

Hi!

# **The Sample Complexity (with a Generative Model)**

**Sham Kakade and Kianté Brantley**

**CS 2824: Foundations of Reinforcement Learning**

# Announcements

- **HW1** is posted now!
  - it cover many concepts from class
- **Reading assignments**
  - please do the readings; **they are helpful for our lectures**
  - next reading assignment posted soon

Today:

# Today:

- Recap: computational complexity
  - Question: Given an MDP  $\mathcal{M} = (S, A, P, r, \gamma)$  can we **exactly compute**  $Q^*$  (or find  $\pi^*$ ) in polynomial time?

# Today:

- Recap: computational complexity
  - Question: Given an MDP  $\mathcal{M} = (S, A, P, r, \gamma)$  can we **exactly compute**  $Q^*$  (or find  $\pi^*$ ) in polynomial time?
- Today: **statistical complexity**
  - Question: Given only sampling access to an unknown MDP  $\mathcal{M} = (S, A, P, r, \gamma)$  how many **observed transitions do we need to estimate**  $Q^*$  (or find  $\pi^*$ )?

Recap

# Summary Table

	Value Iteration	Policy Iteration	LP-based Algorithms
Poly.	$S^2 A \frac{L(P,r,\gamma) \log \frac{1}{1-\gamma}}{1-\gamma}$	$(S^3 + S^2 A) \frac{L(P,r,\gamma) \log \frac{1}{1-\gamma}}{1-\gamma}$	$S^3 A L(P, r, \gamma)$
Strongly Poly.	X	$(S^3 + S^2 A) \cdot \min \left\{ \frac{A^S}{S}, \frac{S^2 A \log \frac{S^2}{1-\gamma}}{1-\gamma} \right\}$	$S^4 A^4 \log \frac{S}{1-\gamma}$

- VI: poly time for **fixed**  $\gamma$ , not strongly poly
- PI: poly and strongly-poly time for **fixed**  $\gamma$
- LP approach: poly and strongly-poly time  
(LP approach is only logarithmic in  $1/(1 - \gamma)$ )

Today

# Sampling Models

(for learning in an unknown MDP)

w.p.r.  $(1-\gamma)$  episode ends.

- **Episodic setting:**
  - in every episode,  $s_0 \sim \mu$ .
  - the learner acts for some finite number of steps and observes the trajectory.
  - The state then resets to  $s_0 \sim \mu$ .

# Sampling Models

(for learning in an unknown MDP)

- **Episodic setting:**

- in every episode,  $s_0 \sim \mu$ .
- the learner acts for some finite number of steps and observes the trajectory.
- The state then resets to  $s_0 \sim \mu$ .

$$s' \sim P(\cdot | s, a)$$

- **Offline RL setting:**

- You have a collection of observed transitions/reward:  $\{(s, a, s', r(s, a))\}$
- You don't have control of the data generating distribution.

# Sampling Models

(for learning in an unknown MDP)

- **Episodic setting:**
  - in every episode,  $s_0 \sim \mu$ .
  - the learner acts for some finite number of steps and observes the trajectory.
  - The state then resets to  $s_0 \sim \mu$ .
- **Offline RL setting:**
  - You have a collection of observed transitions/reward:  $\{(s, a, s', r(s, a))\}$
  - You don't have control of the data generating distribution.
- **Generative model setting:**
  - input:  $(s, a)$  output: a sample  $s' \sim P(\cdot | s, a)$  and  $r(s, a)$
  - provides an “idealized model” to study statistical limits

# Sampling Models

## (for learning in an unknown MDP)

- **Episodic setting:**
  - in every episode,  $s_0 \sim \mu$ .
  - the learner acts for some finite number of steps and observes the trajectory.
  - The state then resets to  $s_0 \sim \mu$ .
- **Offline RL setting:**
  - You have a collection of observed transitions/reward:  $\{(s, a, s', r(s, a))\}$
  - You don't have control of the data generating distribution.
- **Generative model setting:**
  - input:  $(s, a)$  output: a sample  $s' \sim P(\cdot | s, a)$  and  $r(s, a)$
  - provides an “idealized model” to study statistical limits
- **Sample complexity of RL:**  
how many transitions do we need observe in order to find a near optimal policy?

