CS292F Statistical Foundation of Reinforcement Learning **Spring 2021**

Lecture 9: Exploration in Tabular MDPs, April 26

Lecturer: Yu-Xiang Wang *Scribes: Ming Min*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

9.1 Exploration in Tabular MDPs

We now move to the learning in an episodic finite-horizon MDP with non-stationary transitions, i.e. in every episode k, the learner acts for H step starting from a fixed starting state $s_0 \sim \mu$ and, at the end of the H-length episode, the state is reset. $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \{r_h\}_h, \{P_h\}_h, H, s_0\}$ and $\pi = \{\pi_0, \ldots, \pi_{H-1}\}$ depends on time step.

Regret definition:

$$\text{Regret} := \mathbb{E}\left[\sum_{k=0}^{K-1} \text{Regret}_k\right] = \mathbb{E}\left[KV^*(s_0) - \sum_{k=0}^{K-1}\sum_{h=0}^{H-1} r(s_h^k, a_h^k)\right],$$

where the goal of the agent is to minimize her expected cumulative regret over K episodes.

9.2 UCB-VI

9.2.1 Algorithm

UCB-VI algorithm is a model-based approach and requires estimating P. It repeats the following procedure for K episodes:

1. Compute $\hat{P}_h^k$ as the empirical estimates, for all h. It is defined by

$$\text{At } k, h: \quad \hat{P}_h^k(s'|s,a) = \frac{N_h^k(s,a,s')}{N_h^k(s,a)},$$

   where $N_h^k(s,a,s') = \{\text{the number of times these triplets appear from step h to h+1}\} = \sum_{i=0}^{k-1} \mathbb{1}(S_h^i = s, A_h^i = a, S_{h+1}^i = s')$; $N_h^k(s,a) = \sum_{i=1}^{k-1} \mathbb{1}(S_h^i = s, A_h^i = a)$. If there is no state-action pairs, we assume $0/0 := 0$.

2. Compute reward bonus b_h^k for all h, where

$$b_h^k(s,a) = H\sqrt{\frac{L}{N_h^k(s,a)}}, \quad \text{with } L = \log(SAHK/\delta),\ \delta \text{ is the failure probability.}$$

Remark: This Hoeffding style bonus encourages exploring new state-action pairs.

3. Run Value-Iteration on $\{\hat{P}^k_h, r + b^k_h\}_{h=0}^{H-1}$. Starting at H, we perform dynamic programming all the way to $h = 0$:

$$\hat{V}^n_H(s) = 0, \forall s,$$
$$\hat{Q}^n_h(s,a) = \min\{r_h(s,a) + b^n_h(s,a) + \hat{P}(\cdot|s,a) \cdot \hat{V}^n_{h+1}, H\},$$
$$\hat{V}^n_h(s) = \max_a \hat{Q}^n_h(s,a), \quad \pi^n_h(s) = \arg\max_a \hat{Q}^n_h(s,a), \forall\, h, s, a.$$

Remark: It converges in H steps and produces a non-stationary policy indexed by h.

4. Set π^k as the returned policy of VI.

9.2.2 Regret Bound of UCB-VI

Theorem 1. *(Regret Bound of UCB-VI). UCB-VI achieves the following regret bound:*

$$\text{Regret} := \mathbb{E}\left[\sum_{k=0}^{K-1}\left(V^* - V^{\pi^k}\right)\right] \leq 2H^2 S\sqrt{AK \cdot \log(SAH^2K^2)} = \tilde{O}\left(H^2 S\sqrt{AK}\right).$$

Remark. *The regret is not optimal in H, S, but is a simple analysis to start. Ideas for improving it include improving H by using Bernstein's inequality and including S using lemma 3.*

We prove the above theorem in the following with some lemmas introduced first.

