

CSE 542: Statistical Reinforcement Learning

Lecture 2: Sample Complexity with a Generative Model

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Today's Plan

- 1 The Generative Model
- 2 Naive Model-Based Approach
- 3 Sublinear Sample Complexity
- 4 Minimax Optimal Results
- 5 Finite-Horizon Setting

From Planning to Learning

So far: we assumed $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is **known** and studied planning.

Now: P is **unknown**. We observe samples and must learn to act near-optimally.

Generative Model (Simulator Oracle)

At any time, we may query *any* state-action pair (s, a) and receive:

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

Queries are independent — no sequential constraints on what we ask.

This is the most **favorable** data collection model: we have uniform access to all of P . It will serve as a clean baseline for understanding what is statistically possible.

The Central Question

Sample Complexity Question

How many simulator queries are needed to output a policy $\hat{\pi}$ satisfying

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

with probability at least $1 - \delta$?

Naive guess: P has $|\mathcal{S}|^2|\mathcal{A}|$ parameters. Estimating each requires $\sim |\mathcal{S}|$ samples, suggesting $|\mathcal{S}|^2|\mathcal{A}|$ total samples.

The Central Question

Sample Complexity Question

How many simulator queries are needed to output a policy $\hat{\pi}$ satisfying

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

with probability at least $1 - \delta$?

Naive guess: P has $|\mathcal{S}|^2|\mathcal{A}|$ parameters. Estimating each requires $\sim |\mathcal{S}|$ samples, suggesting $|\mathcal{S}|^2|\mathcal{A}|$ total samples.

Main message of this lecture

The true sample complexity scales like

$$\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2} \cdot \text{poly}\left(\frac{1}{1-\gamma}\right),$$

which is **sublinear** in the number of parameters of P .

Today's Plan

- 1 The Generative Model
- 2 Naive Model-Based Approach**
- 3 Sublinear Sample Complexity
- 4 Minimax Optimal Results
- 5 Finite-Horizon Setting

Algorithm: Estimate P , Then Plan

Model-Based Algorithm

- 1 For each (s, a) : query the simulator N times, record $s'_1, \dots, s'_N \sim P(\cdot | s, a)$.
- 2 Build the empirical transition kernel:

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

- 3 Form empirical MDP $\hat{\mathcal{M}}$: same r and γ , transitions \hat{P} .

Though $\hat{P}(s' | s, a)$ is empirical, it is a valid probability kernel:

- For any π define $\hat{Q}^\pi = (I - \gamma \hat{P}^\pi)^{-1} r$.
- $\hat{Q}^* = \max_\pi \hat{Q}^\pi$
- $\hat{\pi}^* \in \arg \max_\pi \hat{Q}^\pi$

$$V^*(s) = \sup_\pi V^\pi(s)$$

$$Q^*(s, a) = \max_\pi Q^\pi(s, a)$$

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a \right]$$

Two Key Lemmas

Our goal is to learn a policy $\hat{\pi}$ such that $\|Q^* - Q^{\hat{\pi}}\|_{\infty} \leq \epsilon$.

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

Lemma 2.2 (Simulation Lemma)

For any stationary policy π :

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

$$\begin{aligned} Q^* - Q^{\hat{\pi}} &= Q^{\pi_*} - Q^{\hat{\pi}} \\ &= Q^{\pi_*} - \hat{Q}^{\pi_*} + \hat{Q}^{\pi_*} - \hat{Q}^{\hat{\pi}} + \hat{Q}^{\hat{\pi}} - Q^{\hat{\pi}} \\ &\leq \max_{\pi} |Q^{\pi} - \hat{Q}^{\pi}| \cdot 2 \end{aligned}$$

(If $\hat{\pi} = \arg\max_{\pi} \hat{Q}^{\pi}$)

↑
element-wise

Simulation Lemma: Proof

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

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

Finally, $(P^\pi Q^\pi)(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)} [V^\pi(s')]$, so $(P^\pi - \hat{P}^\pi) Q^\pi = (P - \hat{P}) V^\pi$. \square

Two Key Lemmas

Our goal is to learn a policy $\hat{\pi}$ such that $\|Q^* - Q^{\hat{\pi}}\|_{\infty} \leq \varepsilon$.

