Adaptive Approximate Policy Iteration

Botao Hao, Nevena Lazić, Yasin Abbasi-Yadkori,
Pooria Joulani, Csaba Szepesvári

DeepMind

Problem Setting

- Markov decision process (MDP) $(\mathcal{X}, \mathcal{A}, r, P)$: observed states $x \in \mathcal{X}$, discrete actions $a \in \{1, ..., |\mathcal{A}|\}$, unknown reward $r(x, a)$ and dynamics $P(\cdot|x, a)$.
- $\pi(\cdot|x)$: policy, distribution over actions in state x.
- $Q_\pi(x, a)$: value of taking action a in state x and then following π.
- **Infinite-horizon undiscounted** setting (average reward), ergodic MDP,

$$J_\pi = \mathbb{E}\Big[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r(x_t, a_t)\Big].$$

- Online single-trajectory learning, analyze regret

$$\text{Regret}_T := \sum_{t=1}^{T} \big(J_* - r(x_t, a_t)\big).$$

where J_* is the average reward of the optimal policy.

Related Work

- Most algorithms with regret guarantees require finite state-action space (tabular MDP), finite horizon, or both

  e.g. REGAL (Bartlett & Tewari 2009), UCRL2 (Jaksch et al. 2010), RLSVI (Osband et al. 2016), SCAL (Fruit et al. 2018), Q-learning (Wei et al. 2019), LSVI (Yang & Wang 2019), OPPO (Cai et al. 2019)...
- Some recent results on infinite-horizon linear MDPs (Wei et al. 2020).

Our work:

- Infinite-horizon, possibly infinite state-space, function approximation
- A variant of approximate policy iteration with sublinear regret

⋆ AAPI: policy iteration with adaptive per-state KL regularization, $\text{Regret}_T = O(T^{2/3})$

Adaptive Approximate Policy Iteration (AAPI)

Initialize $\pi_1(\cdot|x) = \text{Uniform}(\mathcal{A})$, $\widehat{Q}_{\pi_0}(x, a) = 0$.

For $k \in 1, \dots, K = \lfloor T/\tau \rfloor$:

Policy evaluation: execute π_k for τ steps and estimate Q_{π_k}.

Policy improvement: adaptive optimistic FTRL (Mohri & Yang 2016)

$$\pi_{k+1}(\cdot|x) = \underset{\pi\in\Delta_{\mathcal{A}}}{\text{argmax}} \Big\langle \pi, \sum_{i=1}^{k} \hat{Q}_{\pi_i}(x,\cdot) + M_{k+1}(x,\cdot) \Big\rangle - \eta_k(x) R(\pi),$$

where $\eta_k^{-1}(x) = \eta^{-1}\sqrt{2\sum_{i=1}^{k}\Big\|\widehat{Q}_{\pi_i}(x,\cdot) - \widehat{Q}_{\pi_{i-1}}(x,\cdot)\Big\|_\infty^2}$, $R(\pi)$: negative entropy.

- **Observation:** losses (Q-function estimates) are slow-changing.
- Choose $M_{k+1}(x, \cdot) = \hat{Q}_{\pi_k}(x, \cdot)$.

Regret Bound

Condition (policy evaluation error). For each phase $k \in [K]$, denote $D_{\pi_k} = \hat{Q}_{\pi_k} - Q_{\pi_k}$. We assume the following holds with probability $1 - \delta$,

$$\max\Big\{\|D_{\pi_k}\|_{\mu_{\pi^*}\otimes\pi^*}, \|D_{\pi_k}\|_{\mu_{\pi^*}\otimes\pi_k}, \|D_{\pi_k}\|_\infty\Big\} \le \varepsilon_0 + \tilde{C}\sqrt{\frac{\log(1/\delta)}{\tau}},$$

where ε_0 is the irreducible approximation error and $\tilde{C}$ is a problem dependent constant. Additionally, there exists a constant b such that $\hat{Q}_{\pi_k}(x, a) \in [b, b + Q_{\max}]$ for any pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ and $k \in [K]$. Here, μ_{π^*} is the stationary distribution of π^* over the states.

