Hello!

# 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

0-based index. $P_0, P_1, \dots, P_{h-1}$

- Parameterization / Setup

horizon $H < +\infty$

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

$S_0 \sim \mu$

Nonstationary Transition: $P_h (: S \times A \to \Delta S)$ differs for each $h$

$\mu = S_0$

$r_0, \dots, r_{H-1}$ $\quad r_h : S \times A \to \mathbb{R}$ $\quad \mathbb{E}[R_h \mid S_h = s, A_h = a] =: r_h(s,a)$

- Additional notations
  - Q functions $Q_h^\pi, Q_h^*$ $\quad h = 0, 1, \dots, H-1$.

$\boxed{Q_H^{\pi(\cdot,\cdot)} = Q_H^{*(\cdot,\cdot)} = 0 \quad \forall s, a \text{ chart}}$ for convenience

  - V functions $V_h^\pi, V_h^*$
  - Policies $\pi = \{\pi_0, \dots, \pi_{H-1}\}$

$$\boxed{Q_h^\pi(s,a) = r_h(s,a) + \mathbb{E}^\pi_{(s,a) \sim P_h(s',a' \mid s,a)} Q_{h+1}^\pi(s',a') \quad \text{all } h = 0, 1, \dots, H-1}$$

- Observed trajectory data

Execute policy $\pi$:

$\text{trajectory}_k = \{(S_0^k, A_0^k), (S_1^k, A_1^k), \dots, (S_{H-1}^k, A_{H-1}^k)\}$

# Problem setup: online learning of Finite horizon MDPs

$\pi_k = \{\pi_{k,h} \text{ for } h = 0, \dots, H-1\}$

$\pi_{k,h} : S \to \Delta(A)$

- Agent decides on a policy

  $\pi_0$ at 0th episode, ..., $\pi_k$ at $k$th episode

- Collect a trajectory

  $(S_0^k, A_0^k) \cdots (S_{H-1}^k, A_{H-1}^k), \cancel{(S_H, A_H)}$

- Agent updates the policy.

  $\pi_{k+1} \in \text{function}(\text{Hist}_k)$

- Regret definition

  $\pi^*$

  $\pi_h^*(s) = \arg\max_{a \in A} Q_h^*(s, a)$ for each $h$

$$\text{Regret} = \sum_{k=0}^{K-1} \text{Regret}_k = K \cdot V^*(s_0) - \sum_{k=0}^{K-1} V^{\pi_k}(s_0)$$

$$E[\text{Regret}] = K V^*(s_0) - E\sum_{k=0}^{K-1} V^{\pi_k}(s_0)$$

# Recap: The need for strategic exploration

$r(s_i, a) = 0 \quad \forall i < H-1, \ \forall a$

$r(s_{H-1}, a) = 1 \quad \forall a$


Randomized exploration $\quad P^\pi(s_{H-1}) = O\left(\left(\frac{1}{2}\right)^H\right)$

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

- Construct estimates of the transition kernels

  $\hat{P}_h^k$

  model-based approach $\hat{M}^k$

- Design exploration bonuses

  At Episode k: $b_h^k(s,a)$

  $\hat{Q}_h(s,a) + b_h^k(s,a) \geq Q_h^*(s,a)$

  - 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

At $k$, $h$,

$$\hat{P}_h^k(s'|s,a) = \frac{N_h^k(s,a,s')}{N_h^k(s,a)}$$

$N_h^k(s,a,s')$ = # of times these triplets appeared from step $h$ to step $h+1$

$$= \sum_{i=0}^{k-1} \mathbb{1}(S_h^i = s, A_h^i = a, S_{h+1}^i = s')$$

$$N_h^k(s,a) = \sum_{i=1}^{k-1} \mathbb{1}(S_h^i = s, A_h^i = a)$$

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

$$\frac{0}{0} =: 0$$

# What does value iteration do in finite horizon MDPs?

$0 \le r_h(s,a) \le 1$

$\Downarrow$

$\widehat{Q}_h(s,a) \le H$

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

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

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

for $h = H, \dots, 0$

$\widehat{Q}_1^k, \dots, \widehat{Q}_{H-1}^k, \widehat{V}_1^k, \dots, \widehat{V}_{H-1}^k$