Lemma 2.2 (Simulation Lemma)

For any stationary policy π :

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

Lemma 2.3 (Norm Bound)

For any stationary policy π and vector $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$\| \underbrace{(I - \gamma P^{\pi})}^{-1} v \|_{\infty} \leq \frac{1}{1 - \gamma} \|v\|_{\infty}.$$

$$u = (I - \gamma P^{\pi})^{-1} v$$

$$v = (I - \gamma P^{\pi}) u \quad \|v\|_{\infty} \geq \|u\|_{\infty} - \gamma \|P^{\pi} u\|_{\infty} \geq (1 - \gamma) \|u\|_{\infty}$$

$$\|u\|_{\infty} \leq \frac{\|v\|_{\infty}}{1 - \gamma}$$

Two Key Lemmas

Our goal is to learn a policy $\hat{\pi}$ such that $\|Q^* - Q^{\hat{\pi}}\|_{\infty} \leq \varepsilon$.

Lemma 2.2 (Simulation Lemma)

For any stationary policy π :

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

Lemma 2.3 (Norm Bound)

For any stationary policy π and vector $v \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$\|(I - \gamma P^{\pi})^{-1}v\|_{\infty} \leq \frac{1}{1 - \gamma} \|v\|_{\infty}.$$

Combining:
$$\|Q^{\pi} - \hat{Q}^{\pi}\|_{\infty} \leq \frac{\gamma}{1 - \gamma} \|(P - \hat{P})V^{\pi}\|_{\infty}.$$

Hoeffding's Inequality

The central concentration inequality we will use throughout.

Theorem (Hoeffding's Inequality)

Let X_1, \dots, X_N be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \leq X_i \leq b$ almost surely. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\left| \frac{1}{N} \sum_{i=1}^N X_i - \mu \right| \leq (b - a) \sqrt{\frac{\log(2/\delta)}{2N}}.$$

$X_i = \mathbb{1} \{ \text{Fair coin } i \text{ comes up heads} \}$

$$\left| \frac{1}{N} \sum_{i=1}^N X_i - \frac{1}{2} \right| \leq \sqrt{\frac{\log(2/\delta)}{2N}} \lesssim \sqrt{\frac{1}{N}} \quad \delta = .025$$

Hoeffding's Inequality

The central concentration inequality we will use throughout.

Theorem (Hoeffding's Inequality)

Let X_1, \dots, X_N be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $a \leq X_i \leq b$ almost surely. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$:

$$\left| \frac{1}{N} \sum_{i=1}^N X_i - \mu \right| \leq (b - a) \sqrt{\frac{\log(2/\delta)}{2N}}.$$

Key features:

$$\mathbb{P}\left(\left|\frac{1}{N} \sum_{i=1}^N X_i - \mu\right| > (b-a) \sqrt{\frac{\log(2/\delta)}{2N}}\right) \leq \delta$$

- Deviation scales as $\frac{b-a}{\sqrt{N}}$ — the *range* of X_i divided by \sqrt{N} .
- No assumption on the distribution beyond its support.
- Adding a union bound over K events costs a $\log K$ factor: replace $\log(2/\delta)$ with $\log(2K/\delta)$.



$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

Naive Bound: Main Result

Proposition 2.1 (Naive model-based approach)

With $N \gtrsim \frac{|\mathcal{S}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{(1-\gamma)^4 \varepsilon^2}$ samples per (s, a) , so that

$$\text{total samples} \asymp \frac{|\mathcal{S}|^2 |\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta},$$

with probability $\geq 1 - \delta$:

- $\max_{s,a} \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 \leq (1-\gamma)^2 \varepsilon$
- $\left\| Q^\pi - \hat{Q}^\pi \right\|_\infty \leq \varepsilon$ for every policy π simultaneously
- $\left\| \hat{Q}^* - Q^* \right\|_\infty \leq \varepsilon$ and $\left\| Q^{\hat{\pi}^*} - Q^* \right\|_\infty \leq 2\varepsilon$