# Two Fundamental Questions in Sample Complexity

$$P(s|s_a)$$

## 1. The Model Size Question (Sublinear Learning)

- How many parameters do we need to specify the transition kernel  $P$ ?  $O(S^2A)$   
(and how many for the policy?)  $O(SA)$
- Q1: Can we find an  $\epsilon$ -optimal policy with **sublinear** sample complexity?

$$\Omega(S^2A)$$

$$O(SA)$$

transitions suffice?

# Two Fundamental Questions in Sample Complexity

## 1. The Model Size Question (Sublinear Learning)

- How many parameters do we need to specify the transition kernel  $P$ ? (and how many for the policy?)
- Q1: Can we find an  $\epsilon$ -optimal policy with **sublinear** sample complexity?

## 2. The Horizon Question (Error Amplification)

- The Scale: In discounted settings, the values scale as  $1/(1 - \gamma)$ .
- Target: It is natural to measure our additive error  $\epsilon$  relative to this scale.
- Q2: What is the “**horizon amplification**”? i.e. the dependence on  $1/(1 - \gamma)$

# Attempt 1:

the naive model based approach

# The most naive approach: model based

- Today: let us assume access to a [generative model](#)

# The most naive approach: model based

- Today: let us assume access to a **generative model**
- most naive approach to learning:
  - Call our simulator **N times at each state action pair.**
  - Let  $\hat{P}$  be our empirical model:

$$\hat{P}(s' | s, a) = \frac{\text{count}(s', s, a)}{N}$$

where  $\text{count}(s', s, a)$  is the #times  $(s, a)$  transitions to state  $s'$ .

- we know the reward after one call.

(for simplicity, we often assume  $r(s, a)$  is deterministic)

# call to gen  
model.  
SAN

# The most naive approach: model based

- Today: let us assume access to a **generative model**
- most naive approach to learning:
  - Call our simulator **N times at each state action pair.**
  - Let  $\hat{P}$  be our empirical model:

$$\hat{P}(s' | s, a) = \frac{\text{count}(s', s, a)}{N}$$

where  $\text{count}(s', s, a)$  is the #times  $(s, a)$  transitions to state  $s'$ .

- we know the reward after one call.  
(for simplicity, we often assume  $r(s, a)$  is deterministic)
- The total number of calls to our generative model is **SAN**.

# Model accuracy

**Proposition:**  $c$  is an absolute constant.  $\epsilon > 0$ . For  $N \geq \frac{c\gamma}{(1-\gamma)^4} \frac{S \log(cSA/\delta)}{\epsilon^2}$   
and with probability greater than  $1 - \delta$ ,

# Model accuracy

**Proposition:**  $c$  is an absolute constant.  $\epsilon > 0$ . For  $N \geq \frac{c\gamma}{(1-\gamma)^4} \frac{S \log(cSA/\delta)}{\epsilon^2}$  and with probability greater than  $1 - \delta$ ,

- Model accuracy: The transition model is  $\epsilon$  has error bounded as:

$$\max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \leq (1 - \gamma)^2 \epsilon / 2.$$

# Model accuracy

**Proposition:**  $c$  is an absolute constant.  $\epsilon > 0$ . For  $N \geq \frac{c\gamma}{(1-\gamma)^4} \frac{S \log(cSA/\delta)}{\epsilon^2}$

and with probability greater than  $1 - \delta$ ,

- Model accuracy: The transition model is  $\epsilon$  has error bounded as:

$$\max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \leq (1 - \gamma)^2 \epsilon / 2.$$

