upper bound bound this using} \\ \text{an empirical Bernstein inequality}}} + \hat{p}_h^t \pi_{h+1}\hat{e}_{h+1}^{t,\pi}(s,a;r). \end{aligned}$$

with the notation $\pi_{h+1}g(s) = g(s,\pi_{h+1}(s))$

Proof sketch

On the event $\mathcal{E} = \left\{ \mathrm{KL}\left(\hat{p}_h^t(s,a), p_h(s,a)\right) \le \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$, we prove

$$\hat{e}_h^{t,\pi}(s,a;r) \le \underbrace{3\sqrt{\frac{\mathrm{Var}_{\hat{p}_h^t}(\hat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right)} + 15H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}_{(*)}$$

$$\left(1+\frac{1}{H}\right)\hat{p}_h^t\pi_{h+1}\hat{e}_{h+1}^{t,\pi}(s,a;r).$$

Challenge: the empirical variance term depends on the *unobserved* reward function

➜ $(*)$ *cannot* be computed by an algorithm

Proof sketch

On the event $\mathcal{E} = \left\{ \mathrm{KL}\left(\hat{p}_h^t(s,a), p_h(s,a)\right) \le \frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \right\}$, we prove

$$\hat{e}_h^{t,\pi}(s,a;r) \le \underbrace{3\sqrt{\frac{\mathrm{Var}_{\hat{p}_h^t}(\hat{V}_{h+1}^{t,\pi})(s,a;r)}{H^2}\left(\frac{H^2\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\right)} + 15H^2\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}}_{(*)}$$

$$\left(1+\frac{1}{H}\right)\hat{p}_h^t\pi_{h+1}\hat{e}_{h+1}^{t,\pi}(s,a;r).$$

Challenge: the empirical variance term depends on the *unobserved* reward function

➜ $(*)$ *cannot* be computed by an algorithm

Solution: splitting the bonus

Proof sketch

$$\hat{e}_h^{t,\pi}(s,a;r) \le Y_h^{t,\pi}(s,a;r) + W_h^{t,\pi}(s,a)$$

where

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

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

Using notably a Bellman equation for the variance, we prove

$$Y_1^{t,\pi}(s_1,\pi_1(s_1);r) \le 3e\sqrt{W_1^{t,\pi}(s_1,\pi_1(s_1))}$$

$$\Rightarrow \hat{e}_1^{t,\pi}(s,\pi_1(s);r) \le 3e\sqrt{\max_a W_1^t(s_1,a)} + \max_a W_1^t(s,a)$$

Sample complexity: how do we get rid of an H?

By definition of the stopping rule, for any $t < \tau$,

$$\varepsilon \leq 3e\sqrt{W_1^t(s_1, \pi_1^{t+1}(s_1))} + W_1^t(s_1, \pi_1^{t+1}(s_1)).$$

Summing these inequalities yields

$$\tau\varepsilon \leq 3e\sqrt{\tau \underbrace{\sum_{t=0}^{\tau-1} W_1^t(s_1, \pi_1^{t+1}(s_1))}_{(*)}} + \underbrace{\sum_{t=0}^{\tau-1} W_1^t(s_1, \pi_1^{t+1}(s_1))}_{(*)}.$$

A careful sum of the bonuses (w.h.p.)

$$(*) \quad \leq \quad CH^2 \sum_{t=0}^{\tau-1} \sum_{h=1}^{H} \sum_{s,a} p_h^{t+1}(s,a) \frac{\beta(\overline{n}_h^t(s,a), \delta)}{\overline{n}_h^t(s,a) \vee 1}$$

$$\simeq \quad C' H^3 SA \log(\tau) \beta(\tau, \delta)$$

Is $1/n$ any good in practice?

RF-Express ($1/n$ bonuses) versus RF-UCRL ($1/\sqrt{n}$) in a grid-world environment with 15 rooms and 25 states per room

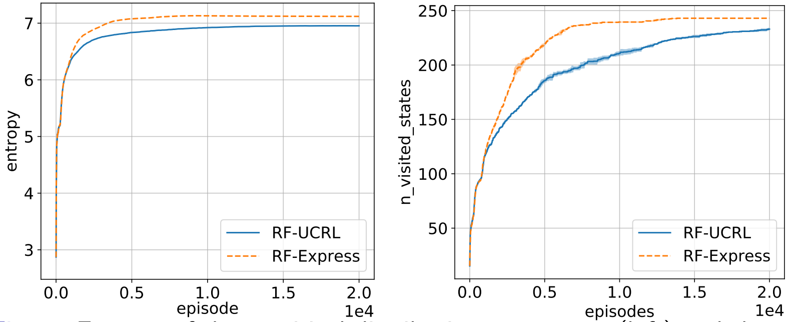


Figure: Entropy of the empirical distribution over states (left) and the number of visited states (right) versus number of episodes.
Horizon $H = 30$, average over 4 runs.

Conclusion

- We got rid of an H in the sample complexity of reward-free exploration algorithms an optimal algorithm
- ... which showcases $1/n$ bonuses

Open questions:

- Can (a variant of) RF-Express attain a *horizon-free* sample complexity under the total bounded reward assumption?
- What are the practical benefits of using $1/n$ bonuses instead of $1/\sqrt{n}$?

see, Domingues et al., *Density-Based Bonuses on Learned Representations for RFE in Deep Reinforcement Learning* @ Unsupervised RL workshop

- Optimal Best Policy Identification in a minimax and problem-dependent sense ?

References

- Azar et al., *Minimax Regret Bounds for Reinforcement Learning*, ICML 2017
- Darwiche Domingues et al., *Episodic Reinforcement Learning in Finite MDPs: Minimax Lower Bounds Revisited* , ALT 2021
- Jin et al., *Reward-Free Exploration for Reinforcement Learning*, ICML 2020
- Jin et al., *Is Q-Learning Provably Efficient?*, NeurIPS 2018
- Jonsson et al., *Planning in Markov Decision Processes with Gap-Dependent Sample Complexity*, NeurIPS 2020
- Fiechter, *Efficient Reinforcement Learning*, COLT 1994
- Kaufmann et al., *Adaptive Reward-Free Exploration*, ALT 2021
- Ménard et al., *Fast active learning for pure exploration in reinforcement learning*, ICML 2021
- Wang et al., *On reward-free reinforcement learning with linear function approximation*, NeurIPS 2020
- Zhang et al., *Task-agnostic exploration in reinforcement learning*, NeurIPS 2020
- Zhang et al., *Nearly Optimal Reward-Free Reinforcement Learning*, ICML 2021