Proof of Proposition 2.1 — Step 1: Model Accuracy

Fix (s, a) , each $s'_i \sim P(\cdot | s, a)$ for $i = 1, \dots, N$ is an i.i.d. draw from a discrete distribution over $|S|$ values. We apply Hoeffding's inequality:

$$\max_{s,a} \left\| P(\cdot | s, a) - \hat{P}(\cdot | s, a) \right\|_1 = \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \langle z, P(\cdot | s, a) - \hat{P}(\cdot | s, a) \rangle$$

$$x \in \mathbb{R}^d$$

$$\begin{aligned} & \max_{z \in \{-1,1\}^d} \langle z, x \rangle \\ &= \max_{z \in \{-1,1\}^d} \sum_{i=1}^d z_i x_i \\ &= \sum_{i=1}^d \max\{x_i, -x_i\} \\ &= \sum_{i=1}^d |x_i| = \|x\|_1 \end{aligned}$$

Proof of Proposition 2.1 — Step 1: Model Accuracy

Fix (s, a) , each $s'_i \sim P(\cdot|s, a)$ for $i = 1, \dots, N$ is an i.i.d. draw from a discrete distribution over $|\mathcal{S}|$ values. We apply Hoeffding's inequality:

$$\begin{aligned} \max_{s,a} \left\| P(\cdot|s, a) - \widehat{P}(\cdot|s, a) \right\|_1 &= \max_{s,a} \max_{z \in \{-1,1\}^{|\mathcal{S}|}} \langle z, P(\cdot|s, a) - \widehat{P}(\cdot|s, a) \rangle \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|\mathcal{S}|}} \sum_{s'} z_{s'} \frac{1}{N} \sum_{i=1}^N (P(s'|s, a) - \mathbf{1}\{s'_i = s'\}) \end{aligned}$$

Proof of Proposition 2.1 — Step 1: Model Accuracy

Fix (s, a) , each $s'_i \sim P(\cdot|s, a)$ for $i = 1, \dots, N$ is an i.i.d. draw from a discrete distribution over $|S|$ values. We apply Hoeffding's inequality:

$$\begin{aligned} \max_{s,a} \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 &= \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \langle z, P(\cdot|s, a) - \hat{P}(\cdot|s, a) \rangle \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \sum_{s'} z_{s'} \frac{1}{N} \sum_{i=1}^N (P(s'|s, a) - \mathbf{1}\{s'_i = s'\}) \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \frac{1}{N} \sum_{i=1}^N (\langle z, P(\cdot|s, a) \rangle - z_{s'_i}) \\ &\leq 2 \sqrt{\frac{\log(2/d)}{2N}} \end{aligned}$$

Proof of Proposition 2.1 — Step 1: Model Accuracy

Fix (s, a) , each $s'_i \sim P(\cdot|s, a)$ for $i = 1, \dots, N$ is an i.i.d. draw from a discrete distribution over $|S|$ values. We apply Hoeffding's inequality:

$$\begin{aligned} \max_{s,a} \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 &= \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \langle z, P(\cdot|s, a) - \hat{P}(\cdot|s, a) \rangle \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \sum_{s'} z_{s'} \frac{1}{N} \sum_{i=1}^N (P(s'|s, a) - \mathbf{1}\{s'_i = s'\}) \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|S|}} \frac{1}{N} \sum_{i=1}^N (\langle z, P(\cdot|s, a) \rangle - z_{s'_i}) \\ &\leq \sqrt{\frac{2 \log(|S| |\mathcal{A}| 2^{|S|} \cdot 2/\delta)}{N}} \lesssim \sqrt{\frac{|S| \log(|S| |\mathcal{A}|/\delta)}{N}}. \end{aligned}$$

$$\sqrt{\frac{|S| \log(2) + \log(|S| |\mathcal{A}|/\delta)}{N}}$$

Proof of Proposition 2.1 — Step 1: Model Accuracy


Fix (s, a) , each $s'_i \sim P(\cdot|s, a)$ for $i = 1, \dots, N$ is an i.i.d. draw from a discrete distribution over $|\mathcal{S}|$ values. We apply Hoeffding's inequality:

