CS 839 Probability and Learning in High Dimension Lecture 19–20 - 04/11,13/2022

Lecture 19–20: Sample Complexity of Reinforcement Learning

Lecturer: Yudong Chen *Scribe: Mingchen Ma, Xindi Lin*

In this two lectures, we introduce three bounds of sample complexity of reinforcement learning with more and more refinement. We will provide the proof of three bounds and see how we got the improvements.

1 Notation

A quick summary of the notation.

1. **MDP:** $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where $S = |\mathcal{S}|, A = |\mathcal{A}|$, $\mathbb{P} \in \mathbb{R}^{SA \times S}, r \in \mathbb{R}^{SA}$
2. **Value and Q functions:** $V^\pi(s) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s\right)$, $Q^\pi(s,a) = \mathbb{E}\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a\right)$
3. **Bellman Equation:** $Q^\pi = r + \gamma \mathbb{P}^\pi Q^\pi$

2 Problem Setup

We have a discounted MDP $M := (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A}$ is the finite action space, $\mathbb{P}$ is the **unknown** transition kernel, r is the reward function, and γ is the discounted factor.

- Let $S := |\mathcal{S}|$ and $A := |\mathcal{A}|$ denote the cardinalities of the state and action spaces, respectively.
- We assume bounded reward, i.e., $\forall s, a : r(s,a) \in [0,1]$. This implies that $\forall s : V^\pi(s) \in [0, 1/(1-\gamma)]$.
- We may view $\mathbb{P}$ as a matrix in $\mathbb{R}^{SA \times S}$ and r as a vector in $\mathbb{R}^{SA}$.
- A γ-discounted MDP is equivalent to a finite horizon MDP with a random horizon $H \sim$ Geometric$(1-\gamma)$. Therefore, $\mathbb{E}[H] = \frac{1}{1-\gamma}$ is called the effective horizon.

Since $\mathbb{P}$ is unknown, we assume we have a **generative model** (a.k.a. sampling oracle, simulator). That is to say, we assume there is a simulator such that for every given pair s, a, one can sample any number of next states s' independently from $\mathbb{P}(\cdot|s,a)$. This is a strong assumption, which allows us to isolate the statistical aspect of RL , ignoring the exploration issues.

Our goal is to study the **sample complexity**, i.e., how many calls of such a simulator we need to get an approximate optimal policy of M. We are interested in the dependence of the sample complexity on S, A and $\frac{1}{1-\gamma}$.

3 Algorithm and Performance Evaluation

We consider a **model-based** apporach. We use samples from generative model to construct an empirical estimation $\hat{\mathbb{P}}$ of the true transition kernel , then compute the corresponding optimal value function/policy of $\hat{\mathbb{P}}$, say using the Value Iteration algorithm mentioned in last lecture. Specifically:

3.1 Algorithm

1. For each (s, a), call the simulator N times; transition to s' for $N(s'|s, a)$ times. The total number of calls is $\bar{N} = NSA$.

2. We compute the empirical transition probability $\hat{\mathbb{P}} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$

$$\hat{\mathbb{P}}(s'|s, a) := \frac{N(s'|s, a)}{N}.$$

   This gives a new MDP $\hat{M} := (\mathcal{S}, \mathcal{A}, \hat{\mathbb{P}}, r, \gamma)$.

3. Based on the empirical MDP, we define the following quantities:

   - $\hat{V}^\pi, \hat{Q}^\pi$: the value function and Q-value function of a policy π evaluated under $\hat{M}$.
   - $\hat{\pi}^*$: the optimal policy of $\hat{M}$.
   - $\hat{V}^*, \hat{Q}^*$: the value function, Q-value function of the optiaml policy $\hat{\pi}^*$ evaluated under $\hat{M}$.

Also recall that Q^* is the Q-function of the optimal policy of the true MDP M.

3.2 Performance Metrics

We are interested in two quantities.

- (Value estimation) How good is $\hat{Q}^*$? That is, how large should $\bar{N}$ be to achieve

$$\left\|\hat{Q}^* - Q^*\right\|_\infty \leq \epsilon.$$

  Here, $\hat{Q}^* = \hat{Q}^{\hat{\pi}^*}$ is the Q-value of the optimal policy $\hat{\pi}^*$ to $\hat{M}$ evaluated under $\hat{M}$, while Q^* is the Q-value of the optimal policy of M evaluated under M.