$\Downarrow$

$\pi_h^{k+1}(s) = \arg\max_a \widehat{Q}_h^k(s,a)$

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

# How do we design exploration bonuses?

Hoeffding style bonus

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

max possible V function

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

# The regret of UCB-VI

(Azar et al., 2017)

(Jaksch et al., 2009)

- Theorem (AJKS Thm 6.1):

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

$\widetilde{O}(\sqrt{H^4S^2AK})$

No regret learning Alg

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

lower bound: $\Omega\left(\sqrt{H^3SAK}\right)$

nonstationary transition MDP

H-horizon

# Step 1: Concentration

Condition on $N_h^k(s,a)$, $S_{h+1}^{(i)}$ for $i = 0, 1, \dots, k-1$ are i.i.d

Lemma 1: for all $h, k, s, a$, w.p. $1-\delta$

$$\left\| \hat{P}_h^k(\cdot|s,a) - P_h^k(\cdot|s,a) \right\|_1 \le O\left(\sqrt{\frac{S \log \frac{SAHK}{\delta}}{N_h^k(s,a)}}\right)$$

By McDiarmid Inequality

Lemma 2 w.p $\ge 1-\delta$, for $\forall s, a, h, k$

$$\left\| \hat{P}_h^k(\cdot|s,a) \cdot V_{h+1}^* - P_h^k(\cdot|s,a) \cdot V_{h+1}^* \right\| \le H \sqrt{\frac{L}{N_h^k(s,a)}}$$

$$\frac{1}{N_h^k(s,a)} \sum_{i=1}^{k-1} \underbrace{V_{h+1}^*(S_{h+1}^i) \mathbb{1}(S_h^i = s, A_h^i = a)}_{\text{i.i.d R.V.}}$$

$\mathbb{P}(\text{Fail}) \le 2\delta$

# Step 2: Optimism

under "not fail"

$\boxed{\text{Lemma 3: } \hat{V}_h^k \geq V_h^* \text{ for all } h = 0, 1, \ldots, H-1, H}$

Proof: Base: $\hat{V}_H^k = V_H^* = 0$

Assume for $h$, $\hat{V}_h^k \geq V_h^*$, we will prove that $\hat{V}_{h-1}^k \geq V_{h-1}^*$

$\hat{V}_{h-1}^k = \max_a \hat{Q}_{h-1}^k(\cdot, a)$, and $\hat{Q}_{h-1}^k(s,a) = \min\{H, \; r_{h-1}(s,a) + \hat{P}_{h-1}(\cdot \mid s,a) \cdot \hat{V}_h^k(\cdot) + b_{h-1}^k(s,a)\}$

Case ① when $H$ is the smaller: $\hat{Q}_{h-1}(s,a) = H \geq Q_{h-1}^*(s,a)$

Case ② when $H$ is not selected

$Q_{h-1}^*(s,a) = r_{h-1}(s,a) + P_{h-1}(\cdot \mid s,a) \cdot V_h^*(\cdot)$

show $\hat{V}_{h-1}^k(s,a) \geq V_{h-1}^*(s,a)$

$\hat{Q}_{h-1}^k(s,a) - Q_{h-1}^*(s,a)$

$= b_{h-1}^k(s,a) + \hat{P}_{h-1}^k(\cdot \mid s,a)^\top \hat{V}_h^k - P_{h-1}(\cdot \mid s,a) V_h^*$

$\geq b_{h-1}^k(s,a) + \left(\hat{P}_{h-1}^k(\cdot \mid s,a) - P_{h-1}(\cdot \mid s,a)\right) \cdot V_h^* \geq b_{h-1}^k(s,a) - H\sqrt{\frac{L}{N_{h-1}^k(s,a)}} \geq 0$