$$\begin{aligned} \max_{s,a} \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 &= \max_{s,a} \max_{z \in \{-1,1\}^{|\mathcal{S}|}} \langle z, P(\cdot|s, a) - \hat{P}(\cdot|s, a) \rangle \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|\mathcal{S}|}} \sum_{s'} z_{s'} \frac{1}{N} \sum_{i=1}^N (P(s'_i|s, a) - \mathbf{1}\{s'_i = s'\}) \\ &= \max_{s,a} \max_{z \in \{-1,1\}^{|\mathcal{S}|}} \frac{1}{N} \sum_{i=1}^N (\langle z, P(\cdot|s, a) \rangle - z_{s'_i}) \\ &\leq \sqrt{\frac{2 \log(|\mathcal{S}||\mathcal{A}|2^{|\mathcal{S}|} \cdot 2/\delta)}{N}} \lesssim \sqrt{\frac{|\mathcal{S}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{N}}. \end{aligned}$$

Setting this $\leq (1 - \gamma)^2 \varepsilon$ requires

$$N \gtrsim \frac{|\mathcal{S}| \log(|\mathcal{S}||\mathcal{A}|/\delta)}{(1 - \gamma)^4 \varepsilon^2},$$

which gives the stated total sample count.

(continued \rightarrow) 

Proof of Proposition 2.1 — Step 2: Value Accuracy

Assume Step 1 holds: $\max_{s,a} \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 \leq (1 - \gamma)^2 \varepsilon$.

Fix any π . By Lemmas 2.2 and 2.3: $\left\| Q^\pi - \hat{Q}^\pi \right\|_\infty \leq \frac{\gamma}{1 - \gamma} \left\| (P - \hat{P})V^\pi \right\|_\infty$.

For each (s, a) , the one-step error satisfies:

$$\left| (P - \hat{P})V^\pi(s, a) \right| \leq \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 \cdot \|V^\pi\|_\infty \leq (1 - \gamma)^2 \varepsilon \cdot \frac{1}{1 - \gamma} = (1 - \gamma)\varepsilon.$$

$$\begin{aligned} \langle x, y \rangle &= \sum_{i=1}^n x_i y_i \\ &\leq \sum_{i=1}^n |x_i| \cdot |y_i| \\ &\leq \max_{i=1, \dots, n} |x_i| \cdot \sum_{i=1}^n |y_i| \\ &:= \|x\|_\infty \cdot \|y\|_1 \end{aligned}$$

Proof of Proposition 2.1 — Step 2: Value Accuracy

Assume Step 1 holds: $\max_{s,a} \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 \leq (1 - \gamma)^2 \varepsilon$.

Fix any π . By Lemmas 2.2 and 2.3: $\left\| Q^\pi - \hat{Q}^\pi \right\|_\infty \leq \frac{\gamma}{1 - \gamma} \left\| (P - \hat{P})V^\pi \right\|_\infty$.

For each (s, a) , the one-step error satisfies:

$$\left| (P - \hat{P})V^\pi(s, a) \right| \leq \left\| P(\cdot|s, a) - \hat{P}(\cdot|s, a) \right\|_1 \cdot \|V^\pi\|_\infty \leq (1 - \gamma)^2 \varepsilon \cdot \frac{1}{1 - \gamma} = (1 - \gamma)\varepsilon.$$

Therefore $\left\| Q^\pi - \hat{Q}^\pi \right\|_\infty \leq \frac{\gamma}{1 - \gamma} \cdot (1 - \gamma)\varepsilon = \gamma\varepsilon \leq \varepsilon$.

Since this holds for every π , it holds simultaneously for all π .

(continued →)

Proof of Proposition 2.1 — Step 3: Near-Optimal Planning

Step 2 gives $\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \varepsilon$ for every π .

Bounding $\|\widehat{Q}^* - Q^*\|_\infty$.

Using $|\sup_\pi f(\pi) - \sup_\pi g(\pi)| \leq \sup_\pi |f(\pi) - g(\pi)|$:

$$\underbrace{|\widehat{Q}^*(s, a) - Q^*(s, a)|} \leq \sup_\pi \underbrace{|\widehat{Q}^\pi(s, a) - Q^\pi(s, a)|} \leq \varepsilon.$$

Proof of Proposition 2.1 — Step 3: Near-Optimal Planning

Step 2 gives $\|Q^\pi - \widehat{Q}^\pi\|_\infty \leq \varepsilon$ for every π .

Bounding $\|\widehat{Q}^* - Q^*\|_\infty$.