- (Policy performance) How good is $\hat{\pi}^*$? That is, how large should $\bar{N}$ be to achieve

$$\left\|Q^{\hat{\pi}^*} - Q^*\right\|_\infty \leq \epsilon.$$

  Here, $Q^{\hat{\pi}^*}$ is the Q-value of the optimal policy $\hat{\pi}^*$ to $\hat{M}$ evaluated under the true MDP M. We want to study if $\hat{\pi}^*$ is actually a good policy for the true MDP.

We will establish three bounds for $\bar{N}$, each being tighter than the previous one, that guarantee achievement of the above goals:

- Naive bound:

$$\bar{N} \geq \frac{1}{(1-\gamma)^4}\frac{S^2A}{\epsilon^2}.$$

- Sublinear bound:

$$\bar{N} \geq \frac{1}{(1-\gamma)^4}\frac{SA}{\epsilon^2}.$$

- Optimal bound:

$$\bar{N} \geq \frac{1}{(1-\gamma)^3}\frac{SA}{\epsilon^2}.$$

Note that the algorithm is the same: the model-based approach. We get better bound using more refined analysis.

4 Naive Bound

Theorem 1 ([AJKS19] Proposition 2.1)**.** *Suppose* $\epsilon \in (0, \frac{1}{1-\gamma})$ *and* $\delta \in (0,1)$*. If we obtain*

$$\bar{N} = NSA \geq \frac{c\gamma}{(1-\gamma)^4} \frac{S^2 A \log(cSA/\delta)}{\epsilon^2}$$

samples from the simulator, then with probability at least $1-\delta$*:*

- *(Model accuracy):* $\max_{s,a} \left\| \mathbb{P}(\cdot|s,a) - \hat{\mathbb{P}}(\cdot|s,a) \right\|_1 \leq (1-\gamma)^2\epsilon$.
- *(Uniform value accuracy): for all policies* π, $\left\| Q^\pi - \hat{Q}^\pi \right\|_\infty \leq \epsilon$.
- *(Policy near-optimality): for the optimal policy* $\hat{\pi}^*$ *of* $\hat{M}$*:*

$$\left\| \hat{Q}^* - Q^* \right\|_\infty \leq \epsilon, \quad \left\| Q^{\hat{\pi}^*} - Q^* \right\|_\infty \leq 2\epsilon.$$

4.1 Technical Lemma for Proving Theorem 1

Lemma 1 below can be used to control the error of the model estimate $\hat{\mathbb{P}}$ w.r.t. the true model $\mathbb{P}$. Below one should think of $q = \mathbb{P}(\cdot|s,a)$.

Lemma 1 (Concentration; [AJKS19] Proposition A.8)**.** *Let* q *be a discrete distribution on* $\{1,\ldots,S\}$*. Write* q *as a vector in* $\mathbb{R}^S$ *with* $q_j = Pr[Z_1 = j]$*. Let* $Z_1,\ldots,Z_N \overset{iid}{\sim} q$*. Define the empirical distribution* $\hat{q} \in \mathbb{R}^S$ *as* $\hat{q}_j = \sum_{i=1}^N \frac{1\{Z_i=j\}}{N}$*. For any* $\epsilon > 0$*, we have*

$$\begin{aligned} Pr\left[\|\hat{q} - q\|_1 \geq \sqrt{S}\left(\tfrac{1}{\sqrt{N}} + \epsilon\right) \right] &\leq Pr\left[\|\hat{q} - q\|_2 \geq \frac{1}{\sqrt{N}} + \epsilon \right] \\ &\leq e^{-N\epsilon^2}. \end{aligned}$$

Proof Here, we give a sketch of the proof.

Step 1: The function $(Z_1,\ldots,Z_N) \to \|\hat{q} - q\|_2$ satisfies bounded differences property with parameter $O(1/N)$. By McDiarmids inequality, we can get

$$\Pr\left[\|\hat{q} - q\|_2 - \mathbb{E}\|\hat{q} - q\|_2 \geq \epsilon \right] \leq e^{-N\epsilon^2}.$$

Step 2: Using Jensen's inequality we can show $\mathbb{E}\|\hat{q} - q\|_2 = O(1/\sqrt{N})$. Combining the two bounds, we can prove the statement. For complete proof see, e.g., [HKZ12], Proposition 19. □

Next, we want to control the error of the evaluating a policy under two MDPs. One additional piece of notation: For a given stochastic policy π, define the (row-stochastic) matrix $\mathbb{P}^\pi \in \mathbb{R}^{SA \times SA}$ by

$$\mathbb{P}^\pi_{(s,a),(s',a')} = \mathbb{P}(s'|s,a)\pi(a'|s').$$

This matrix $\mathbb{P}^\pi$ is the probability transition matrix over state-action pairs, where $\mathbb{P}^\pi_{(s,a),(s',a')}$ is the probability of transitioning from (s,a) to (s',a') under the policy π.

