CS292F StatRL Lecture 10
Exploration in Linear MDPs

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Recap: Lecture 9

- Exploration in Tabular MDPs
  - Problem setup: Episodic Finite-H MDP with non-stationary transitions.
  - Regret definition

- UCB-VI (Azar et al., 2017)
  - A model-based approach; requires estimating P.

- Ideas in the proof
  - Concentration
  - Optimism via exploration bonus in value iterations
  - A few other tricks

This lecture: Exploration in Reinforcement Learning

- Why is it challenging?
  - The reward depends on both s, a
  - Unlike the generative model setting, we cannot just choose any s to explore.
  - The data needs to be actively collected

- We will study
  - Tabular MDP
  - **Linear MDPs**
  - Both in the finite horizon episodic setting.

Recap: episodic finite horizon MDPs with non-stationary transitions

- Problem setup / notations
  - MDP: $$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \{r_h\}_h, \{P_h\}_h, H, s_0\}$$
  - Policy depends on time step $$\pi = \{\pi_0, \ldots, \pi_{H-1}\}$$

    $\pi_h : \mathcal{S} \to \mathcal{A}$

    $\Delta(\mathcal{A})$
  - Performance measure $$\text{Regret} := \mathbb{E}\left[\sum_{k=0}^{K-1}\left(V^\star - V^{\pi^k}\right)\right]$$

    $V^* := V_0^{\pi^*}(s_0)$

    $V^{\pi^k} := V_0^{\pi^k}(s_0)$

Recap: Linear function approximation in TD-learning

- Why do we need it?
  - Recall the PACMAN example.

Let's say we discover through experience that this state is bad:

In naïve q-learning, we know nothing about this state:

Or even this one!

  - Describe the state by its features, and value functions linear in features
    - $V_w(s) = w_1 f_1(s) + w_2 f_2(s) + \ldots + w_n f_n(s)$
    - $Q_w(s,a) = w_1 f_1(s,a) + w_2 f_2(s,a) + \ldots + w_n f_n(s,a)$

How do we formally analyze this?

$\|f - Q^*\|_\infty \leq \epsilon$

$\Downarrow$

$\pi_f = \arg\max_a f(s,a)$

$\boxed{V^* - V^{\pi_f} \leq \frac{\epsilon}{1-\gamma}}$

- What are the assumptions to make?
  - **Q*(s,a) approximately linear?**
  - **Q$^\pi$(s,a) is approximately linear for all π?**
  - Q*(s,a) is exactly linear? $\phi(s,a)$

    $Q^*(s,a) = (\theta^*)^T \phi(s,a)$
  - Q$^\pi$(s,a) is exactly linear for all π?

Exponential sample complexity / regret lower bounds for the approximate case…

(Du, Kakade, Wang, Yang, 2019) Is a good representation sufficient for sample efficient reinforcement learning? $\Omega(e^H)$

Linear MDPs

\# of parameters

$H(d + |S| \cdot d)$

- Exists feature map $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$
  - Such that:

$\in \mathbb{R}^{|S| \times |S| \cdot |A|}$

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

$(\theta_h^\star)^T \; \phi(s,a)$

d $\quad |S|(|A|)$

$|S| \quad \mu_h^\star$

Dictionary

$= P_h(\cdot|\cdot,\cdot)$

- Implies a low-rank assumption in large-MDP case

$\phi(s,a) \quad \phi(s_1,a_1)$

(Jin et al., 2020) Provably efficient reinforcement learning with linear function approximation

Tabular MDPs are instances of linear MDPs

$\delta_{s,a}(\cdot) = \begin{cases} 1 & \text{if input is } s,a \\ 0 & \text{otherwise} \end{cases}$

$|S||A|$, $|S|$, $P_h(s'|s,a)$

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

- Choose d $= |S||A|$
- Feature map to be:

$\phi(s,a) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in R^{|S||A|}$ $\leftarrow (s,a)$

Linear MDPs imply that the Q function for any policy is linear

- Optimal Q* function:

$$Q_h^*(s,a) = r(s,a) + P_h(\cdot|s,a) \cdot V_{h+1}^*(\cdot)$$

$$= \theta_h^* \cdot \phi(s,a) + (\mu_h^* \phi(s,a))^T V_{h+1}^*(\cdot)$$

$$= (\theta^* + \mu_h^{*T} V_{h+1}^*(\cdot))^T \phi(s,a) = w^{*T} \phi(s,a)$$

with any $f: S \to \mathbb{R}$

$f: V_{h+1}^{\pi}(\cdot)$

- Claim 7.2 (AJKS) For any function of the state, the Bellman backup is a linear in the feature.