- Uniform value accuracy: For all policies  $\pi$ ,

$$\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \epsilon / 2$$

# Model accuracy

**Proposition:**  $c$  is an absolute constant.  $\epsilon > 0$ . For  $N \geq \frac{c\gamma}{(1-\gamma)^4} \frac{S \log(cSA/\delta)}{\epsilon^2}$  and with probability greater than  $1 - \delta$ ,

- Model accuracy: The transition model is  $\epsilon$  has error bounded as:

$$\max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \leq (1 - \gamma)^2 \epsilon / 2.$$

- Uniform value accuracy: For all policies  $\pi$ ,

$$\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \epsilon / 2$$

- Near optimal planning: Suppose that  $\hat{\pi}^*$  is the optimal policy in  $\widehat{M}$ .

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

# Matrix Expressions

- View  $P$  as a matrix of size  $SA \times S$  (rows are probability distributions)

# Matrix Expressions

- View  $P$  as a matrix of size  $SA \times S$  (rows are probability distributions)
- Define  $P^\pi$  to be the transition matrix on state-action pairs (for deterministic  $\pi$ ):

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

# Matrix Expressions

- View  $P$  as a matrix of size  $SA \times S$  (rows are probability distributions)
- Define  $P^\pi$  to be the transition matrix on state-action pairs (for deterministic  $\pi$ ):

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

$$Q^\pi (s, \pi(s)) = V^\pi(s)$$

- With this notation,  
 $Q^\pi = r + \gamma P V^\pi$   
 $Q^\pi = r + \gamma P^\pi Q^\pi$

$$P^\pi Q^\pi = P V^\pi$$

# Matrix Expressions

- View  $P$  as a matrix of size  $SA \times S$  (rows are probability distributions)
- Define  $P^\pi$  to be the transition matrix on state-action pairs (for deterministic  $\pi$ ):

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

- With this notation,

$$Q^\pi = r + \gamma P V^\pi$$

$$Q^\pi = r + \gamma P^\pi Q^\pi$$

- Also,

$$Q^\pi = (I - \gamma P^\pi)^{-1} r$$

(where one can show the inverse exists)

# “Simulation” Lemma

“Simulation Lemma”: For all  $\pi$ ,

$$Q^\pi - \widehat{Q}^\pi = \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi$$

# “Simulation” Lemma

“Simulation Lemma”: For all  $\pi$ ,

$$Q^\pi - \widehat{Q}^\pi = \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi$$

**Proof:** Using our matrix equality for  $Q^\pi$ , we have:

$$\begin{aligned} Q^\pi - \widehat{Q}^\pi &= Q^\pi - (I - \gamma \widehat{P}^\pi)^{-1}r \\ &= (I - \gamma \widehat{P}^\pi)^{-1}((I - \gamma \widehat{P}^\pi) - (I - \gamma P^\pi))Q^\pi \\ &= \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P^\pi - \widehat{P}^\pi)Q^\pi \\ &= \gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi \end{aligned}$$

# Proof of Claim 1

# Proof of Claim 1

- Concentration of a distribution in the  $\ell_1$  norm:
  - For a fixed  $s, a$ . With pr greater than  $1 - \delta$ ,

$$\|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \leq c \sqrt{\frac{S \log(1/\delta)}{N}}$$

with  $N$  samples used to estimate  $\widehat{P}(\cdot | s, a)$ .

# Proof of Claim 1

- Concentration of a distribution in the  $\ell_1$  norm:
  - For a fixed  $s, a$ . With pr greater than  $1 - \delta$ ,

$$\|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \leq c \sqrt{\frac{S \log(1/\delta)}{N}}$$

with  $N$  samples used to estimate  $\widehat{P}(\cdot | s, a)$ .

- The first claim now follows by the union bound.