Using $|\sup_\pi f(\pi) - \sup_\pi g(\pi)| \leq \sup_\pi |f(\pi) - g(\pi)|$:

$$|\widehat{Q}^*(s, a) - Q^*(s, a)| \leq \sup_\pi |\widehat{Q}^\pi(s, a) - Q^\pi(s, a)| \leq \varepsilon.$$

Bounding $\|Q^{\widehat{\pi}^*} - Q^*\|_\infty$.

Since $\widehat{\pi}^*$ is optimal in $\widehat{\mathcal{M}}$, $\widehat{Q}^{\widehat{\pi}^*} = \widehat{Q}^*$. By triangle inequality and Step 2:

$$\|Q^{\widehat{\pi}^*} - Q^*\|_\infty \leq \|Q^{\widehat{\pi}^*} - \widehat{Q}^*\|_\infty + \|\widehat{Q}^* - Q^*\|_\infty \leq \varepsilon + \varepsilon = 2\varepsilon.$$

□

What the Warmup Achieves — and What It Costs

Proposition 2.1 gives a **very** strong guarantee:

- **Uniform** value accuracy: $\widehat{Q}^\pi \approx Q^\pi$ for every policy π .
- Near-optimal planning follows as a corollary.

The cost of uniformity

Uniform accuracy requires estimating each $P(\cdot|s, a)$ accurately in ℓ_1 . Each such distribution has $|\mathcal{S}|$ free parameters, so we need $N \gtrsim |\mathcal{S}|$ samples per pair — giving $|\mathcal{S}|^2|\mathcal{A}|$ total samples.

Key observation: For planning we only need a single near-optimal policy. We do not need $\widehat{Q}^\pi \approx Q^\pi$ for every π — just for π^* . Can we exploit this to reduce the sample count?

Today's Plan

- 1 The Generative Model
- 2 Naive Model-Based Approach
- 3 Sublinear Sample Complexity**
- 4 Minimax Optimal Results
- 5 Finite-Horizon Setting

Targeting Q^* Directly

Rather than estimating Q^π for all π , we focus on a single object: Q^* .

Lemma 2.5 (Planning-Focused Comparison)

Let π^* be optimal in M and $\hat{\pi}^*$ optimal in \hat{M} . Then componentwise:

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

Why this matters: both bounds involve $(P - \hat{P})V^*$, where V^* is the *fixed* optimal value function of the true MDP — independent of the data.

This is essential: applying Hoeffding to $(P - \hat{P})V^{\hat{\pi}^*}$ would be **incorrect** because $V^{\hat{\pi}^*}$ is *random* (it depends on the same samples used to form \hat{P}).

Sublinear Sample Complexity: Main Result

Proposition 2.4 (Crude value bound)

Fix $\delta \in (0, 1)$. With N samples per (s, a) , with probability $\geq 1 - \delta$:

$$\|Q^* - \hat{Q}^*\|_\infty \leq \frac{\gamma}{(1 - \gamma)^2} \sqrt{\frac{2 \log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}.$$

Sample complexity to achieve $\|Q^* - \hat{Q}^*\|_\infty \leq \varepsilon$:

$$\text{total samples} \asymp \frac{|\mathcal{S}||\mathcal{A}|}{(1 - \gamma)^4 \varepsilon^2} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta}.$$

- $|\mathcal{S}||\mathcal{A}|$ not $|\mathcal{S}|^2|\mathcal{A}|$ — a factor of $|\mathcal{S}|$ saved over the naive approach.
- Same $(1 - \gamma)^{-4}$ horizon dependence — we will improve this next.

Proof of Proposition 2.4

By Lemma 2.5 and Lemma 2.3:

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

Proof of Proposition 2.4

By Lemma 2.5 and Lemma 2.3:

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

Bounding $\|(P - \hat{P})V^*\|_\infty$. Fix (s, a) . The quantity

$$(P - \hat{P})V^*(s, a) = \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^*(s')] - \frac{1}{N} \sum_{i=1}^N V^*(s'_i)$$

is the deviation between the true mean and sample mean of $V^*(s')$, where V^* is **fixed**.

