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

Lecture 10: Exploration in Linear MDPs, April 28

Lecturer: Yu-Xiang Wang *Scribes: Yichen Feng*

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

10.1 Recap

1. Exploration in Tabular MDPs:
   - Problem setup: 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.

2. UCB-VI (Azar et al., 2017)
   A model-based approach; requires estimating P. It repeats the following procedure for K episodes:
   (a) Compute $\hat{P}_h^k$ as the empirical estimates, for all h;
   (b) Compute reward bonus b_h^k for all h;
   (c) Run Value-Iteration on $\{\hat{P}_h^k, r + b_h^k\}_{h=0}^{H-1}$;
   (d) Set π^k as the returned policy of VI.

3. Proof of Regret Bound of UCBVI: $\tilde{O}\left(H^2 S\sqrt{AK}\right)$.

10.2 Linear MDPs

The structural assumption is a linear structure in both reward and the transition.

Definition 1. *The transition and and reward of a linear MDP follows:*

$$r_h(s,a) = \theta_h^* \cdot \phi(s,a), \quad P_h(\cdot|s,a) = \mu_h^* \cdot \phi(s,a), \quad \forall h,$$

where ϕ is a known state-action feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, and $\mu_h^* \in \mathbb{R}^{|\mathcal{S}| \times d}$. Here ϕ, θ_h^* are known to the learner,while μ^* is unknown.

The definition implies a low-rank assumption in large-MDP case, since it says the transition matrix $P_h \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ has rank ar most d.

Example. *Tabular MDPs are instances of linear MDPs by choosing $d = |\mathcal{S}||\mathcal{A}|$ and the feature map* $\phi(s,a) = \delta_{s,a}(\cdot) = \begin{cases} 1, & \text{if input is } (s,a) \\ 0, & \text{otherwise} \end{cases}$.

Claim 2. *Linear MDPs imply that the Q-function for any policy is linear.*

Proof. For the optimal Q^* function:

$$\begin{aligned} Q_h^*(s,a) &= r(s,a) + P_h(\cdot|s,a)\cdot V_{h+1}^* \\ &= \theta_h^* \cdot \phi(s,a) + (\mu_h^* \phi(s,a))^T V_{h+1}^* \\ &= \left(\theta_h^* + (\mu_h^* V_{h+1}^*)\right)^T \phi(s,a) \\ &= (w_h)^T \phi(s,a), \end{aligned}$$

where $w_h := \theta_h^* + (\mu_h^* V_{h+1}^*$. Then it is clear that $Q_h^*(s,a)$ is a linear function with respect to $\phi(s,a)$ and the optimal policy is simply $\pi_h^*(s) = \arg\max_a (w_h)^T \phi(s,a)$ for all $h = 0, \ldots, H-1$. □

10.3 UCB-VI for linear MDPs

10.3.1 Algorithm

Since $\mu_h^* \phi(s_h^i, a_h^i) = P_h(\cdot|s_h^i, a_h^i)$, and $\delta(s_{h+1}^i)$ is an unbiased estimator of $P_h(\cdot|s_h^i, a_h^i)$ conditioned on s_h^i, a_h^i. It is reasonable to learn μ^* by regression from $\phi(s_h^i, a_h^i)$ to $\delta(s_{h+1}^i)$. Thus it leads to the ridge linear regression:

$$||\hat{\mu}_h^n||_F = \arg\min_{\mu \in \mathbb{R}^{|\mathcal{S}|\times d}} \sum_{i=0}^{n-1} ||\mu\phi(s_h^i, a_h^i) - \delta(s_{h+1}^i)||_2^2 + \lambda||\mu||_F^2,$$

which has the closed form solution given in (10.1).

Algorithm: in every round,

1. Run Ridge regression for estimating the model:
$$\hat{\mu}_h^n = \sum_{i=0}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i, a_h^i)^T (\Gamma_h^n)^{-1}. \tag{10.1}$$
2. Construct the exploration bonuses:
$$b_h^n(s,a) = \beta\sqrt{\phi(s,a)^T(\Lambda_h^n)^{-1}\phi(s,a)}.$$
3. Run optimistic value iterations, and update greedy policy:
$$\begin{aligned} &\hat{V}_H^n(s) = 0, \forall s, \\ &\hat{Q}_h^n(s,a) = r_h(s,a) + b_h^n(s,a) + \hat{P}(\cdot|s,a)\cdot \hat{V}_{h+1}^n \\ &\qquad = \theta_h^* \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^T(\Lambda_h^n)^{-1}\phi(s,a)} + \phi(s,a)^T(\hat{\mu}_h^n)^T \hat{V}_{h+1}^n \\ &\qquad = \beta\sqrt{\phi(s,a)^T(\Lambda_h^n)^{-1}\phi(s,a)} + \left(\theta_h^* + (\hat{\mu}_h^n)^T \hat{V}_{h+1}^n\right)^T \phi(s,a), \\ &\hat{V}_h^n(s) = \min\{\max_a \hat{Q}_h^n(s,a), H\}, \quad \pi_h^n(s) = \arg\max_a \hat{Q}_h^n(s,a). \end{aligned}$$