# Proof of Claim 2 (&3)

## Proof of Claim 2 (&3)

For the second claim,

$$\|Q^\pi - \widehat{Q}^\pi\|_\infty = \|\gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi\|_\infty$$

$$\begin{aligned} & \|(\mathbb{I} - \gamma P^\pi)^{-1}x\|_\infty \\ & \leq \underbrace{\frac{1}{1-\gamma}} \|x\|_\infty \end{aligned}$$

# Proof of Claim 2 (&3)

For the second claim,

$$\begin{aligned}\|Q^\pi - \widehat{Q}^\pi\|_\infty &= \|\gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi\|_\infty \\ &\leq \frac{\gamma}{1 - \gamma} \|(P - \widehat{P})V^\pi\|_\infty \\ &\leq \frac{\gamma}{1 - \gamma} \left( \max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \right) \|V^\pi\|_\infty \\ &\leq \frac{\gamma}{(1 - \gamma)^2} \max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1\end{aligned}$$

(why is the first inequality true?)

# Proof of Claim 2 (&3)

For the second claim,

$$\begin{aligned}\|Q^\pi - \widehat{Q}^\pi\|_\infty &= \|\gamma(I - \gamma \widehat{P}^\pi)^{-1}(P - \widehat{P})V^\pi\|_\infty \\ &\leq \frac{\gamma}{1 - \gamma} \|(P - \widehat{P})V^\pi\|_\infty \\ &\leq \frac{\gamma}{1 - \gamma} \left( \max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1 \right) \|V^\pi\|_\infty \\ &\leq \frac{\gamma}{(1 - \gamma)^2} \max_{s,a} \|P(\cdot | s, a) - \widehat{P}(\cdot | s, a)\|_1\end{aligned}$$

(why is the first inequality true?)

The proof for the Claim 3 immediately follows from the second claim.

## Attempt 2:

obtaining sublinear sample complexity

idea: use concentration only on  $V^\star$

# Reference sheet (defs/notation)

# Reference sheet (defs/notation)

- Remember: # samples from generative model =  $SAN$

# Reference sheet (defs/notation)

- Remember: # samples from generative model =  $SAN$
- $P^\pi$  is the transition matrix on state-action pairs for a deterministic policy  $\pi$ :

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

# Reference sheet (defs/notation)

- Remember: # samples from generative model =  $SAN$
- $P^\pi$  is the transition matrix on state-action pairs for a deterministic policy  $\pi$ :

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s') \\ 0 & \text{if } a' \neq \pi(s') \end{cases}$$

- With this notation,

$$Q^\pi = r + \gamma P V^\pi, \quad Q^\pi = r + \gamma P^\pi Q^\pi, \quad Q^\pi = (I - \gamma P^\pi)^{-1} r$$

# Reference sheet (defs/notation)

- Remember: # samples from generative model =  $SAN$

- $P^\pi$  is the transition matrix on state-action pairs for a deterministic policy  $\pi$ :

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s) \\ 0 & \text{if } a' \neq \pi(s) \end{cases}$$

- With this notation,

$$Q^\pi = r + \gamma P V^\pi, \quad Q^\pi = r + \gamma P^\pi Q^\pi, \quad Q^\pi = (I - \gamma P^\pi)^{-1} r$$

- $\frac{1}{1-\gamma}(I - \gamma P^\pi)^{-1}$  is a matrix whose rows are probability distributions (why?)

# Reference sheet (defs/notation)

- Remember: # samples from generative model =  $SAN$

- $P^\pi$  is the transition matrix on state-action pairs for a deterministic policy  $\pi$ :

$$P_{(s,a),(s',a')}^\pi := \begin{cases} P(s' | s, a) & \text{if } a' = \pi(s) \\ 0 & \text{if } a' \neq \pi(s) \end{cases}$$

- With this notation,

$$Q^\pi = r + \gamma P V^\pi, \quad Q^\pi = r + \gamma P^\pi Q^\pi, \quad Q^\pi = (I - \gamma P^\pi)^{-1} r$$

- $\frac{1}{1-\gamma}(I - \gamma P^\pi)^{-1}$  is a matrix whose rows are probability distributions (why?)