Since $0 \leq V^* \leq \frac{1}{1-\gamma}$, **Hoeffding's inequality** + union bound over $|\mathcal{S}||\mathcal{A}|$ pairs gives, with prob. $\geq 1 - \delta$:

$$\|(P - \hat{P})V^*\|_\infty \leq \frac{1}{1-\gamma} \sqrt{\frac{2 \log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N}}.$$

Substituting yields the stated bound.

From Q^* Accuracy to Policy Performance

We have $\left\| \widehat{Q}^* - Q^* \right\|_{\infty} \leq \Delta$ where $\Delta = \frac{\gamma}{(1-\gamma)^2} \sqrt{\frac{2 \log(\cdot)}{N}}$, and $\widehat{\pi}^*$ is greedy w.r.t. \widehat{Q}^* :
 $\widehat{\pi}^*(s) = \arg \max_a \widehat{Q}^*(s, a)$.

Recall Lemma 1.11 (Q-error amplification, Lecture 1):

$$V^* - V^{\widehat{\pi}^*} \leq \frac{2 \left\| \widehat{Q}^* - Q^* \right\|_{\infty}}{1-\gamma} \cdot \mathbf{1}.$$

Substituting:

$$\left\| V^* - V^{\widehat{\pi}^*} \right\|_{\infty} \leq \frac{2\Delta}{1-\gamma} = \frac{2\gamma}{(1-\gamma)^3} \sqrt{\frac{2 \log(\cdot)}{N}}.$$

To achieve an ε -optimal policy ($\frac{2\Delta}{1-\gamma} \leq \varepsilon$):

$$\text{total samples} \asymp \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^6 \varepsilon^2} \log(\cdot).$$

Horizon amplification

Going from Q^* accuracy to policy performance costs an *extra* $(1-\gamma)^{-2}$ via Lemma 1.11. This can be avoided with a sharper argument (Theorem 2.6).

Today's Plan

- 1 The Generative Model
- 2 Naive Model-Based Approach
- 3 Sublinear Sample Complexity
- 4 Minimax Optimal Results**
- 5 Finite-Horizon Setting

Minimax Upper Bound

The crude bound uses **Hoeffding**, which accounts only for the *range* of V^* . A sharper **variance-sensitive** (Bernstein) analysis improves the horizon dependence.

Theorem 2.6 (Minimax upper bound, discounted case)

With

$$\text{total samples} \gtrsim \frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \varepsilon^2} \log \frac{|\mathcal{S}||\mathcal{A}|}{\delta},$$

the model-based estimator achieves both $\left\| \hat{Q}^* - Q^* \right\|_{\infty} \leq \varepsilon$ and $\left\| Q^* - Q^{\hat{\pi}^*} \right\|_{\infty} \leq \varepsilon$ with probability $\geq 1 - \delta$.

Two improvements over Proposition 2.4:

- $(1-\gamma)^{-3}$ vs. $(1-\gamma)^{-4}$: exploiting that $V^*(s')$ has variance $\lesssim \left(\frac{\gamma}{1-\gamma}\right)^2 \ll \left(\frac{1}{1-\gamma}\right)^2$.
- Policy performance matches Q^* accuracy with *no* extra horizon factor — the amplification is avoided.

Matching Lower Bound

The model-based rate is **minimax optimal**.

Theorem 2.7 (Minimax lower bound, discounted case)

There exists a family of MDPs such that any algorithm achieving $\|\hat{Q}^* - Q^*\|_\infty \leq \varepsilon$ with probability $\geq 1 - \delta$ on every MDP in the family must use at least

$$\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3 \varepsilon^2} \log \frac{1}{\delta}\right)$$

samples.

Conclusion:

- The plug-in model-based approach is **minimax optimal**.
- No algorithm can do better in $|\mathcal{S}||\mathcal{A}|$, ε , or $(1 - \gamma)$ (up to log factors).

Sample complexity with a generative model

To find a policy $\hat{\pi}$ with $\|Q^* - Q^{\hat{\pi}}\|_{\infty} \leq \varepsilon$ (with high probability):

$$\text{total samples} \asymp \frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2} \cdot \text{poly}\left(\frac{1}{1-\gamma}\right)$$

This bound is **tight**: no algorithm can do better (up to the precise polynomial in the horizon).