10.3.2 Regret Analysis

Theorem 3. *(Regret Bound). Choose*

$$\lambda = 1, \quad \beta = Hd\left(\sqrt{\log(H/\delta)} + \sqrt{\log(W+H)} + \sqrt{\log B} + \sqrt{\log d} + \sqrt{\log N}\right) = \tilde{O}(Hd),$$

UCB-VI achieves the following regret bound:

$$\mathbb{R}\left[NV^* - \sum_{i=0}^{N} V^{\pi_n}\right] \leq \tilde{O}(H^2\sqrt{d^3N}),$$

where N is the number of episodes.

Proof. **Sketch of the regret analysis**: We will use several results/lemmas, which is presented in Section 10.3.3, to prove the regret bound. We first use optimism to upper bound per-episode regret and use simulation lemma to decompose the per-episode regret. Then by uniform concentration and information gain bound, we can attain the regret bound.

first, we upper bound the per-episode regret (for episode n) as follows:

$$\begin{aligned}
V^* - V^{\pi_n} &= V_0^*(s_0) - V_0^{\pi_n}(s_0) \\
\text{(by optimism)} &\leq \hat{V}_0^{\pi_n}(s_0) - V_0^{\pi_n}(s_0) \\
\text{(by simulation lemma)} &\leq \sum_{h=0}^{H-1} \mathbb{E}^{\pi_n}\left[b_h^n(S_h, A_h) + \left(\hat{P}_h^n(\cdot|S_h, A_h) - P_h(\cdot|S_h, A_h)\right) \cdot \hat{V}_{h+1}^n\right] \\
\text{(by uniform concentration lemma)} &\leq \sum_{h=0}^{H-1} \mathbb{E}^{\pi_n}\left[2b_h^n(S_h, A_h)\right] \quad \text{if } \hat{V}_{h+1}^n \in \mathcal{F}.
\end{aligned}$$

Then with high probability:

$$V^* - V^{\pi_n} \leq \sum_{h=1}^{H-1} \mathbb{E}^{\pi_n}\left[2b_h^n(S_h, A_h)|\text{hist}_n\right],$$

and the total regret:

$$\begin{aligned}
\text{Regret} &= \mathbb{E}\left[\sum_{n=0}^{N-1}(V^* - V^{\pi_n})\right] = \mathbb{E}\left[\sum_{n=0}^{N-1}(V^* - V^{\pi_n})\mathbb{1}(\text{Not Fail})\right] + \mathbb{E}\left[\sum_{n=0}^{N-1}(V^* - V^{\pi_n})\mathbb{1}(\text{Fail})\right] \\
&\leq \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} 2b_h^n(S_h^n, A_h^n)\mathbb{1}(\text{Not Fail})\right] + \delta NH \\
&\leq \mathbb{E}\left[\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} 2b_h^n(S_h^n, A_h^n)\right] + \delta NH \qquad (10.2) \\
\text{(see below)} &\leq \tilde{O}(H^2\sqrt{d^3N}). \qquad (10.3)
\end{aligned}$$

The last step is because, by information gain bound and note that $\beta = \tilde{O}(Hd)$,

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{h=0}^{H-1} b_h^n(S_h^n, A_h^n) &= \beta \sum_{n=0}^{N-1}\sum_{h=0}^{H-1} \sqrt{\phi(S_h^n, A_h^n)(\Lambda_h^n)^{-1}\phi(S_h^n, A_h^n)} \\ &\leq \beta \sum_{h=0}^{H-1} \sqrt{N \sum_{n=0}^{N-1} \phi(S_h^n, A_h^n)(\Lambda_h^n)^{-1}\phi(S_h^n, A_h^n)} \\ &= \tilde{O}(dH^2\sqrt{Nd\log N}) = \tilde{O}(H^2\sqrt{d^3N}).\end{aligned}$$

□

10.3.3 Lemmas for Regret Analysis

It remains to prove for UCB-VI:

1. Uniform convergence bound;
2. Optimism;
3. Information gain bound as Lemma 7.12 in [AJKS].

Lemma 4. *(Uniform Concentration/ Uniform Convergence). With probability* $1-\delta$, $\forall\ s, a, h, n$

$$\sup_{f\in\mathcal{F}} \left(\hat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)\right)\cdot f \leq \beta||\phi(s,a)||_{(\Lambda_h^n)^{-1}} =: b_h^n(s,a)$$

Proof. Using Lemma 7.3 in [AJKS]:

$$\hat{\mu}_h^n - \mu_h^* = -\lambda\mu_h^*(\Lambda_h^n)^{-1} + \sum_{i=1}^{n-1}\epsilon_h^i\phi(s_h^i, a_h^i)^T(\Lambda_h^n)^{-1}.$$