- $\widehat{Q}^*$ : optimal value in estimated model  $\widehat{M}$ .
- $\widehat{\pi}^*$ : optimal policy in  $\widehat{M}$ .
- $Q^{\widehat{\pi}^*}$ : (true) value of estimated policy.

# Attempt 2: Sublinear Sample Complexity

# Attempt 2: Sublinear Sample Complexity

**Proposition:** (Crude Value Bound) With probability greater than  $1 - \delta$ ,

$$\|Q^* - \widehat{Q}^*\|_\infty \leq \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{2 \log(2SA/\delta)}{N}}$$

# samples

$$\|Q^* - \widehat{Q}^{\pi^*}\|_\infty \leq \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{2 \log(2SA/\delta)}{N}}$$

SA/N

# Attempt 2: Sublinear Sample Complexity

$Q^* = Q \Rightarrow^*$

**Proposition:** (Crude Value Bound) With probability greater than  $1 - \delta$ ,

$$\|Q^* - \widehat{Q}^*\|_\infty \leq \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{2 \log(2SA/\delta)}{N}}$$

$Q^* = Q \Rightarrow^*$

$$\|Q^* - \widehat{Q}^{\pi^*}\|_\infty \leq \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{2 \log(2SA/\delta)}{N}}$$

What about the value of the policy?

$$\|Q^* - Q^{\widehat{\pi}^*}\|_\infty \leq \frac{\gamma}{(1-\gamma)^3} \sqrt{\frac{2 \log(2SA/\delta)}{N}}$$

# Sample Size Corollaries

**Corollary:** for  $\epsilon < 1$ , provided  $N \geq \frac{c}{(1-\gamma)^4} \frac{\log(cSA/\delta)}{\epsilon^2}$  then

$$\|Q^* - \widehat{Q}^*\|_\infty \leq \epsilon \text{ (with prob. greater than } 1 - \delta)$$

# Sample Size Corollaries

**Corollary:** for  $\epsilon < 1$ , provided  $N \geq \frac{c}{(1-\gamma)^4} \frac{\log(cSA/\delta)}{\epsilon^2}$  then  
 $\|Q^* - \widehat{Q}^*\|_\infty \leq \epsilon$  (with prob. greater than  $1 - \delta$ )

What about the policy?

**Corollary:** for  $\epsilon < 1$ , provided  $N \geq \frac{c}{(1-\gamma)^6} \frac{\log(cSA/\delta)}{\epsilon^2}$  then  
 $\|Q^* - Q^{\widehat{\pi}^*}\|_\infty \leq \epsilon$  (with prob. greater than  $1 - \delta$ )

# Component-wise Bounds Lemma

**Lemma:** we have that

$$Q^* - \widehat{Q}^* \leq \gamma(I - \gamma \widehat{P}^{\pi^*})^{-1}(P - \widehat{P})V^*$$

$$Q^* - \widehat{Q}^* \geq \gamma(I - \gamma \widehat{P}^{\widehat{\pi}^*})^{-1}(P - \widehat{P})V^*$$

$$\widehat{P} V^*$$

# Component-wise Bounds Lemma

$$E_{s \sim P(\cdot|s, a)} [V^*(s)] - E_{s \sim \hat{P}(\cdot|s, a)} [V^*(s)]$$

**Lemma:** we have that

$$Q^* - \hat{Q}^* \leq \gamma(I - \gamma \hat{P}^{\pi^*})^{-1}(P - \hat{P})V^*$$

$$Q^* - \hat{Q}^* \geq \gamma(I - \gamma \hat{P}^{\hat{\pi}^*})^{-1}(P - \hat{P})V^*$$

**Proof:**

For the first claim, the optimality of  $\pi^*$  in  $M$  implies:

$$Q^* - \hat{Q}^* = Q^{\pi^*} - \hat{Q}^{\hat{\pi}^*} \leq Q^{\pi^*} - \hat{Q}^{\pi^*} = \gamma(I - \gamma \hat{P}^{\pi^*})^{-1}(P - \hat{P})V^*,$$

using the simulation lemma in the final step.

