Motivation

In many RL domains, executing a new policy is expensive. **Off-policy RL**: Find $\pi_*$ given only off-line data from $\mu$.

Challenging... Relax

**Limited Adaptivity RL**: Find $\pi_*$ with online data from $\{\mu_1, \ldots, \mu_n\}$ for some small $n$.

Local Policy Switch

Setup: episodic MDP with horizon $H$, play $K$ episodes

Def: The number of local policy switches for an RL algorithm is

$$N_{\text{switch}} = \sum_{k=1}^{K} \left| \{ (h, s) : \pi^k_h(s) \neq \pi^{k+1}_h(s) \} \right|$$

where $\pi^k_h$ is the (deterministic) policy it plays at episode $K$.

Smaller $N_{\text{switch}} \Rightarrow$ closer to off-policy

Prior work: Q-Learning with UCB-Hoeffding exploration\textsuperscript{(1)}:

$$\tilde{O}(\sqrt{H^4SAT})$$ regret, but $N_{\text{switch}} = \Theta(HSK)$ linear in $K$ 😞

Any sublinear regret algo such that $N_{\text{switch}}$ sublinear in $K$?

Algorithm: Q-Learning with UCB2 Scheduling

Idea: update the policy according to $Q$ only when $Q$ has been updated $\tau(r) = (1 + \alpha)^r$ times.

**Algorithm 2** Q-learning with UCB2 scheduling

input  Parameter $\alpha \in (0, 1)$ and $c > 0$.

Initialize: $Q_h(x, a) \leftarrow H$, $Q_h \leftarrow Q_h^0$, $N_h(x, a) \leftarrow 0$ for all $(x, a, h) \in S \times A \times [H]$.

for episode $k = 1, \ldots, K$ do

Receive $x_1$.

for step $h = 1, \ldots, H$ do

Take action $a_h \leftarrow \arg \max_{a'} Q_h(x_h, a')$, and observe $x_{h+1}$.

$\tilde{Q}_h(x_h, a_h) \leftarrow (1 - \alpha_t)\tilde{Q}_h(x_h, a_h) + \alpha_t [r_h(x_h, a_h) + \tilde{V}_{h+1}(x_{h+1}) + b_t]$. // Update $Q \setminus Q_h$ via Q-Learning

$\tilde{V}_h(x_h) \leftarrow \min \{ H, \max_{a' \in A} Q_h(x_h, a') \}$

if $t = \tau(r)$ for some $r$ then

(Update policy) $Q(x_h, \cdot) \leftarrow \tilde{Q}(x_h, \cdot)$. // Set $Q$ to be $\tilde{Q}$ occasionally according to UCB2 scheduling

end if

end for

end for

Theoretical Result

**Theorem 1**: Q-Learning with UCB2 scheduling achieves regret $\tilde{O}(\sqrt{H^4SAT})$ and policy switch bound

$$N_{\text{switch}} \leq O(H^3SA \log(K/A))$$

which is logarithmic in $K$. 😊

Proof highlight: improved propagation of error argument under delayed Q updates.

“**Theorem**” 2 (lower bound): Any sublinear regret or PAC algorithm must have

$$N_{\text{switch}} \geq \Omega(HSA)$$

Mild gap: $O(H^2 \log(K/A))$, conjecture that log is also necessary & gap is at most $O(H^2)$

Discussion & Future Work

- Algorithms with tighter regret bounds (e.g. tighten the $H^4$)?
- Close the gap between lower and upper bound.
- Model-based algorithms with limited adaptivity? Better bounds?

References