Recap: LinUCB in linear bandits

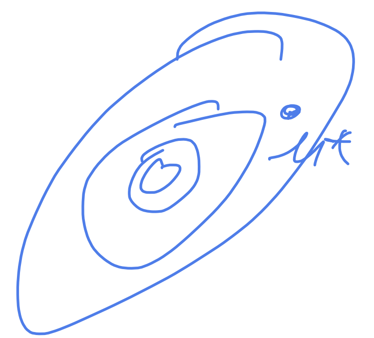
- In every round:
  1. Use ridge regression to estimate the model parameters

$$\hat{\mu}_t = \arg\min_{\theta} \sum_{i=1}^{t-1} (x_i^\top \theta - r_i)^2 + \lambda \|\theta\|^2$$

  2. Construct an ellipsoidal confidence set of the parameters

$$\mathrm{Ball}_t = \left\{ \mu \;\middle|\; (\mu - \hat{\mu}_t)^\top \Sigma_t (\mu - \hat{\mu}_t) \le \beta_t \right\}$$

where $\Sigma_t = \sum_{i=1}^{t-1} x_i x_i^\top + \lambda I_d$



  3. Choose actions that maximize UCB

$$x_t = \arg\max_{x \in D} \max_{\mu \in \mathrm{Ball}_t} \langle x, \mu \rangle$$

UCB-VI for Linear MDPs

[θ^* is known]

parameter $\mu^* \in R^{|S|\times d}$

- In every round:
  1. Run Ridge regression for estimating the model

Feature

$\|\cdot\|_F = \|\mathrm{vec}(\cdot)\|_2$

label

$\widehat{\mu}_h^n \in R^{|S|\times d}$

$$\widehat{\mu}_h^n = \mathrm{argmin}_{\mu\in\mathbb{R}^{|S|\times d}} \sum_{i=0}^{n-1} \left\|\mu\phi(s_h^i,a_h^i) - \delta(s_{h+1}^i)\right\|_2^2 + \lambda\|\mu\|_F^2.$$

$= P(\cdot \mid s_h^i, a_h^i) \in R^{|S|}$

$\delta_{s_{h+1}^i} \in R^{|S|}$

$$\widehat{\mu}_h^n = \sum_{i=0}^{n-1} \delta(s_{h+1}^i)\phi(s_h^i,a_h^i)^\top(\Lambda_h^n)^{-1}$$

$\Lambda_h^n = \sum_{i=1}^{n-1} \phi(s_h^i,a_h^i)\phi(s_h^i,a_h^i)^T + \lambda I$

$E[\delta_{s_h^i}] = P(\cdot \mid s_h^i, a_h^i)$

  2. Construct the exploration bonuses

$$b_h^n(s,a) = \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)},$$

width of $\phi(s,a)$

  3. Run optimistic value iterations, and update greedy policy

Optimistic value iterations

$r_h(s,a)$ $\quad b_h^n(s,a)$ $\quad \widehat{P}_h(\cdot|s,a)\cdot \widehat{V}^n_{h+1}$

$$
\begin{aligned}
&\widehat{V}_H^n(s) = 0, \forall s,\\
&\widehat{Q}_h^n(s,a) = \theta^\star_h \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1}\phi(s,a)} + \phi(s,a)^\top (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n\\
&\qquad = \beta\sqrt{\phi(s,a)^\top (\Lambda_h^n)^{-1}\phi(s,a)} + (\theta^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n)^\top \phi(s,a),\\
&\widehat{V}_h^n(s) = \min\{\max_a \widehat{Q}_h^n(s,a),\ H\}, \quad \pi_h^n(s) = \mathrm{argmax}_a\, \widehat{Q}_h^n(s,a).
\end{aligned}
$$