Lemma 2. *With probability at least $1-\delta$, for all h, k, s, a,*

$$||\hat{P}^k_h(\cdot|s,a) - P^*_h(\cdot|s,a)||_1 \leq \sqrt{\frac{S\log(SAHK/\delta)}{N^k_h(s,a)}}.$$

Lemma 3. *With probability at least $1-\delta$, for all h, k, s, a,*

$$|\hat{P}^k_h(\cdot|s,a) \cdot V^*_{h+1} - P^*_h(\cdot|s,a) \cdot V^*_{h+1}| \leq H\sqrt{\frac{L}{N^k_h(s,a)}}, \quad L = \log(SAHK/\delta).$$

From above, we know that the probability of the inequalities fail is 2δ, i.e., $P(\text{Fail}) \leq 2\delta$.

Lemma 4. *(Optimism). Assume the above inequality in Lemma 3 is true. For all episode k, we have:*

$$\hat{V}^k_h \geq V^*_h, \quad \forall\, h = 0, 1, \ldots, H-1, H.$$

Proof. Prove via induction.
Base: $\hat{V}^k_H = V^*_H = 0$.
Assume for h, $\hat{V}^k_h \geq V^*_h$, we will prove that $\hat{V}^k_{h-1} \geq V^*_{h-1}$. Note that $\hat{V}^k_{h-1} = \max_a \hat{Q}^k_{h-1}(\cdot, a)$, and

$$\hat{Q}^k_{h-1}(s,a) = \min\{H, r_{h-1}(s,a) + b^k_{h-1}(s,a) + \hat{P}^k_{h-1}(\cdot|s,a) \cdot \hat{V}^k_h\}$$
$$Q^*_{h-1}(s,a) = r_{h-1}(s,a) + b^k_{h-1}(s,a) + P^*_{h-1}(\cdot|s,a) \cdot V^*_h.$$

- When H is smaller: $\hat{Q}^k_{h-1}(s,a) = H \geq Q^*_{h-1}(s,a)$.

- When H is not selected:

$$\begin{aligned}\hat{Q}^k_{h-1}(s,a) - Q^*_{h-1}(s,a) &= b^k_{h-1}(s,a) + \hat{P}^k_{h-1}(\cdot|s,a)\cdot \hat{V}^k_h - P^*_{h-1}(\cdot|s,a)\cdot V^*_h \\ &\geq b^k_{h-1}(s,a) + \left(\hat{P}^k_{h-1}(\cdot|s,a) - P^*_{h-1}(\cdot|s,a)\right)\cdot V^*_h \\ &\geq b^k_{h-1}(s,a) - H\sqrt{\frac{L}{N^k_{h-1}(s,a)}} \\ &\geq 0,\end{aligned}$$

where the first inequality is from the inductive hypothesis, and the second is by lemma 3.

- Thus for any s,

$$\hat{V}^k_{h-1}(s) = \max_a \hat{Q}^k_{h-1}(s,a) \geq \hat{Q}^k_{h-1}(s,a^*) \geq Q^*_{h-1}(s,a^*) = V^*_{h-1}(s).$$

□

Finally, we can prove the main theorem for the regret bound.

Proof. **Proof of Theorem 1.**
Recall the finite horizon simulation lemma from HW1 Q5:

$$\hat{V}^\pi_0 - V^\pi_0 = \sum_{h=0}^{H-1} \mathbb{E}^\pi\left[\hat{r}^\pi_h(S_h) - r^\pi(S_h) + \left(\hat{P}^\pi_h(\cdot|S_h) - P^\pi_h(\cdot|S_h)\right)\cdot \hat{V}^\pi_{h+1}(\cdot)\right].$$