Theorem. Consider an ergodic MDP and suppose the policy evaluation error condition holds. By choosing the phase length $\tau = (\tilde{C}/\rho t_{\text{mix}}^3)^{2/3}T^{2/3}$, we have with probability at least $1 - 1/T$,

$$R_T = \tilde{O}\left(t_{\text{mix}}^2(\rho\tilde{C}^2)^{1/3}T^{2/3} + T\varepsilon_0\right),$$

where ρ is the distribution mismatch coefficient and $\tilde{O}(\cdot)$ hides universal constants and poly-logarithmic factors.

Proof Hints

- **Regret decomposition** Since the policy is only updated at the end of each phase of length τ, we have $\pi_t = \pi_k$ for $t \in \{\tau(k-1), \dots, \tau k\}$. Thus, the pseudo-regret term can be rewritten as

$$\sum_{t=1}^{T}\Big(J_* - J_{\pi_t}\Big) = \tau\sum_{k=1}^{K}\Big(J_* - J_{\pi_k}\Big).$$

By the **performance difference lemma**, we have

$$J_* - J_{\pi_k} = \langle \mu_{\pi^*}, Q_{\pi_k}(\cdot, \pi_*) - Q_{\pi_k}(\cdot, \pi_k)\rangle .$$

Bridging by empirical estimations, we decompose it into $R_{1T} + R_{2T}$, where

$$R_{1T} = \tau\sum_{k=1}^{K}\Big\langle \mu_{\pi^*}, Q_{\pi_k}(\cdot, \pi_*) - \hat{Q}_{\pi_k}(\cdot, \pi_*)\Big\rangle + \tau\sum_{k=1}^{K}\Big\langle \mu_{\pi^*}, \hat{Q}_{\pi_k}(\cdot, \pi_k) - Q_{\pi_k}(\cdot, \pi_k)\Big\rangle,$$

$$R_{2T} = \tau\sum_{k=1}^{K}\Big\langle \mu_{\pi^*}, \hat{Q}_{\pi_k}(\cdot, \pi^*) - \hat{Q}_{\pi_k}(\cdot, \pi_k)\Big\rangle.$$

Remark. **R_{1T} is bounded by policy evaluation error.**

- **Online learning reduction.** Minimizing R_{2T} can be cast into an online learning problem. For each state $x \in \mathcal{X}$, we view $\pi_k(\cdot|x)$ as the prediction vector and $\hat{Q}_{\pi_k}(x, \cdot)$ as the loss vector. At each round t, adaptive optimistic FTRL has the following form:

$$f_{t+1} = \underset{f\in\mathcal{F}}{\text{argmin}}\langle f, \sum_{s=1}^{t} q_s + M_{t+1}\rangle + \eta_t\mathcal{R}(f), \eta_t = \eta\sqrt{\sum_{s=1}^{t}\|q_s - M_s\|_*^2},$$

Lemma. Choose $\eta = \sqrt{2/\mathcal{R}(f^*)}$ and denote $R_{\max} = \max_f \mathcal{R}(f)$. The cumulative regret for AO-FTRL is upper-bounded by

$$\tilde{R}_T \le \sqrt{2R_{\max}\sum_{t=1}^{T}\|q_t - M_t\|_*^2} - \sum_{t=1}^{T}\frac{\eta_t}{4}\|f_t - f_{t+1}\|^2 + \langle M_{T+1}, f^* - f_{T+1}\rangle.$$

Choose: $q_t = Q_{\pi_k}, M_t = Q_{\pi_{k-1}}, f_t = \pi_{k-1}(\cdot|x), f_{t+1} = \pi_k(\cdot|x)$.

- **A key observation:** For any two successive policies π_{k-1} and π_k, the following holds for any state-action pair (x, a),

$$\Big|Q_{\pi_k}(x, a) - Q_{\pi_{k-1}}(x, a)\Big| \le t_{\text{mix}}^2\log_2^2(K)\max_x\big\|\pi_{k-1}(\cdot|x) - \pi_k(\cdot|x)\big\|_1 + \frac{2}{K^3}.$$