See notes for the proof of second claim.

Proof: (& key idea for sublinearity)

# Proof: (& key idea for sublinearity)

- Proof of the first claim:

# Proof: (& key idea for sublinearity)

- Proof of the first claim:

- By comp. lemma:  $\|Q^* - \widehat{Q}^*\|_\infty \leq \frac{\gamma}{1-\gamma} \|(P - \widehat{P})V^*\|_\infty$

# Proof: (& key idea for sublinearity)

- Proof of the first claim:

- By comp. lemma:  $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma} \|(P - \widehat{P})V^\star\|_\infty$

- Recall  $\|V^\star\|_\infty \leq 1/(1-\gamma)$ .

# Proof: (& key idea for sublinearity)

- Proof of the first claim:

- By comp. lemma:  $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma} \|(P - \widehat{P})V^\star\|_\infty$

- Recall  $\|V^\star\|_\infty \leq 1/(1-\gamma)$ .

- By Hoeffding's inequality and the union bound,

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

which holds with probability greater than  $1 - \delta$ .

# Proof: (& key idea for sublinearity)

- Proof of the first claim:

- By comp. lemma:  $\|Q^* - \widehat{Q}^*\|_\infty \leq \frac{\gamma}{1-\gamma} \|(P - \widehat{P})V^*\|_\infty$

- Recall  $\|V^*\|_\infty \leq 1/(1-\gamma)$ .

- By Hoeffding's inequality and the union bound,

$$\|(P - \widehat{P})V^*\|_\infty = \max_{s,a} \left| E_{s' \sim P(\cdot|s,a)}[V^*(s')] - E_{s' \sim \widehat{P}(\cdot|s,a)}[V^*(s')] \right|$$

$$\leq \frac{1}{1-\gamma} \sqrt{\frac{2 \log(2SA/\delta)}{N}}$$

which holds with probability greater than  $1 - \delta$ .

- Proof of second claim is similar (see the book)

Handwritten notes in blue ink:

- $\frac{\text{Var}(f)}{\sqrt{N}}$
- $E_{x \sim P}(f) - E[f] \leq \frac{\text{Bound}(f)}{\sqrt{N}}$
- $\frac{\text{Var}(V^*)}{\sqrt{N}}$

# Attempt 3:

minimax optimal sample complexity

idea: better variance control

(“near”) Minimax Optimal Sample Complexity

# (“near”) Minimax Optimal Sample Complexity

**Theorem:** With probability greater than  $1 - \delta$ ,

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

where  $c$  is an absolute constant.

# Minimax Optimal Sample Complexity

# Minimax Optimal Sample Complexity

**Corollary:** for  $\epsilon < 1$ , provided  $N \geq \frac{c}{(1-\gamma)^3} \frac{\log(cSA/\delta)}{\epsilon^2}$  then

$$\|Q^* - \widehat{Q}^*\|_\infty \leq \epsilon \text{ (with prob. greater than } 1 - \delta)$$

# Minimax Optimal Sample Complexity

**Corollary:** for  $\epsilon < 1$ , provided  $N \geq \frac{c}{(1-\gamma)^3} \frac{\log(cSA/\delta)}{\epsilon^2}$  then

$$\|Q^* - \widehat{Q}^*\|_\infty \leq \epsilon \text{ (with prob. greater than } 1 - \delta)$$

What about the policy?

Naively, we need  $N/(1-\gamma)^2$  more samples.