Lemma 2 (Simulation Lemma; [AJKS19] Lemma 2.2)**.** *For all* π*,* $Q^\pi - \hat{Q}^\pi = \gamma\left(I - \gamma\hat{\mathbb{P}}^\pi\right)^{-1}\left(\mathbb{P} - \hat{\mathbb{P}}\right)V^\pi$.

Proof From Bellman equation for Q-functions:

$$Q^\pi = r + \gamma \mathbb{P}^\pi Q^\pi, \quad \text{equivalently, } Q^\pi = (I - \gamma \mathbb{P}^\pi)^{-1} r. \tag{1}$$

Here $I - \gamma \mathbb{P}^\pi$ is invertible because eigenvalues of $\mathbb{P}^\pi$ are bounded by 1 (Perron-Frobenius) and $\gamma < 1$. So

$$\begin{aligned} Q^\pi - \hat{Q}^\pi &= Q^\pi - \big(I - \gamma \hat{\mathbb{P}}^\pi\big)^{-1} r \\ &= \big(I - \gamma \hat{\mathbb{P}}^\pi\big)^{-1} \Big[\big(I - \gamma \hat{\mathbb{P}}^\pi\big) - (I - \gamma \mathbb{P}^\pi)\Big] Q^\pi \\ &= \gamma \big(I - \gamma \hat{\mathbb{P}}^\pi\big)^{-1} \big(\mathbb{P}^\pi - \hat{\mathbb{P}}^\pi\big) Q^\pi = \gamma \big(I - \gamma \hat{\mathbb{P}}^\pi\big)^{-1} \big(\mathbb{P} - \hat{\mathbb{P}}\big) V^\pi. \end{aligned}$$

Here, the second equation follows by Bellman equation. □

The matrix $(I - \gamma \mathbb{P}^\pi)^{-1}$ plays an important role in MDPs. We have the following lemma on the ℓ_∞ operator norm of this matirx.

Lemma 3 (Operator Norm Lemma; [AJKS19] Lemma 2.3)**.** *For all π and $v \in \mathbb{R}^{SA}$, $\big\|(I - \gamma \mathbb{P}^\pi)^{-1} v\big\|_\infty \le \frac{1}{1-\gamma} \|v\|_\infty$. That is, $\|(I - \gamma \mathbb{P}^\pi)^{-1}\|_{\ell_\infty \to \ell_\infty} \le \frac{1}{1-\gamma}$.*

Proof Let $w := (I - \gamma \mathbb{P}^\pi)^{-1} v$. Then

$$\begin{aligned} \|v\|_\infty &= \|(I - \gamma \mathbb{P}^\pi)\, w\|_\infty \\ &\ge \|w\|_\infty - \gamma \|\mathbb{P}^\pi w\|_\infty \\ &\ge \|w\|_\infty - \gamma \|w\|_\infty, \end{aligned}$$

where last step holds since each element of $\mathbb{P}^\pi w$ is a weighted average of w. Rearranging terms proves the lemma. □

4.2 Proof of Theorem 1

- Model accuracy:

  We first fix a pair (s, a). By Lemma 1, we have

  $$\mathbb{P}\left(\left\|\mathbb{P}(\cdot|s,a) - \hat{\mathbb{P}}(\cdot|s,a)\right\|_1 \ge \sqrt{S}\left(\frac{1}{\sqrt{N}} + \epsilon\right)\right) \le e^{-N\epsilon^2}.$$

  By union bound, we get

  $$\mathbb{P}\left(\exists (s,a) : \left\|\mathbb{P}(\cdot|s,a) - \hat{\mathbb{P}}(\cdot|s,a)\right\|_1 \ge \sqrt{S}\left(\frac{1}{\sqrt{N}} + \epsilon\right)\right) \le SAe^{-N\epsilon^2}.$$

  By choosing $N = \frac{c\gamma}{(1-\gamma)^4} \frac{S \log(cSA/\delta)}{\epsilon^2}$, we get

  $$\mathbb{P}\left(\forall (s,a) : \left\|\mathbb{P}(\cdot|s,a) - \hat{\mathbb{P}}(\cdot|s,a)\right\|_1 \le (1-\gamma)^2 \epsilon\right) \ge 1 - \delta.$$

  This proves the model accuracy.