Then the regret in the k-th episode:

$$\begin{aligned}\text{Regret}_k &= V^*_0(s_0) - V^{\pi_k}_0(s_0) \\ \text{(by optimism)} &\leq \hat{V}^{\pi_k}_0(s_0) - V^{\pi_k}_0(s_0) \\ \text{(by simulation lemma)} &\leq \sum_{h=0}^{H-1}\mathbb{E}^{\pi_k}\left[\hat{r}_h(S_h,A_h) - r(S_h,A_h) + \left(\hat{P}_h(\cdot|S_h,A_h) - P_h(\cdot|S_h,A_h)\right)\cdot \hat{V}^{\pi_k}_{h+1}\right] \\ &= \sum_{h=0}^{H-1}\mathbb{E}^{\pi_k}\left[b^k_h(S_h,A_h) + \left(\hat{P}_h(\cdot|S_h,A_h) - P_h(\cdot|S_h,A_h)\right)\cdot \hat{V}^{\pi_k}_{h+1}\right] \\ &\leq \sum_{h=0}^{H-1}\mathbb{E}^{\pi_k}\left[2H\sqrt{\frac{SL}{N^k_h(S_h,A_h)}}\right] \\ &= 2H\sqrt{SL}\,\mathbb{E}\left[\sum_{h=0}^{H-1}\sqrt{\frac{1}{N^k_h(S^k_h,A^k_h)}}\middle|\text{hist}_k\right],\end{aligned}$$

where in the last term the expectation is taken with respect to the trajectory and condition on all history $H(<k)$ up to and including the end of episode $k-1$. The last inequality is by lemma 2,

$$\left(\hat{P}_h(\cdot|S_h,A_h) - P_h(\cdot|S_h,A_h)\right)\cdot \hat{V}^{\pi_k}_{h+1} \leq ||\hat{P}_h(\cdot|S_h,A_h) - P_h(\cdot|S_h,A_h)||_1 ||\hat{V}^{\pi_k}_{h+1}||_\infty \leq \sqrt{\frac{SL}{N^k_h(S_h,A_h)}}\cdot H.$$

Then the total regret:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{K-1} \text{Regret}_k\right] &= \mathbb{E}\left[\sum_{k=0}^{K-1} V^*(s_0) - V^{\pi_k}(s_0)\right] \\
&= \mathbb{E}\left[\left(\sum_{k=0}^{K-1} V^*(s_0) - V^{\pi_k}(s_0)\right) \mathbb{1}(\text{Not Fail})\right] + \mathbb{E}\left[\left(\sum_{k=0}^{K-1} V^*(s_0) - V^{\pi_k}(s_0)\right) \mathbb{1}(\text{Fail})\right] \\
&\leq \mathbb{E}\left[\left(\sum_{k=0}^{K-1} V^*(s_0) - V^{\pi_k}(s_0)\right) \mathbb{1}(\text{Not Fail})\right] + 2\delta \cdot K \cdot H \\
&\leq 2H\sqrt{SL} \cdot \mathbb{E}\left[\sum_{k=0}^{K-1} \sum_{h=0}^{H-1} \frac{1}{\sqrt{N_h^k(S_h^k, A_h^k)}}\right] + 2\delta \cdot KH.
\end{aligned}$$

The expectation in the first term $= \sum_{h=0}^{H-1} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{i=1}^{N_h^k(s,a)} \frac{1}{\sqrt{i}} \leq \sum_{h=0}^{H-1} \sum_{(s,a)} 2\sqrt{N_h^k(s,a)}$ from last lecture. We conclude that

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=0}^{K-1} \text{Regret}_k\right] &\leq 2H\sqrt{SL} \cdot \mathbb{E}\left[\sum_{h=0}^{H-1} \sum_{(s,a)} 2\sqrt{N_h^k(s,a)}\right] + 2\delta \cdot KH \\
&\leq 2H\sqrt{SL} \cdot 2 \sum_{h=0}^{H-1} \sqrt{SA \cdot \sum_{(s,a)} N_h^k(s,a)} + 2\delta \cdot KH \\
&\leq 2H\sqrt{SL} \cdot 2H\sqrt{SAK} + 2\delta \cdot KH \\
&\leq 4H^2 S\sqrt{AKL} + 2\delta \cdot KH \\
&= \tilde{O}(H^2 S\sqrt{AK}), \quad \text{choose } \delta = \frac{1}{KH}.
\end{aligned}$$

□