Then

$$\begin{aligned}\left((\hat{\mu}_h^n - \mu_h^*)\cdot\phi(s,a)\right)^T V &= \phi(s,a)^T(\hat{\mu}_h^n - \mu_h^*)^T V \\ &= -\lambda\phi(s,a)^T(\Lambda_h^n)^{-1}(\mu_h^*)^T V + \left(\sum_{i=1}^{n-1}\epsilon_h^i\phi(s_h^i, a_h^i)^T(\Lambda_h^n)^{-1}\right)^T \cdot V.\end{aligned}$$

The first term is:

$$\begin{aligned}\text{Bias} &= -\lambda\phi(s,a)^T(\Lambda_h^n)^{-1}(\mu_h^*)^T V \\ &= -\lambda\phi(s,a)^T(\Lambda_h^n)^{-1/2}(\Lambda_h^n)^{-1/2}(\mu_h^*)^T V \\ &\leq \lambda||\phi(s,a)||_{(\Lambda_h^n)^{-1}}||(\mu_n^*)^T V||_{(\Lambda_h^n)^{-1}} \\ &\leq \sqrt{d}H||\phi(s,a)||_{(\Lambda_h^n)^{-1}}.\end{aligned}$$

The second term:

$$\begin{aligned}
\text{Variance} &= \phi(s,a)^T(\Lambda_h^n)^{-1}\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(\epsilon_h^i)^T\cdot V \\
&\le ||\phi(s,a)||_{(\Lambda_h^n)^{-1}}\cdot||\sum_{i=1}^{n-1}\phi(s_h^i,a_h^i)(\epsilon_h^i)^T V||_{(\Lambda_h^n)^{-1}} \\
\text{(by "self-normalized bound")} &\le ||\phi(s,a)||_{(\Lambda_h^n)^{-1}}\cdot 3H\sqrt{\log\left(\frac{H\det(\Lambda_h^n)^{1/2}\det(\lambda I)^{-1/2}}{\delta}\right)} \\
\text{(by information gain)} &\le ||\phi(s,a)||_{(\Lambda_h^n)^{-1}}\cdot 3H\sqrt{d\log N}.
\end{aligned}$$

So that

$$\left((\hat{\mu}_h^n-\mu_h^*)\cdot\phi(s,a)\right)^T V \le ||\phi(s,a)||_{(\Lambda_h^n)^{-1}}\cdot(\sqrt{d}H+3H\sqrt{d\log N}).$$

Recall that $\left(\hat{P}_h^n(\cdot|s,a)-P_h(\cdot|s,a)\right)\cdot f=\phi(s,a)^T(\hat{\mu}_h^n-\mu_h^*)^T f$ and use covering number theorem, for any $f\in\mathcal{F}$, there exists a $V\in\mathcal{N}_\epsilon$, such that $||f-V||_\epsilon\le\epsilon$. It remains one more step to finish the proof and it is shown in the next lecture notes. □

Lemma 5. *("Optimism" from Optimism Value Iteration)*

$$\hat{V}_0^{\pi_n}(s_0)\ge V_0^*(s_0)$$

.

Proof. (Use Induction.) If $\hat{V}_{h+1}^n(s)\ge\hat{V}_{h+1}^*(s)$ for all s, then

$$\begin{aligned}
\hat{Q}_h^n(s,a)-Q_h^*(s,a) &= r(s,a)+\beta\sqrt{\phi(s,a)^T(\Lambda_h^n)^{-1}\phi(s,a)}+\phi(s,a)^T(\hat{\mu}_h^n)^T\hat{V}_{h+1}^n \\
&\quad -r(s,a)-\phi(s,a)^T(\mu_h^*)^T V_{h+1}^* \\
\text{apply inductive hypothesis} &\ge \beta\sqrt{\phi(s,a)^T(\Lambda_h^n)^{-1}\phi(s,a)}+\phi(s,a)^T(\hat{\mu}_h^n-\mu_n^*)\hat{V}_{h+1}^n \\
&\ge 0,
\end{aligned}$$

if we choose β such that with high probability

$$\beta\sqrt{\phi(s,a)^T(\Lambda_h^n)^{-1}\phi(s,a)}\ge\left|\phi(s,a)^T(\hat{\mu}_h^n-\mu_n^*)\hat{V}_{h+1}^n\right|,$$

according to Lemma 4. □

Lemma 6. *(Information Gain Bound)* $\forall S_h^n, A_h^n$ *sequence*

$$\sum_{n=0}^{N-1}\phi(S_h^n,A_h^n)\Lambda_h^{n-1}\phi(S_h^n,A_h^n)=\tilde{O}(d\log N)$$

Challenge

We cannot use union bound because we have an infinite number of value functions. We will use the covering number idea to solve for it and prove Lemma 4.

References

[AJKS] Agarwal, Jiang, Kakade and Sun, "Reinforcement Learning: Theory and Algorithms," *unpublished working draft*, Dec 2020.