- Value accuracy: Applying Lemma 2 (simulation) and Lemma 3 (operator norm), we have

$$\begin{aligned}\left\|Q^\pi - \hat{Q}^\pi\right\|_\infty &= \left\|\gamma(I-\gamma\hat{\mathbb{P}}^\pi)^{-1}(\mathbb{P}-\hat{\mathbb{P}})V^\pi\right\|_\infty \le \frac{\gamma}{1-\gamma}\left\|(\mathbb{P}-\hat{\mathbb{P}})V^\pi\right\|_\infty \\ &\le \frac{\gamma}{1-\gamma}\left(\max_{s,a}\left\|\mathbb{P}(\cdot|s,a)-\hat{\mathbb{P}}(\cdot|s,a)\right\|_1\right)\|V^\pi\|_\infty && \text{Holder's inequality} \\ &\le \frac{\gamma}{1-\gamma}\cdot(1-\gamma)^2\epsilon\cdot\frac{1}{1-\gamma} = \gamma\epsilon. && \text{Model accuracy bound}\end{aligned}$$

- Policy near-optimality, 1st inequality: $\forall s, a$, we have

$$\begin{aligned}\left|\hat{Q}^*(s,a) - Q^*(s,a)\right| &= \left|\sup_\pi \hat{Q}^\pi(s,a) - \sup_\pi Q^\pi(s,a)\right| \\ &\le \sup_\pi\left|\hat{Q}^\pi(s,a) - Q^\pi(s,a)\right| \le \epsilon.\end{aligned}$$

- Policy near-optimality, 2nd inequality: using triangle inequality and the last two bounds, we have

$$\left\|Q^{\hat{\pi}^*} - Q^*\right\|_\infty \le \left\|Q^{\hat{\pi}^*} - \hat{Q}^{\hat{\pi}^*}\right\|_\infty + \left\|\hat{Q}^{\hat{\pi}^*} - Q^*\right\|_\infty \le \epsilon + \epsilon.$$

5 Sublinear Bound

We notice that previous approach estimates *every entry* of $\mathbb{P}$, and V^π for *every policy* π. What we actually care about, however, is V^* and π^*, for which estimating every entry of $\mathbb{P}$ is an overkill. With this observation in mind, we can improve the sample complexity from S^2A to SA.

Theorem 2 ([AJKS19] Proposition 2.4)**.** *Define*

$$\Delta_{\delta,N} := \frac{\gamma}{(1-\gamma)^2}\sqrt{\frac{2\log(2SA/\delta)}{N}}.$$

With probability at lest $1-\delta$*,*

$$\left\|\hat{Q}^* - Q^*\right\|_\infty \le \Delta_{\delta,N}, \quad \left\|\hat{Q}^{\pi^*} - Q^*\right\|_\infty \le \Delta_{\delta,N}, \quad \left\|Q^{\hat{\pi}^*} - Q^*\right\|_\infty \le \frac{1}{1-\gamma}\Delta_{\delta,N}.$$

This theorem implies that $\|Q^* - \hat{Q}^*\|_\infty \le \epsilon$ if $\bar{N} \gtrsim \frac{\gamma^2}{(1-\gamma)^4}\frac{SA\log(SA/\delta)}{\epsilon^2}$.

5.1 Technical Lemma for Proving Theorem 2

Lemma 4 (Component-wise bounds; [AJKS19] Lemma 2.5)**.** *We have*

$$\begin{aligned}Q^* - \hat{Q}^* &\le \gamma(I-\gamma\hat{\mathbb{P}}^{\pi^*})^{-1}(\mathbb{P}-\hat{\mathbb{P}})V^*, \\ Q^* - \hat{Q}^* &\ge \gamma(I-\gamma\hat{\mathbb{P}}^{\hat{\pi}^*})^{-1}(\mathbb{P}-\hat{\mathbb{P}})V^*,\end{aligned}$$

where the inequalities above are component-wise vector inequalities.

