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

$$\pi = \{\pi_0, \ldots, \pi_{H-1}\}$$

  - Performance measure

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

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?

- 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?
  - $Q^\pi(s,a)$ 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?

Linear MDPs

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

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

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

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

Tabular MDPs are instances of linear MDPs

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

- Choose d
- Feature map to be:

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

- Optimal Q* function:

- 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

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

  2. Construct an ellipsoidal confidence set of the parameters

$$\text{Ball}_t = \left\{ \mu \mid (\mu - \hat{\mu}_t)^T \Sigma_t (\mu - \hat{\mu}_t) \le \beta_t \right\}$$

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

  3. Choose actions that maximize UCB

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

UCB-VI for Linear MDPs

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

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

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

  2. Construct the exploration bonuses

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

  3. Run optimistic value iterations, and update greedy policy

Optimistic value iterations

$$\widehat{V}_H^n(s) = 0, \forall s,$$

$$\widehat{Q}_h^n(s,a) = \theta^\star \cdot \phi(s,a) + \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + \phi(s,a)^\top(\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n$$

$$= \beta\sqrt{\phi(s,a)^\top(\Lambda_h^n)^{-1}\phi(s,a)} + (\theta^\star + (\widehat{\mu}_h^n)^\top \widehat{V}_{h+1}^n)^\top \phi(s,a),$$

$$\widehat{V}_h^n(s) = \min\{\max_a \widehat{Q}_h^n(s,a),\ H\}, \quad \pi_h^n(s) = \mathrm{argmax}_a\, \widehat{Q}_h^n(s,a).$$

Regret bound

- Choose

$$\beta = Hd\left(\sqrt{\ln\frac{H}{\delta}} + \sqrt{\ln(W+H)} + \sqrt{\ln B} + \sqrt{\ln d} + \sqrt{\ln N}\right)$$

$$\lambda = 1$$

- Regret

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

Sketch of the regret analysis

Sketch of the regret analysis

Sketch of the regret analysis

It remains to prove

- 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

Let us start with pointwise convergence for a fixed V

- Recall: Hoeffding's inequality + Union bound

- Recall: Redge regression

Error of ridge regression estimate

- Lemma 7.3 AJKS

$$\widehat{\mu}_h^n - \mu_h^\star = -\lambda \mu_h^\star \left(\Lambda_h^n\right)^{-1} + \sum_{i=1}^{n-1} \epsilon_h^i \phi(s_h^i, a_h^i)^\top \left(\Lambda_h^n\right)^{-1}.$$

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

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?

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

- This is a martingale difference sequence.

- Thus by the "Self-Normalized bound":

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

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.

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