Regret bound

- Choose $\beta = Hd\left(\sqrt{\ln\frac{H}{\delta}} + \sqrt{\ln(W+H)} + \sqrt{\ln B} + \sqrt{\ln d} + \sqrt{\ln N}\right) \asymp \tilde{O}(1+d)$

  $\lambda = 1$

- Regret $\underline{\tilde{O}\left(H^2\sqrt{d^3 N}\right)}$ N is the # of episodes

Sketch of the regret analysis

Per-episode Regret:

$$V^* - V^{\pi_n} = V_0^*(s_0) - V_0^{\pi_n}(s_0) \overset{\text{optimism}}{\le} \hat{V}_0^{\pi_n}(s_0) - V_0^{\pi_n}(s_0)$$

Simulation Lemma $\rightarrow$ $$\le \sum_{h=0}^{H-1} \mathbb{E}^{\pi_n}\Big[b_h^n(s_h,A_h) + \big(\hat{P}_h^n(\cdot|s_h,A_h) - P_h(\cdot|s_h,A_h)\big)\cdot \hat{V}_{h+1}^n \Big]$$

$$\le \sum_{h=0}^{H-1} \mathbb{E}^{\pi_n}\big[b_h^n(s_h,A_h) \big]$$

Lemma (Uniform Concentration) w.p $1-\delta$ $\forall s,a,h,n$,

$$\sup_{f\in F} \big(\hat{P}_h^n(\cdot|s,a) - P_h(\cdot|s,a)\big)\cdot f \le \beta\|\phi(s,a)\|_{\Lambda_h^{n-1}} =: b_h^n(s,a) \quad \text{if } \hat{V}_{h+1}^n \in F$$

w.h.p $$V^* - V^{\pi_n} \le \mathbb{E}^{\pi_n}\Big[\sum_{h=0}^{H-1} 2 b_h^n(S_h^n, A_h^n) \,\Big|\, hist_n\Big]$$

$$\text{Regret} = \mathbb{E}\Big[\sum_{i=0}^{N-1} V^* - V^{\pi_i}\Big] = \mathbb{E}\Big[\sum_{i=0}^{N-1}(V^*-V^{\pi_i})\mathbb{1}(\text{Not fail})\Big] + \mathbb{E}\Big[\sum_{i=0}^{N-1}(V^*-V^{\pi_i})\mathbb{1}(\text{fail})\Big]$$

$$\le \mathbb{E}\Big[\sum_{i=1}^{N-1}\sum_{h=0}^{H} 2 b_h^n(S_h^i,A_h^i)\cancel{\mathbb{1}(\text{Not fail})}\Big] + \underbrace{\delta\cdot N\cdot H}_{\text{worst possible failure mode}}$$

Sketch of the regret analysis

tabular case

$$\sqrt{\frac{1}{N_h^n(s,a)}}$$

$$\sum_{i=1}^{N_{s,a}} \frac{1}{\sqrt{i}}$$

$$\sum_{n=0}^{N-1} \sum_{h=0}^{H-1} b_h^n(s_h^n, a_h^n) = \beta \sum_{n=0}^{N-1} \sum_{h=0}^{H-1} \sqrt{\phi(s_h^n, a_h^n)^\top (\Lambda_h^n)^{-1} \phi(s_h^n, a_h^n)}$$

$$\leq \beta \sum_{h=0}^{H-1} \sqrt{N \sum_{n=0}^{N-1} \phi(s_h^n, a_h^n)^\top (\Lambda_h^n)^{-1} \phi(s_h^n, a_h^n)} \simeq dH^2 \sqrt{N d \log N}$$

Lemma (Information Gain bound)

$\forall\, s_h^n, a_h^n$ sequence $\displaystyle\sum_{n=0}^{N-1} \phi(s_h^n, a_h^n)^\top (\Lambda_h^n)^{-1} \phi(s_h^n, a_h^n) = \tilde{O}(d \log N)$

$$= \tilde{O}(H^2 \sqrt{d^3 N})$$

$$\beta \simeq \tilde{O}(Hd)$$

Sketch of the regret analysis

It remains to prove

$$\sup_f \left| (\hat{P} - P) \cdot f \right| \le \varepsilon$$

w.p. $1-\delta$

- 1. Uniform convergence bound

- 2. "Optimism"