Proof

1st inequality: By optimality of $\hat{\pi}^*$ to $\hat{M}$ and Simulation Lemma 2, we have

$$Q^* - \hat{Q}^* \le Q^{\pi^*} - \hat{Q}^{\pi^*} = \gamma\big(I-\gamma\hat{\mathbb{P}}^{\pi^*}\big)^{-1}(\mathbb{P}-\hat{\mathbb{P}})V^*.$$

2nd inequality: we have

$$\begin{aligned}
Q^* - \hat{Q}^* &= Q^* - \left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1} r && \text{by eq. (1)} \\
&= \left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1} \left[\left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right) - \left(I - \gamma \mathbb{P}^{\pi^*}\right)\right] Q^* && \text{by eq. (1)} \\
&= \gamma \left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1} \left(\mathbb{P}^{\pi^*} - \hat{\mathbb{P}}^{\hat{\pi}^*}\right) Q^* \\
&\geq \gamma \left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1} \left(\mathbb{P}^{\pi^*} - \hat{\mathbb{P}}^{\pi^*}\right) Q^* \\
&= \gamma \left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1} \left(\mathbb{P} - \hat{\mathbb{P}}\right) V^*,
\end{aligned}$$

where the inequality follows from $\hat{\mathbb{P}}^{\hat{\pi}^*} Q^* \leq \hat{\mathbb{P}}^{\pi^*} Q^*$ (proved below) and entry-wise non-negativity of $\left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1}$ (proved below).

Proof of $\hat{\mathbb{P}}^{\hat{\pi}^*} Q^* \leq \hat{\mathbb{P}}^{\pi^*} Q^*$: Since $\pi^*(s') = \arg\max_a Q^*(s', a), \forall s'$ (π^* chooses the action a that maximizes $Q^*(s', a)$ for every fixed s'), we have

$$\begin{aligned}
(\hat{\mathbb{P}}^{\pi^*} Q^*)_{s,a} &= \sum_{(s',a')} \hat{\mathbb{P}}^{\pi^*}_{(s,a),(s',a')} Q^*_{(s',a')} \\
&= \sum_{(s',a')} \hat{\mathbb{P}}(s' \mid s, a) \pi^*(a' \mid s') Q^*_{(s',a')} \\
&\geq \sum_{(s',a')} \hat{\mathbb{P}}(s' \mid s, a) \hat{\pi}^*(a' \mid s') Q^*_{(s',a')} \\
&= (\hat{\mathbb{P}}^{\hat{\pi}^*} Q^*)_{s,a}.
\end{aligned}$$

Proof that $\left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1}$ is entry-wise non-negative: We can expand $\left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1}$ using the Neuman series:

$$\begin{aligned}
\left[(1-\gamma) \cdot \left(I - \gamma \mathbb{P}^{\pi}\right)^{-1}\right]_{(s,a),(s',a')} &= (1-\gamma) \sum_{t=0}^{\infty} \left(\gamma \mathbb{P}^{\pi}\right)^t \\
&= (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^{\pi} \left(s_t = s', a_t = a' \mid s_0 = s, a_0 = a\right), \qquad (2)
\end{aligned}$$

which shows that every entry of $\left(I - \gamma \hat{\mathbb{P}}^{\hat{\pi}^*}\right)^{-1}$ is non-negative. □

Remark From equation (2), we see that for each (s, a), we have $\sum_{s',a'} \left[(1-\gamma) \cdot \left(I - \gamma \mathbb{P}^{\pi}\right)^{-1}\right]_{(s,a),(s',a')} = 1$. Therefore, $\left[(1-\gamma) \cdot \left(I - \gamma \mathbb{P}^{\pi}\right)^{-1}\right]_{(s,a),(\cdot,\cdot)}$ is a probability measure on $\mathcal{S} \times \mathcal{A}$, called the occupancy measure.

5.2 Proof of Theorem 2

By Lemmas 2 and 3, we have

$$\left\| Q^* - \hat{Q}^{\pi^*} \right\|_\infty = \left\| \gamma (I - \gamma \hat{\mathbb{P}}^{\pi^*})^{-1} (\mathbb{P} - \hat{\mathbb{P}}) V^* \right\|_\infty \leq \frac{\gamma}{1-\gamma} \left\| (\mathbb{P} - \hat{\mathbb{P}}) V^* \right\|_\infty.$$

By Lemmas 4 and 3, we have

$$\left\| Q^* - \hat{Q}^* \right\|_\infty \leq \frac{\gamma}{1-\gamma} \left\| (\mathbb{P} - \hat{\mathbb{P}}) V^* \right\|_\infty.$$