$\hat{V}_h^k(s) = \max_a \hat{Q}_h^k(s,a)$

$\geq \hat{Q}_h^k(s, a^*) \geq Q_h^*(s, a^*)$

$= V_h^*(s)$ $\square$

# Finite horizon simulation lemma (from HW1)

Q5.

$$\left\| \left| \hat{V}^{\pi}(s) - V^{\pi}(s) \right| \le \text{small} \right\|$$

$$\hat{V}_h^{\pi} = \hat{r}_h + \hat{P}_h^{\pi} \cdot \hat{V}_{h+1}^{\pi}$$

$$V_h^{\pi} = r_h + P_h^{\pi} V_h^{\pi}$$

Take difference

$$\underbrace{\hat{V}_h^{\pi} - V_h^{\pi}}_{\text{recursively apply the difference}} = \hat{r}_h - r_h + \overbrace{P_h^{\pi}}^{\text{add}} \underline{(\hat{V}_{h+1}^{\pi} - V_{h+1}^{\pi})} + \overbrace{(\hat{P}_h^{\pi} - P_h^{\pi})}^{\text{Subtract}} \cdot \hat{V}_{h+1}^{\pi} \quad (*)$$

$$\hat{V}_0^{\pi} - V_0^{\pi} = \hat{r}_0 - r_0 + P_0^{\pi} (\hat{V}_1^{\pi} - V_1^{\pi}) + (\hat{P}_0^{\pi} - P_0^{\pi}) \cdot \hat{V}_1^{\pi}$$

$$P_0^{\pi} \left[ \hat{r}_1 - r_1 + (\hat{P}_1^{\pi} - P_1^{\pi}) \cdot \hat{V}_1^{\pi} + P_1^{\pi} (\hat{V}_2^{\pi} - V_2^{\pi}) \right]$$

$$\hat{V}_0^{\pi} - V_0^{\pi} = \sum_{h=0}^{H-1} \underset{\text{True MDP}}{\mathbb{E}^{\pi}} \left[ \hat{r}_h^{\pi}(s_h) - r^{\pi}(s_h) + \left( \hat{P}_h^{\pi}(\cdot \mid s_h) - P_h^{\pi}(\cdot \mid s_h) \right) \cdot \hat{V}_{h+1}(\cdot) \right]$$

$$= \sum_{h=0}^{H-1} \mathbb{E}^{\pi} \left[ \sum_a \pi_h(a \mid s_h) \left( \hat{r}_h(s_h, a_h) - r_h(s_h, a_h) + \hat{P}_h(\cdot \mid s_h, a_h) - P(\cdot \mid s_h, a_h) \right) \cdot \hat{V}_{h+1}(\cdot) \right]$$

# Regret in kth Episode

Simulation lemma and also take $\min(H, \cdot)$ make it smaller

$$\text{Regret}_k = V_0^*(s_0) - V_0^{\pi_k}(s_0) \leq \hat{V}_0^{\pi_k}(s_0) - V_0^{\pi_k}(s_0) \leq \sum_{h=0}^{H-1} \mathbb{E}^{\pi_k}\Big[\hat{r}_h(s_h, A_h) - r(s_h, A_h) + (\hat{P}_h(\cdot|s_h, A_h) - P_h(\cdot|s_h, A_h)) \cdot \hat{V}_{h+1}^{\pi_k}\Big]$$

By optimism

Condition on $\text{Hist}_k$

$$= \sum_{h=0}^{H-1} \mathbb{E}^{\pi_k}\Big[b_h^k(s_h, A_h) + \underbrace{(\hat{P}_h(\cdot|s_h, A_h) - P_h(\cdot|s_h, A_h)) \cdot \hat{V}_{h+1}^{\pi_k}}\Big]$$

$$\leq \|\hat{P}_h(\cdot|s_h, A_h) - P_h(\cdot|s_h, A_h)\|_1 \|\hat{V}_{h+1}^{\pi_k}\|_\infty \leq \sqrt{\frac{SL}{N_h^k(s_h, A_h)}} \cdot H$$