  The same induction argument as in the UCB-VI for tabular MDP
  (Read Lemma 7.10 in AJKS)

- 3. "Information gain" bound

  The same argument as in the Linear Bandits case.
  (Read Lemma 7.12 in AJKS)

“Optimism” from Optimistic Value Iterations

$\boxed{\text{Claim: } \hat{V}_0^{\pi_n}(s_0) \geq V_0^*(s_0)}$

If $\hat{V}^n_{h+1}(s) \geq V^*_{h+1}(s)$,

$$\hat{Q}^n_h(s,a) - Q^*_h(s,a) = \cancel{r(s,a)} + \beta\sqrt{\phi(s,a)^T(\Lambda^n_h)^{-1}\phi(s,a)} + \phi(s,a)^T(\hat{\mu}^n_h)^T\hat{V}^n_{h+1} - \cancel{r(s,a)} - \phi(s,a)^T(\mu^*_h)^T V^*_{h+1}$$

" ≥ 0 "

applying inductive Hypothesis $\rightarrow$ $\geq \beta\sqrt{\phi(s,a)^T(\Lambda^n_h)^{-1}\phi(s,a)} + \phi(s,a)^T(\hat{\mu}^n_h - \mu^*_h)^T\hat{V}^n_{h+1}$

if we choose β s.t. w.h.p. $\underline{\beta\sqrt{\phi(s,a)^T(\Lambda^n_h)^{-1}\phi(s,a)}} \geq \underline{\left|\phi(s,a)^T(\hat{\mu}^n_h - \mu^*_n)^T\hat{V}^n_{h+1}\right|}$

$\Downarrow$

$$\sup_{f \in F} \phi(s,a)^T(\hat{\mu}^n_h - \mu^*_h)^T f \leq$$

$\boxed{\hat{V}^n_{h+1} \in F}$

when F is finite, $\underline{\text{Hoeffding's}}$ + $\underline{\text{union bound}}$ $\delta : \frac{\delta}{|F|}$

Let us start with pointwise convergence for a fixed V

- Recall: Hoeffding's inequality + Union bound

- Recall: Ridge regression

$\{(S_h^i, A_h^i, S_{h+1}^i) \text{ for } i = 1, 2, \cdots, n-1\}$

$$\hat{\mu}_h^n = \underset{\mu \in \mathbb{R}^{|S| \times d}}{\text{argmin}} \sum_{i=0}^{n-1} \| \mu \phi(S_h^i, A_h^i) - \underbrace{\delta_{S_{h+1}^i}(\cdot)}_{} \|_2^2 + \lambda \|\mu\|_F^2$$

$$\hat{\mu}_h^n = \sum_{i=1}^{n-1} \delta_{S_{h+1}^i}(\cdot) \cdot \phi(S_h^i, A_h^i)^T (\Lambda_h^n)^{-1}$$

$|S|$ d d

for a fixed $V \in \mathbb{R}^{|S|}$

$$(\hat{P}_h(\cdot|s,a) - P_h(\cdot|s,a)) \cdot V \in \mathbb{R}$$

$$\|$$

$$\phi(s,a)^T \left(\sum_{i=1}^{n-1} (\Lambda_h^n)^{-1} \phi(S_h^i, A_h^i) \langle \delta_{S_{h+1}^i}, V \rangle \right.$$

$$\langle \delta_{S_{h+1}^i} - \mu^*, V \rangle)$$

ε_h^i

$$\hat{P}_h^n(\cdot|s,a) = (\hat{\mu}_h^n \cdot \phi(s,a))$$

$$P_h(\cdot|s,a) = \mu^* \cdot \phi(s,a)$$

Error of ridge regression estimate

$$\epsilon_h^i = \delta_{s_{h+1}^i} - P(\cdot \mid s_h^i, a_h^i)$$

$$\mu_h^\star \cdot \phi(s_h^i, a_h^i)$$

- Lemma 7.3 AJKS

$$\widehat{\mu}_h^n - \mu_h^\star = \underbrace{-\lambda \mu_h^\star (\Lambda_h^n)^{-1}}_{\text{bias}} + \underbrace{\sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top (\Lambda_h^n)^{-1}}_{\text{variance}} . \quad \in \mathbb{R}^{|S| \times d}$$