It remains to bound $\left\|(\mathbb{P}-\hat{\mathbb{P}})V^*\right\|_\infty$. Instead of using Lemma 1, we bound this term by applying Hoeffding's inequality and union bound. Doing so gives that with probability at least $1-\delta$,

$$\begin{aligned}\left\|(\mathbb{P}-\hat{\mathbb{P}})V^*\right\|_\infty &= \max_{s,a}\left|\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[V^*(s')\right]-\mathbb{E}_{s'\sim\hat{\mathbb{P}}(\cdot|s,a)}\left[V^*(s')\right]\right| \\ &\leq \frac{1}{1-\gamma}\sqrt{\frac{2\log\left(2SA/\delta\right)}{N}}. \end{aligned} \tag{3}$$

Combining the last three display equations proves Theorem 2.

Remark Reason for bound improvement: When we derived the naive bound, we controlled the difference $\mathbb{P}-\hat{\mathbb{P}}$ between two distributions, which has SA^2 entries. Here, we instead bound the quantity $\mathbb{P}V^*-\hat{\mathbb{P}}V^*$, which is the difference between the *expectation* of two distributions and only has SA entries.

In general, estimating a functional (e.g., expectation, variance, entropy) of a distribution is often easier than estimating the entire distribution. In our proof, we are estimating the expectation under $\mathbb{P}$ rather than the distribution $\mathbb{P}$ itself.

6 Optimal Bound

We can further improve the effective horizon factor $\frac{1}{(1-\gamma)^4}$ in the sublinear bound (Theorem 2) to $\frac{1}{(1-\gamma)^3}$. Instead using Hoeffding's inequality to bound $\left\|(\mathbb{P}-\hat{\mathbb{P}})V^*\right\|_\infty$. This is done in the following two theorems.

Theorem 3 (Value Accuracy; [AJKS19] Theorem 2.6)**.** *With probability at least* $1-\delta$*,*

$$\left\|Q^*-\hat{Q}^*\right\|_\infty \leq \gamma\sqrt{\frac{c}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}}+\frac{c\gamma}{(1-\gamma)^3}\frac{\log(cSA/\delta)}{N}.$$

Theorem 4 (Policy performance; [AJKS19] Theorem 2.8)**.** *For* $\epsilon\leq\sqrt{\frac{1}{1-\gamma}}$*, if*

$$\bar{N}\geq\frac{c}{(1-\gamma)^3}\frac{SA\log(cSA/\delta)}{\epsilon^2},$$

then with probability at least $1-\delta$*,* $\left\|Q^*-\hat{Q}^*\right\|_\infty\leq\epsilon$ *and* $\left\|Q^*-Q^{\hat{\pi}^*}\right\|_\infty\leq\epsilon$*.*

Before we prove these two theorems, we compare with the following minimax lower bound on the sample complexity.

Theorem 5 (Minimax lower bound; [AJKS19] Theorem 2.8)**.** *If an algorithm achieves* $\left\|Q^*-\hat{Q}^*\right\|_\infty\leq\epsilon$ *with probability at least* $1-\delta$*, then it must require a total number of samples*

$$\bar{N}\geq\frac{c}{(1-\gamma)^3}\frac{SA\log(cSA/\delta)}{\epsilon^2}.$$

This lower bound indicates that the sample complexity bound in Theorem 4 is a minimax optimal. The proof of Theorem 5 can be found in [AMK12] and [SWW+18].

6.1 Proof of Theorem 3

Our **key idea** is to keep track of variance. In particular, we establish the following component-wise bound by applying scalar Bernstein inequality

$$\left|(\mathbb{P}-\hat{\mathbb{P}})V^*\right| \overset{\text{component-wise}}{\leq} \sqrt{\frac{2\log(2SA/\delta)}{N}}\sqrt{\mathrm{Var}_{\mathbb{P}}\left[V^*\right]}+\mathcal{O}(\frac{1}{N}),$$

where $\mathrm{Var}_{\mathbb{P}}\left[V^*\right]\in\mathbb{R}^{SA}$ is the variance vector, defined as

$$\mathrm{Var}_{\mathbb{P}}\left[V^*\right](s,a) := \mathrm{Var}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[V^*(s')\right], \quad \forall s,a.$$

Note that this is a tighter bound than what we used in the proof of Theorem 2, where bound the $\|\cdot\|_\infty$ norm of the same quantity using Hoeffding.

Then we combine with the following lemma to bound the variance, thereby completing the proof of Theorem 3.