**A different thm gives:** With the same  $N$ ,

$$\|Q^* - Q^{\widehat{\pi}^*}\|_\infty \leq \epsilon \text{ (with prob. greater than } 1 - \delta)$$

# Minimax Optimal Sample Complexity

**Corollary:** for  $\epsilon < 1$ , provided  $N \geq \frac{c}{(1-\gamma)^3} \frac{\log(cSA/\delta)}{\epsilon^2}$  then

$$\|Q^* - \widehat{Q}^*\|_\infty \leq \epsilon \text{ (with prob. greater than } 1 - \delta)$$

What about the policy?

Naively, we need  $N/(1-\gamma)^2$  more samples.

**A different thm gives:** With the same  $N$ ,

$$\|Q^* - Q^{\widehat{\pi}^*}\|_\infty \leq \epsilon \text{ (with prob. greater than } 1 - \delta)$$

**Lower Bound:** We can't do better.

## Revisiting proof attempt 2: where is there slop?

- **Proof of the first claim:**

- By comp. lemma:  $\|Q^\star - \widehat{Q}^\star\|_\infty \leq \frac{\gamma}{1-\gamma} \|(P - \widehat{P})V^\star\|_\infty$

- Recall  $\|V^\star\|_\infty \leq 1/(1-\gamma)$ .

- By Hoeffding's inequality and the union bound,

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

which holds with probability greater than  $1 - \delta$ .

- Proof of second claim is similar (see the book)

# Proof sketch: part 1

- From Bernstein's ineq, with pr. greater than  $1 - \delta$ , we have (component-wise):

$$|(P - \widehat{P})V^*| \leq \sqrt{\frac{2 \log(2SA/\delta)}{N}} \sqrt{\text{Var}_P(V^*)} + \frac{1}{1 - \gamma} \frac{2 \log(2SA/\delta)}{3N} \vec{1}$$

# Proof sketch: part 1

- From Bernstein's ineq, with pr. greater than  $1 - \delta$ , we have (component-wise):

$$|(P - \widehat{P})V^*| \leq \sqrt{\frac{2 \log(2SA/\delta)}{N}} \sqrt{\text{Var}_P(V^*)} + \frac{1}{1 - \gamma} \frac{2 \log(2SA/\delta)}{3N} \vec{1}$$

- How to use this: again from “Component-wise Bounds” lemma,

$$Q^* - \widehat{Q}^* \leq \gamma \|(I - \gamma \widehat{P}^{\pi^*})^{-1} (P - \widehat{P})V^*\|_\infty \leq ??$$

# Proof sketch: part 1

- From Bernstein's ineq, with pr. greater than  $1 - \delta$ , we have (component-wise):

$$|(P - \widehat{P})V^*| \leq \sqrt{\frac{2 \log(2SA/\delta)}{N}} \sqrt{\text{Var}_P(V^*)} + \frac{1}{1 - \gamma} \frac{2 \log(2SA/\delta)}{3N} \vec{1}$$

- How to use this: again from “Component-wise Bounds” lemma,

$$Q^* - \widehat{Q}^* \leq \gamma \|(I - \gamma \widehat{P}^{\pi^*})^{-1} (P - \widehat{P})V^*\|_\infty \leq ??$$

- Therefore

$$Q^* - \widehat{Q}^* \leq \gamma \sqrt{\frac{2 \log(2SA/\delta)}{N}} \|(I - \gamma \widehat{P}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V^*)}\|_\infty$$

+ "lower order term"

# Bellman Equation for the (total) Variance

- **Variance:**  $\text{Var}_P(V)(s, a) := \text{Var}_{P(\cdot|s,a)}(V)$

Component wise variance:  $\text{Var}_P(V) := P(V)^2 - (PV)^2$

# Bellman Equation for the (total) Variance

- **Variance:**  $\text{Var}_P(V)(s, a) := \text{Var}_{P(\cdot|s,a)}(V)$

Component wise variance:  $\text{Var}_P(V) := P(V)^2 - (PV)^2$

- Let's keep around the MDP M subscripts.

