Adaptive Approximate Policy Iteration
Botao Hao, Nevena Lazić, Yasin Abbassi-Yadkori, Pooria Joulani, Csaba Szepesvári

**Problem Setting**
- Markov decision process (MDP): $(X, A, r, P)$; observed states $x \in X$, discrete actions $a \in \{1, \ldots, |A|\}$, unknown reward $r(x, a)$ and dynamics $P(x'|x, a)$.
- $\pi(x)$: policy, distribution over actions in state $x$.
- $Q(x, a)$: value of taking action $a$ in state $x$ and then following $\pi$.
- Infinite-horizon undiscounted setting (average reward), ergodic MDP, $Q$-function, finite horizon, or both.

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

Online single-trajectory learning, analyze regret $\text{Regret}_T := \sum_{t=1}^{T} (J_t - r(x_t, a_t))$.

where $J_t$ 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 (Osebi 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 $\pi_1(x) = \text{Uniform}(A)$, $\hat{Q}_{\pi_1}(x, a) = 0$.

For $k = 1, \ldots, K = \lceil T/\tau \rceil$:
- **Policy evaluation**: execute $\pi_k$ for $\tau$ steps and estimate $Q_{\pi_k}$.
- **Policy improvement**: adaptive optimistic FTRL (Mohri & Yang 2016)

$$\pi_{k+1}(x) = \arg\max \left\{ \pi, \sum_{x} Q_{\pi}(x, a) + M_{\pi}(x, a) \right\} - q_k(x) R(x),$$

where $q_k(x) = \eta_k \cdot \sqrt{\sum_{x} \left| \hat{Q}_{\pi_k}(x, :)-Q_{\pi_k}(x, :)) \right|^2_R(x)}$ negative entropy.

- **Observation**: losses (Q-function estimates) are slow-changing.
- Choose $M_{\pi_k}(x, :)=Q_{\pi_k}(x, :)$.

**Regret Bound**

**Condition (policy evaluation error)**

For each phase $k \in [K]$, denote $D_{\pi_k} = Q_{\pi_k} - Q_{\pi_k}$. We assume the following holds with probability $1-\delta$,

$$\max \left\{ \|D_{\pi_{k+1}}\|_{\text{KL}}, \|D_{\pi_{k+1}}\|_{\text{lin}}, \|D_{\pi_{k+1}}\|_{\text{lip}} \right\} \leq c_0 + C \sqrt{\log(1/\delta)} / \tau,$$

where $c_0$ is the irreducible approximation error and $C$ is a problem dependent constant.

Additionally, there exists a constant $\delta$ such that $Q_{\pi_k}(x, a) \in [h, h+M_{\pi_k}]$ for any pair $(x, a) \in X \times A$ and $k \in [K]$. Here, $\pi_k$ is the stationary distribution of $\pi_k$ over the states.

**Theorem.** Consider an ergodic MDP and suppose the policy evaluation error condition holds. By choosing the phase length $\tau = (C/\mu_{\min})^{1/2} (20C^2)^{1/3}$, we have with probability at least $1-1/T$,

$$R_T = O \left( \epsilon_{\text{det}}^{2/3} (20C^2)^{1/3} + Tc_0 \right),$$

where $\epsilon$ is the distribution mismatch coefficient and $O(\cdot)$ 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 $\tau$, we have $\pi_t = \pi_k$ for $t \in (\tau(k-1), \tau(k))$. Thus, the pseudo-regret term can be rewritten as

$$\sum_{t=1}^{T} (J_t - J_k).$$

By the performance difference lemma, we have

$$J_t - J_k = (J_t - J_k) + (J_k - J_k).$$

Bridging by empirical estimations, we decompose it into $R_{\text{FTL}} + R_{\text{FTR}}$, where

$$R_{\text{FTL}} = \tau \sum_{k=1}^{K} (\mu_{\text{f}}, Q_{\pi_k}(x, a) - Q_{\pi_k}(x, a)) + \tau \sum_{k=1}^{K} (\mu_{\text{f}}, Q_{\pi_k}(x, a) - Q_{\pi_k}(x, a)),$$

$$R_{\text{FTR}} = \tau \sum_{k=1}^{K} (\mu_{\text{f}}, Q_{\pi_k}(x, a) - Q_{\pi_k}(x, a)).$$

**Remark.** $R_T$ is bounded by policy evaluation error.

- **Online learning reduction.** Minimizing $R_T$ can be cast into an online learning problem. For each state $x \in X$, we view $\pi_t(x)$ as the prediction vector and $Q_{\pi_t}(x, a)$ as the loss vector. At each round $t$, adaptive optimistic FTRL has the following form:

$$f_{t+1} = \arg\min_{f_{t+1}} \left\{ \sum_{s=1}^{T} q_s + M_{\pi_t} \right\} + \eta T R(f), \eta = \eta \sum_{t=1}^{T} \|q_t - M_{\pi_t}\|^2_{\text{F}}.$$

**Lemma.** Choose $\eta = \sqrt{2T R(f)}$. Denote $R_{\text{FTR}} = \max_{T} R(f)$. The cumulative regret for AO-FTRL is upper-bounded by

$$R_T \leq 2\sum_{t=1}^{T} \|q_t - M_{\pi_t}\|_{\text{F}}^2 + \sum_{t=1}^{T} \|f_t - f_{t+1}\|_{\text{F}}^2 + (M_{\pi_t}, f^* - f_{t+1}).$$

**Choose:** $q_1 = Q_{\pi_1}, M = Q_{\pi_1}, f_1 = \pi_{x-1}(x), f_{t+1} = \pi_{x}(x)$.

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

$$\left| Q_{\pi_{x-1}}(x, a) - Q_{\pi_{x}}(x, a) \right| \leq \epsilon_{\text{det}}^{2/3} (20C^2)^{1/3} + T c_0.$$