Lemma 5 (Weighted variance bound)**.** *For any policy π:*

$$\left\|(I-\gamma\mathbb{P}^\pi)^{-1}\sqrt{\mathrm{Var}_{\mathbb{P}}\left[V^\pi\right]}\right\|_\infty \leq \sqrt{\frac{2}{(1-\gamma)^3}}.$$

6.2 Technical Intuition for Refinement

How would one come up with the above refined proof in the first place? It is not a priori obvious which part of proof of sublinear bound (Theorem 2) is not tight. It is also non-trivial to realize that keeping track of the variance is the right way to go. Below we give an informal argument that provides some intuition on why this is the right way.

Recall that in the proof of Theorem 2, we established the following inequalities:

$$\begin{aligned}
\left\|Q^*-\hat{Q}^{\pi^*}\right\|_\infty &= \left\|\gamma(I-\gamma\hat{\mathbb{P}}^{\pi^*})^{-1}(\mathbb{P}-\hat{\mathbb{P}})V^*\right\|_\infty && \text{Simulation Lemma 2} \\
&\leq \frac{\gamma}{1-\gamma}\left\|(\mathbb{P}-\hat{\mathbb{P}})V^*\right\|_\infty && \text{Operator norm Lemma 3} \quad (4) \\
&\leq \frac{\gamma}{1-\gamma}\frac{1}{1-\gamma}\sqrt{\frac{2\log\left(2SA/\delta\right)}{N}} && \text{Hoeffding} \quad (5)
\end{aligned}$$

Both (4) and (5) are tight in the worst case; we argue that they cannot be tight *simultaneously*. Here, (4) is tight when $(\mathbb{P}-\hat{\mathbb{P}})V^*$ is proportional to the constant vector $\mathbb{1}$, which is the top eigenvector of the row-stochastic matrix $(1-\gamma)(I-\gamma\hat{\mathbb{P}}^{\pi^*})^{-1}$ (see (2)).[1]

If $(\mathbb{P}-\hat{\mathbb{P}})V^*$ is a constant vector, then the variance $\mathrm{Var}_{s'\sim\mathbb{P}(\cdot|s,a)}[V^*(s')]$ is also constant across (s,a). In turn, this means the transition probabilities $\mathbb{P}(\cdot|s,a)$ does not depend on (s,a), in which case the values $Q^*(s,a)$ and $V^*(s)$ are similar across (s,a). Consequently, we must have $\mathrm{Var}_{\mathbb{P}}\left[V^*\right]\ll\frac{1}{1-\gamma}$, so the Hoeffding's inequality used in (5) is not tight.

The above argument suggests that in order to improve the bound, we should not separate the two steps (4) and (5).

6.3 Proof of Theorem 4

Theorem 3 controls $\left\|\hat{Q}^*-Q^*\right\|_\infty$. To translate this to a bound on $V^*-V^{\hat{\pi}^*}$, one may be tempted to apply the the crude bound below.

[1]For a row-stochastic matrix A with $\sum_j A_{ij}=1$, we have $A\mathbb{1}=\mathbb{1}$).

Lemma 6 (Q error amplification; Lemma 1.11 in [AJKS19]). *Let* $\mathbf{1} \in \mathbb{R}^S$ *be the all-one vector. Then*

$$V^* - V^{\hat{\pi}^*} \le \frac{2}{1-\gamma} \left\| \hat{Q}^* - Q^* \right\|_\infty \mathbf{1}$$

Applying this lemma gives $\left\| Q^* - Q^{\hat{\pi}^*} \right\|_\infty \le \frac{\|\hat{Q}^* - Q^*\|_\infty}{1-\gamma}$, which has a superfluous effective horizon factor $\frac{1}{1-\gamma}$.

To remove $\frac{1}{1-\gamma}$, it turns out that we need a better understanding of the quantity $(\mathbb{P} - \hat{\mathbb{P}})\hat{V}^*$. The main challenge here is the **probabilistic dependency** between $\hat{\mathbb{P}}$ and $\hat{V}^*$. To address this issue, the key idea is to use the so-called **leave-one-out** analysis (which is called absorbing MDP in [AKY20]). Below we sketch this analysis.