Define  $\Sigma_M^\pi$  as the (total) variance of the discounted reward:

$$\Sigma_M^\pi(s, a) := E \left[ \left( \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - Q_M^\pi(s, a) \right)^2 \middle| s_0 = s, a_0 = a \right]$$

# Bellman Equation for the (total) Variance

- **Variance:**  $\text{Var}_P(V)(s, a) := \text{Var}_{P(\cdot|s,a)}(V)$

Component wise variance:  $\text{Var}_P(V) := P(V)^2 - (PV)^2$

- Let's keep around the MDP M subscripts.

Define  $\Sigma_M^\pi$  as the (total) variance of the discounted reward:

$$\Sigma_M^\pi(s, a) := E \left[ \left( \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) - Q_M^\pi(s, a) \right)^2 \middle| s_0 = s, a_0 = a \right]$$

- **Bellman equation for the total variance:**

$$\Sigma_M^\pi = \gamma^2 \text{Var}_P(V_M^\pi) + \gamma^2 P^\pi \Sigma_M^\pi$$

# Key Lemma

**Lemma:** For any policy  $\pi$  and MDP  $M$ ,

$$\left\| (I - \gamma P^\pi)^{-1} \sqrt{\text{Var}_P(V_M^\pi)} \right\|_\infty \leq \sqrt{\frac{2}{(1 - \gamma)^3}}$$

Proof idea: convexity + Bellman equations for the variance.

# Putting it all together

**Proof sketch:** we have two MDPs  $M$  and  $\widehat{M}$ . need to bound:

# Putting it all together

**Proof sketch:** we have two MDPs  $M$  and  $\widehat{M}$ . need to bound:

$$\|(I - \gamma \widehat{P}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V^*)}\|_\infty = \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_M^{\pi^*})}\|_\infty$$

$$\leq \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_{\widehat{M}}^{\pi^*})}\| + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1 - \gamma)^3}} + \text{"lower order"}$$

# Putting it all together

**Proof sketch:** we have two MDPs  $M$  and  $\widehat{M}$ . need to bound:

$$\|(I - \gamma \widehat{P}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V^*)}\|_\infty = \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_M^{\pi^*})}\|_\infty$$

$$\leq \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_{\widehat{M}}^{\pi^*})}\| + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1 - \gamma)^3}} + \text{"lower order"}$$

First equality above: just notation

# Putting it all together

**Proof sketch:** we have two MDPs  $M$  and  $\widehat{M}$ . need to bound:

$$\|(I - \gamma \widehat{P}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V^*)}\|_\infty = \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_M^{\pi^*})}\|_\infty$$

$$\leq \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_{\widehat{M}}^{\pi^*})}\| + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1 - \gamma)^3}} + \text{"lower order"}$$

First equality above: just notation

Second step: concentration  $\rightarrow$  we need to quantify:

$$\sqrt{\text{Var}_P(V_M^{\pi^*})} \approx \sqrt{\text{Var}_P(V_{\widehat{M}}^{\pi^*})}$$

# Putting it all together

**Proof sketch:** we have two MDPs  $M$  and  $\widehat{M}$ . need to bound:

$$\|(I - \gamma \widehat{P}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V^*)}\|_\infty = \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_M^{\pi^*})}\|_\infty$$

$$\leq \|(I - \gamma P_{\widehat{M}}^{\pi^*})^{-1} \sqrt{\text{Var}_P(V_{\widehat{M}}^{\pi^*})}\| + \text{"lower order"}$$

$$\leq \sqrt{\frac{2}{(1 - \gamma)^3}} + \text{"lower order"}$$

First equality above: just notation

Second step: concentration  $\rightarrow$  we need to quantify:

$$\sqrt{\text{Var}_P(V_M^{\pi^*})} \approx \sqrt{\text{Var}_P(V_{\widehat{M}}^{\pi^*})}$$

Last step: previous slide