- **Sublinear:** $|\mathcal{S}||\mathcal{A}|$, not $|\mathcal{S}|^2|\mathcal{A}|$ — we never need to learn P globally.
- **Horizon:** polynomial in $\frac{1}{1-\gamma}$, reflecting error compounding over time.
- **Algorithm:** estimate \hat{P} , then plan. Simple and optimal.

Today's Plan

- 1 The Generative Model
- 2 Naive Model-Based Approach
- 3 Sublinear Sample Complexity
- 4 Minimax Optimal Results
- 5 Finite-Horizon Setting**

Finite-Horizon MDP: Setup and Result

Recall the finite-horizon MDP: stages $h = 0, \dots, H - 1$, transition kernels $\{P_h\}$, rewards $\{r_h\}$.

Generative model: query any triple (s, a, h) to get $s' \sim P_h(\cdot | s, a)$.

Plug-in estimator: draw N samples at each (s, a, h) to form \hat{P}_h .

Total samples: $H|\mathcal{S}||\mathcal{A}| \cdot N$.

Theorem 2.8 (Finite-horizon minimax upper bound)

To achieve $\left\| Q_0^* - \hat{Q}_0^* \right\|_{\infty} \leq \varepsilon$ with probability $\geq 1 - \delta$, it suffices to take

$$\text{total samples} \asymp \frac{H^3 |\mathcal{S}| |\mathcal{A}|}{\varepsilon^2} \log \frac{|\mathcal{S}| |\mathcal{A}|}{\delta}.$$

This rate is minimax optimal (up to logarithmic factors).

Under the rough correspondence $H \approx \frac{1}{1-\gamma}$, this matches the discounted bound.

Finite-Horizon: From \widehat{Q}^* to Policy Performance

$\widehat{\pi}^*$ takes action $\widehat{\pi}_h^*(s) = \arg \max_a \widehat{Q}_h^*(s, a)$ at each step h .

For any h and s , decompose the sub-optimality:

$$\begin{aligned} V_h^*(s) - V_h^{\widehat{\pi}^*}(s) &\leq \underbrace{\max_a Q_h^*(s, a) - Q_h^*(s, \widehat{\pi}_h^*(s))}_{\leq 2\|Q_h^* - \widehat{Q}_h^*\|_\infty \text{ (since } \widehat{\pi}_h^* \text{ maximizes } \widehat{Q}_h^*)} + \underbrace{P_h(V_{h+1}^* - V_{h+1}^{\widehat{\pi}^*})(s, \widehat{\pi}_h^*(s))}_{\leq \|V_{h+1}^* - V_{h+1}^{\widehat{\pi}^*}\|_\infty} \end{aligned}$$

Unrolling (with $V_H^* = V_H^{\widehat{\pi}^*} = 0$):

$$\|V_0^* - V_0^{\widehat{\pi}^*}\|_\infty \leq 2 \sum_{h=0}^{H-1} \|Q_h^* - \widehat{Q}_h^*\|_\infty.$$

Horizon amplification (finite-horizon)

The proof of Theorem 2.8 bounds $\|Q_h^* - \widehat{Q}_h^*\|_\infty \lesssim (H-h)/\sqrt{N}$ at each stage h .

Substituting: $\|V_0^* - V_0^{\widehat{\pi}^*}\|_\infty \lesssim H^2/\sqrt{N}$ — an extra factor of H vs. value estimation alone. As in the discounted case, the sub-optimality bullet of Theorem 2.8 shows this H amplification can be avoided with a sharper analysis.

Offline Policy Evaluation: Setup

Fix a policy π that we wish to evaluate. Suppose instead of a generative model, we have a **fixed offline dataset**:

$$\mathcal{D} = \{(s, a, h, s') : (s, a, h) \text{ observed } N_h(s, a) \text{ times}\}.$$

Counts $N_h(s, a)$ are **non-uniform** — set by how data was collected, not by us.

Algorithm: plug-in estimator $\hat{P}_h(s'|s, a) = \frac{\#\{s' \text{ seen at } (s, a, h)\}}{N_h(s, a)}$, followed by

backward induction to compute \hat{V}^π .