- The quantity of interest is a inner product with this:

$$\left[(\hat{\mu}_h^n - \mu_h^n) \cdot \phi(s,a)\right]^\top V = \phi(s,a)^\top (\hat{\mu}_h^n - \mu_h^n)^\top \cdot V$$

$$\text{bias} = -\lambda\, \phi(s,a)^\top (\Lambda_h^n)^{-1} \mu_h^{\star\top} V$$

$$= -\lambda\, \phi(s,a)^\top (\Lambda_h^n)^{-\frac{1}{2}} (\Lambda_h^n)^{-\frac{1}{2}} \mu_h^{\star\top} V$$

$$\le \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \|\mu_h^{\star\top} V\|_{(\Lambda_h^n)^{-1}} \le \lambda \|\phi(s,a)\|_{(\Lambda_h^n)^{-1}} \sqrt{d}\, H$$

$$\text{Variance} = \bigcirc^\top V$$

Recap: Self-normalized Martingale concentration bound.

Lemma (Self-Normalized Bound for Vector-Valued Martingales)

(Abassi et. al '11) Suppose $\{\varepsilon_i\}_{i=1}^{\infty}$ *are mean zero random variables (can be generalized to martingales), and* ε_i *is bounded by* σ*. Let* $\{X_i\}_{i=1}^{\infty}$ *be a stochastic process. Define* $\Sigma_t = \Sigma_0 + \sum_{i=1}^{t} X_i X_i^\top$*. With probability at least* $1-\delta$*, we have for all* $t \geq 1$*:*

$$\left\| \sum_{i=1}^{t} X_i \varepsilon_i \right\|_{\Sigma_t^{-1}}^2 \leq \sigma^2 \log\left(\frac{\det(\Sigma_t)\det(\Sigma_0)^{-1}}{\delta^2} \right).$$

Apply the above concentration

- How?

$s_{h+1} \sim P(\cdot|s,a)$

$$\mathbb{E}\left[V^\top \epsilon_h^i | \mathcal{H}_h^i\right] = 0, \quad |V^\top \epsilon_h^i| \leq \|V\|_\infty \|\epsilon_h^i\|_1 \leq 2H, \forall h, i.$$

- This is a martingale difference sequence.

- Thus by the “Self-Normalized bound”:

"information gain"

$$\left\| \sum_{i=0}^{n-1} \phi(s_h^i, a_h^i)(V^\top \epsilon_h^i) \right\|_{(\Lambda_h^n)^{-1}} \leq 3H \sqrt{\ln \frac{H \det(\Lambda_h^n)^{1/2} \det(\lambda I)^{-1/2}}{\delta}}.$$

Challenge: we cannot use union bound because we have an infinite number of value functions

- A covering number argument.

- Covering number: the number of balls with radius ε that is needed to cover all points in a set.

$\tilde{F} \subset F, \ |\tilde{F}| \le N_\varepsilon$

$\forall f \in F.$

$\exists \tilde{f} \in \tilde{F} \ \ \boxed{d(f, \tilde{f}) \le \varepsilon}$

$$(\hat{P}-P)^T \hat{V} = \underbrace{(\hat{P}-P)^T(\hat{V}-\tilde{V})}_{\le \varepsilon} + \underbrace{(\hat{P}-P)^T(\tilde{V})}_{\text{pointwise + union bound}}$$

Family of value functions we consider

$$f_{w,\beta,\Lambda}(s) = \min\left\{\max_a \left(w^\top \phi(s,a) + \beta\sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)}\right),\ H\right\}, \forall s \in \mathcal{S}.$$

$$\mathcal{F} = \{f_{w,\beta,\Lambda} : \|w\|_2 \le L, \beta \in [0,B], \sigma_{\min}(\Lambda) \ge \lambda\}.$$

What is a finite set to cover this class such that for every f in this set, there is a function in the finite set, such that they are ε-close in sup-norm?

Covering number calculations

From covering number to a uniform convergence bound