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

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:

$$s' \sim P(\cdot \mid s, a), \qquad r(s, a).$$

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 $\widehat{\pi}$ satisfying

$$\left\| Q^* - Q^{\widehat{\pi}} \right\|_\infty \le \varepsilon$$

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 $\widehat{\pi}$ satisfying

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

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

Algorithm: Estimate P, Then Plan

Model-Based Algorithm

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

$$\widehat{P}(s' \mid s, a) \;=\; \frac{\#\{i : s'_i = s'\}}{N}.$$

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

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

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

Two Key Lemmas

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

Lemma 2.2 (Simulation Lemma)

For any stationary policy π:

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

Simulation Lemma: Proof

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

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

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

Two Key Lemmas

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

Lemma 2.2 (Simulation Lemma)

For any stationary policy π:

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

Lemma 2.3 (Norm Bound)

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

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

Two Key Lemmas

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

Lemma 2.2 (Simulation Lemma)

For any stationary policy π:

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

Lemma 2.3 (Norm Bound)

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

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

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

Hoeffding's Inequality

The central concentration inequality we will use throughout.

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_N$ 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 $1-\delta$:

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

Hoeffding's Inequality

The central concentration inequality we will use throughout.

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_N$ 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:

- 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 $\log(2K/\delta)$.

Naive Bound: Main Result

Proposition 2.1 (Naive model-based approach)

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

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

Proof of Proposition 2.1 — Step 1: Model Accuracy

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

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

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:

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

$$= \max_{s,a} \max_{z \in \{-1,1\}^{|\mathcal{S}|}} \sum_{s'} z_{s'} \frac{1}{N} \sum_{i=1}^{N} \left(P(s_i'|s,a) - \mathbf{1}\{s_i' = s'\} \right) \rangle$$

Proof of Proposition 2.1 — Step 1: Model Accuracy

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

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

Proof of Proposition 2.1 — Step 1: Model Accuracy

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

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

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:

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

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 →)

Proof of Proposition 2.1 — Step 2: Value Accuracy

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

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

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

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

Proof of Proposition 2.1 — Step 2: Value Accuracy

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

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

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

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

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

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 $\left\| Q^\pi - \widehat{Q}^\pi \right\|_\infty \leq \varepsilon$ *for every* π.

Bounding $\left\| \widehat{Q}^* - Q^* \right\|_\infty$**.**
Using $\left|\sup_\pi f(\pi) - \sup_\pi g(\pi)\right| \leq \sup_\pi |f(\pi) - g(\pi)|$:

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

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

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

Bounding $\left\| \widehat{Q}^* - Q^* \right\|_\infty$**.**
Using $\left| \sup_\pi f(\pi) - \sup_\pi g(\pi) \right| \le \sup_\pi |f(\pi) - g(\pi)|$:

$$\left| \widehat{Q}^*(s,a) - Q^*(s,a) \right| \;\le\; \sup_\pi \left| \widehat{Q}^\pi(s,a) - Q^\pi(s,a) \right| \;\le\; \varepsilon.$$

Bounding $\left\| Q^{\widehat{\pi}^*} - Q^* \right\|_\infty$**.**
Since $\widehat{\pi}^*$ is optimal in $\widehat{\mathcal{M}}$, $\widehat{Q}^{\widehat{\pi}^*} = \widehat{Q}^*$. By triangle inequality and Step 2:

$$\left\| Q^{\widehat{\pi}^*} - Q^* \right\|_\infty \;\le\; \left\| Q^{\widehat{\pi}^*} - \widehat{Q}^* \right\|_\infty + \left\| \widehat{Q}^* - Q^* \right\|_\infty \;\le\; \varepsilon + \varepsilon \;=\; 2\varepsilon.$$

□

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

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 $\widehat{\pi}^*$ optimal in $\widehat{\mathcal{M}}$. Then componentwise:

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

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

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

Sublinear Sample Complexity: Main Result

Proposition 2.4 (Crude value bound)

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

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

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

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

- $|\mathcal{S}||\mathcal{A}|$ 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:

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

Proof of Proposition 2.4

By Lemma 2.5 and Lemma 2.3:

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

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

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

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

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

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

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

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 $\left\| \widehat{Q}^* - Q^* \right\|_\infty \leq \varepsilon$ *and* $\left\| Q^* - Q^{\widehat{\pi}^*} \right\|_\infty \leq \varepsilon$ 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 $\left\|\widehat{Q}^* - Q^*\right\|_\infty \leq \varepsilon$ 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).

Bottom Line

Sample complexity with a generative model

To find a policy $\widehat{\pi}$ with $\left\| Q^* - Q^{\widehat{\pi}} \right\|_\infty \le \varepsilon$ (with high probability):

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

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 $\widehat{P}$, then plan. Simple and optimal.

Today's Plan

Finite-Horizon MDP: Setup and Result

Recall the finite-horizon MDP: stages $h = 0, \ldots, H-1$, 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 $\widehat{P}_h$.
Total samples: $H|\mathcal{S}||\mathcal{A}| \cdot N$.

Theorem 2.8 (Finite-horizon minimax upper bound)

To achieve $\left\| Q_0^* - \widehat{Q}_0^* \right\|_\infty \leq \varepsilon$ 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:

$$V^*_h(s) - V^{\widehat{\pi}^*}_h(s) \le \underbrace{\max_a Q^*_h(s,a) - Q^*_h(s,\widehat{\pi}^*_h(s))}_{\le 2\left\|Q^*_h - \widehat{Q}^*_h\right\|_\infty \text{ (since } \widehat{\pi}^*_h \text{ maximizes } \widehat{Q}^*_h\text{)}} + \underbrace{P_h\left(V^*_{h+1} - V^{\widehat{\pi}^*}_{h+1}\right)(s,\widehat{\pi}^*_h(s))}_{\le \left\|V^*_{h+1} - V^{\widehat{\pi}^*}_{h+1}\right\|_\infty}$$

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

$$\left\|V^*_0 - V^{\widehat{\pi}^*}_0\right\|_\infty \le 2\sum_{h=0}^{H-1}\left\|Q^*_h - \widehat{Q}^*_h\right\|_\infty .$$

Horizon amplification (finite-horizon)

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

Substituting: $\left\|V^*_0 - V^{\widehat{\pi}^*}_0\right\|_\infty \lesssim H^2/\sqrt{N}$ — 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.