Leave-one-out analysis in the proof of Theorem 4: Fix an arbitrary state s_0. We construct a new empirical MDP $\hat{M}_0 = (\mathcal{S}, \mathcal{A}, \hat{\mathbb{P}}_0, r, \gamma)$ that is identical to the original empirical MDP $\hat{M}$ except that s_0 is an absorbing state. Explicitly,

$$\hat{\mathbb{P}}_0(\cdot|s,a) = \begin{cases} \hat{\mathbb{P}}(\cdot|s,a) & s \neq s_0 \\ \mathbb{1}_{s_0} & s = s_0 \end{cases}$$

Let $\hat{V}_0^*$ be the optimal value function for $\hat{M}_0$.

Remark We note that the new empirical MDP $\hat{M}_0$ is used only in analysis, not in algorithm.

To proceed, our first obervation is that by construction, $\hat{\mathbb{P}}(\cdot|s_0,a)$ and $\hat{V}_0^*$ are independent. Therefore, using standard concentration inequalities it is easy to bound the quantity

$$\left[(\mathbb{P} - \hat{\mathbb{P}})\hat{V}_0^*\right]_{s_0,a} = \sum_{s'} \left(\mathbb{P}(s'|s_0,a) - \hat{\mathbb{P}}(s'|s_0,a)\right) \hat{V}_0^*(s').$$

Our second observation is that $\hat{V}_0^*$ and $\hat{V}^*$ are close, since we only alter one state when constructing $\hat{M}_0$. Therefore, the above bound on $\left[(\mathbb{P} - \hat{\mathbb{P}})\hat{V}_0^*\right]_{s_0,a}$ implies bound on $\left[(\mathbb{P} - \hat{\mathbb{P}})\hat{V}^*\right]_{s_0,a}$, the quantity we want to control in the first place.

Repeating the above analysis for every s_0, we can obtain a tight bound on $\|(\mathbb{P} - \hat{\mathbb{P}})\hat{V}^*\|_\infty$.

7 Further Reading on Sample Complexity of RL

Model-based approach:

- Optima value bound (Thm 3) is due to [GAMK13]
- Optimal policy bound (Thm 4) is due to [AKY20]
- Lower bound (Thm 5) is due to [AMK12]
- Thm 4 holds for $\epsilon \in \left(0, \frac{1}{\sqrt{1-\gamma}}\right]$. Improvement to $\epsilon \in \left(0, \frac{1}{1-\gamma}\right]$: [LWC$^+$20]

Model-free approach:

- Q-learning is sub-optimal, with sample complexity $\propto \frac{SA}{(1-\gamma)^4}$, both in theory and numerically: [Wai19a]
- Q-learning with *variance reduction* is optimal, with sample complexity $\propto \frac{SA}{(1-\gamma)^3}$: [SWW$^+$18], [Wai19b]

8 What's Next

The sample complexity bounds we have established in this lecture are proportional to S and A, the cardinalities of the state and action spaces. For problems with large S and A, these bounds would require a large number of samples. Using function approximation is one way to address this issue. We will study this approach in the next few lectures.

Below we give a preview of a probabilistic tool that we will use when analyzing function approximation approaches. Consider the sum $\sum_{s=1}^{k} \phi_s \epsilon_s$, where ϵ_s are i.i.d. Rademacher random variables and $\phi_s = 1, \forall s$. By Hoeffding bound, we have

$$\frac{|\sum_{s=1}^{k} \phi_s \epsilon_s|}{\sqrt{\sum_{s=1}^{k} \phi_s^2}} \sim \sqrt{\log(\frac{1}{\delta})}.$$

with high probability. It turns out the above bound still holds even when ϕ_s are arbitrary i.i.d. random variables (which may be unbounded and heavy-tailed). This result is called a self-normalized concentration inequality.

For intuitive understanding, consider the ratio

$$\frac{|\sum_{s=1}^{k} X_s|}{\sqrt{\sum_{s=1}^{k} X_s^2}},$$

which is the sum of (potentially heavy-tailed) random variables normalized by their squared variation. If the numerator has a heavy tail, then the denominator will also have a heavy tail. It turns out these two effect cancel out with each other, so the ratio has a good concentration property.

References

[AJKS19] Alekh Agarwal, Nan Jiang, Sham M Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. 2019.

[AKY20] Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pages 67–83. PMLR, 2020.

[AMK12] Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kappen. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

[GAMK13] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax pac bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91(3):325–349, 2013.

[HKZ12] Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[LWC+20] Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Breaking the sample size barrier in model-based reinforcement learning with a generative model. *Advances in neural information processing systems*, 33:12861–12872, 2020.

[SWW+18] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018.

[Wai19a] Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

[Wai19b] Martin J Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.