CS292F StatRL Lecture 9
Exploration in Tabular MDPs

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Logistic notes

- HW1 due today.

- HW2 is posted on the course website.
  - Q1: A simple coding question
  - Q2: An alternative rate-optimal algorithm for MAB.
  - Q3: Exploration in tabular RL

Recap: Lecture 8

- Linear Bandits
  - Problem setup
  - Regret definition

- Optimism in the face of uncertainty
  - LinUCB algorithm
  - Bounding sum of square regrets with information gain.
  - a self-normalized Martingale Concentration

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: Finite horizon MDPs

- Parameterization / Setup

$$M = (\mathcal{S}, \mathcal{A}, \{P\}_h, \{r\}_h, H, \mu)$$

- Additional notations
  - Q functions
  - V functions
  - Policies
- Observed trajectory data

Problem setup: online learning of Finite horizon MDPs

- Agent decides on a policy
- Collect a trajectory
- Agent updates the policy.
- Regret definition

Recap: The need for strategic exploration

UCB-VI: model-based learning by optimistic value Iterations

- Construct estimates of the transition kernels

- Design exploration bonuses

  - Idea: based on the uncertainty in the transition kernel estimates

- Update the policy by optimistic value iteration

How do we estimate the model parameters (P and r)?

- Simple plug-in estimator

- What happens if we observe no state-action pairs?

What does value iteration do in finite horizon MDPs?

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

$$\widehat{Q}_h^k(s,a) = \min\left\{ r_h(s,a) + b_h^k(s,a) + \widehat{P}_h^k(\cdot|s,a) \cdot \widehat{V}_{h+1}^k,\ H \right\},$$

$$\widehat{V}_h^k(s) = \max_a \widehat{Q}_h^k(s,a), \pi_h^k(s) = \mathrm{argmax}_a\, \widehat{Q}_h^k(s,a), \forall h, s, a.$$

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

How do we design exploration bonuses?

$$b_h^k(s,a) = H\sqrt{\frac{L}{N_h^k(s,a)}} \quad \text{where} \quad L := \ln\left(SAHK/\delta\right)$$

- Intuitively, this encourages exploring new state-action pairs.
- Idea: propagate errors from the estimated transitions over to the rewards.

The regret of UCB-VI

- Theorem (AJKS Thm 6.1):

$$\textit{Regret} := \mathbb{E}\left[\sum_{k=0}^{K-1}\left(V^\star - V^{\pi^k}\right)\right] \leq 2H^2 S\sqrt{AK\cdot\ln(SAH^2K^2)} = \widetilde{O}\left(H^2 S\sqrt{AK}\right)$$

- This is not optimal in H, S, but a simple analysis to start. We will talk about how to improve it towards the end.

Step 1: Concentration

Step 2: Optimism

Finite horizon simulation lemma (from HW1)

Regret in kth Episode

Total regret

Ideas for improving the dependence on S and H

Final notes about exploration in Tabular MDPs

- Optimal rates:
  - Non-stationary transitions
  - Stationary transitions

- State of the art:
  - Stationary case: MVP O(sqrt{H^2SAK} + H^2S^2A)
    - Zhang, Ji and Du (2020) https://arxiv.org/pdf/2009.13503.pdf
    - Modified the episode reward bound from [0,1] to [0,H] to be consistent with this lecture
  - Nonstationary case: O(sqrt(H^3SAK) + H^4S^2A)
  - Q-learning: Jin et al., Bai et al., optimal rates in Zhang et al. (2020)

- Open problem:
  - Is it possible to get rid of the S dependence in the low-order terms.