Lemma 1

$$\leq \sum_{h=0}^{H-1} \mathbb{E}^{\pi_k}\left[2H\sqrt{\frac{SL}{N_h^k(s_h, A_h)}}\right]$$

$$= \underbrace{2H\sqrt{SL}} \cdot \underline{\mathbb{E}\left[\sum_{h=0}^{H-1} \sqrt{\frac{1}{N_h^k(s_h^k, A_h^k)}} \,\middle|\, \text{hist}_k\right]}$$

# Total regret

$$\mathbb{E}\left[\sum_{k=0}^{K-1} \text{Regret}_k\right] = \mathbb{E}\left[\sum V^*_{(s_0)} - V^{\pi_k}_{(s_0)}\right] = \mathbb{E}\left[\mathbb{1}(\text{Not fail})\sum_{k=1}^{K-1} V^* - V^{\pi_k}\right] + \mathbb{E}\left[\mathbb{1}(\text{fail})\sum_{k=1}^{K-1}(V^* - V^{\pi_k})\right]$$

$$\le \mathbb{E}\left[\left(\sum_{k=1}^{K-1} V^* - V^{\pi_k}\right)\mathbb{1}(\text{Not fail})\right] + 2\delta \cdot K \cdot H$$

$$\le \mathbb{E}\left[\underbrace{\sum_{k=0}^{K-1}\sum_{h=0}^{H-1}\frac{\boxed{H\sqrt{SL}}}{\sqrt{N_h^k(s_h^k, a_h^k)}}}_{(**)}\right] + 2\delta \cdot K \cdot H \quad (\triangle)$$

Regret$_k$ bound holds under "Not fail"

$$(**) = \sum_{h=0}^{H-1}\sum_{s,a\in S\times A}\sum_{i=1}^{N_h^k(s,a)}\frac{1}{\sqrt{i}} \le \sum_{h=0}^{H-1} 2\sqrt{N_h^k(s,a)}$$

From last lecture.

$$(\triangle) \le \mathbb{E}\left[\sum_{h=1}^{H-1}\sum_{s,a} 2\sqrt{N_h^k(s,a)}\right] + 2\delta \cdot KH$$

$$\le 2\sum_{h=1}^{H-1}\sqrt{SA \cdot \sum_{s,a} N_h^k(s,a)} + 2\delta K \cdot H$$

$$\le 2H\sqrt{SA \cdot K} + 2\delta KH \le 4H\sqrt{SAK}$$

Choose $\delta = \frac{1}{KH}$

Regret $\le 4H^2 S\sqrt{AK}$

# Ideas for improving the dependence on S and H

$\Omega(\sqrt{H^3 SAK})$

$O(\sqrt{H^4 S^2 AK})$

Improve $S$: $\left(\hat{P}_h^k(\cdot|s_h,a_h) - P_h^{\pi_k}(\cdot|s_h,a_h)\right)\left(\hat{V}_{h+1}^{\pi_k} - V_{h+1}^* + V_{h+1}^*\right)$

$= \underbrace{(\hat{P}-P)\cdot V^*}_{\substack{\text{Lemma 2} \\ H\sqrt{\frac{L}{N_{(s,a)}}}}} + \sqrt{\frac{SL}{N}} \cdot H\sqrt{\frac{SL}{N}} \quad \frac{1}{\sqrt{N}}$

$= H\sqrt{\frac{L}{N}} + \frac{SHL}{N}$

Improve $H$: Bernstein's inequality

# Final notes about exploration in Tabular MDPs

- Optimal rates:
  - Non-stationary transitions $\Omega(\sqrt{H^3SAK})$
  - Stationary transitions $\Omega(\sqrt{H^2SAK})$

- 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.