Question: What is $\mathbb{E}[(\hat{V}_0^\pi(s_0) - V_0^\pi(s_0))^2]$ as a function of $\{N_h(s, a)\}$?

Error Propagation in Backward Induction

Evaluate π by backward induction: $\widehat{V}_H^\pi = 0$; for $h = H - 1, \dots, 0$:

$$\widehat{V}_h^\pi(s) = r_h(s, \pi(s)) + \widehat{P}_h \widehat{V}_{h+1}^\pi(s, \pi(s)).$$

Let $\delta_h \triangleq \widehat{V}_h^\pi - V_h^\pi$. Subtracting the true recursion:

$$\delta_h(s) = \underbrace{(\widehat{P}_h - P_h)V_{h+1}^\pi(s, \pi(s))}_{\xi_h(s) : \text{estimation error at step } h} + \widehat{P}_h \delta_{h+1}(s, \pi(s)).$$

Unrolling from $h = 0$ (approximating $\widehat{P} \approx P$ for simplicity):

$$\delta_0(s_0) \approx \sum_{h=0}^{H-1} \sum_{s,a} w_h^\pi(s, a) \cdot \xi_h(s),$$

where $w_h^\pi(s, a) \triangleq \Pr_\pi[s_h = s, a_h = a \mid s_0]$ is the **visitation probability** of π .

Each step's error is **weighted by how often π visits** that state-action pair.

MSE as a Function of Coverage

Since $\xi_h(s) = (\hat{P}_h - P_h)V_{h+1}^\pi(s, \pi(s))$ is an average of i.i.d. terms with zero mean, and the samples at different (s, a, h) are independent:

$$\begin{aligned}\text{MSE} &= \mathbb{E}[\delta_0(s_0)^2] \approx \sum_{h,s,a} [w_h^\pi(s, a)]^2 \cdot \text{Var}[\xi_h(s)] \\ &= \sum_{h,s,a} [w_h^\pi(s, a)]^2 \cdot \frac{\text{Var}_{P_h(\cdot|s, \pi(s))}(V_{h+1}^\pi)}{N_h(s, a)} \\ &\leq H^2 \sum_{h,s,a} \frac{[w_h^\pi(s, a)]^2}{N_h(s, a)},\end{aligned}$$

using $0 \leq V_{h+1}^\pi \leq H$ so that the variance is at most H^2 .

Key formula

$$\text{MSE} \lesssim H^2 \sum_{h,s,a} \frac{[w_h^\pi(s, a)]^2}{N_h(s, a)}.$$

The Concentration Coefficient

Since $\sum_{s,a} w_h^\pi(s, a) = 1$ for each fixed h :

$$\sum_{s,a} \frac{[w_h^\pi(s, a)]^2}{N_h(s, a)} = \sum_{s,a} w_h^\pi(s, a) \cdot \frac{w_h^\pi(s, a)}{N_h(s, a)} \leq \max_{s,a} \frac{w_h^\pi(s, a)}{N_h(s, a)}.$$

Summing over h defines the **concentration coefficient**:

$$\sum_{h,s,a} \frac{[w_h^\pi(s, a)]^2}{N_h(s, a)} \leq \underbrace{\sum_h \max_{s,a} \frac{w_h^\pi(s, a)}{N_h(s, a)}}_{C(\pi, \mathcal{D})}.$$

Intuition: $C(\pi, \mathcal{D})$ is small when the data covers π 's trajectory well.

- $N_h \propto w_h^\pi$ (data matches π): $C = H$, $\text{MSE} \lesssim H^3 / N_{\text{total}}$.
- $N_h(s, a) \ll w_h^\pi(s, a)$ at visited pairs: C large, evaluation unreliable.

Looking ahead: $\hat{\pi}^*$ is optimal in $\hat{\mathcal{M}}$, so $\hat{V}^{\hat{\pi}^*} \geq \hat{V}^\pi$ for any π . If $C(\hat{\pi}^*, \mathcal{D})$ is small, then $\hat{V}^{\hat{\pi}^*} \approx V^{\hat{\pi}^*}$ and hence $V^{\hat{\pi}^*}$ is near-optimal. But $\hat{\pi}^*$ is unknown before seeing data — offline RL bounds C for the *best competing policy*, a **coverage condition** we will revisit in later lectures.