Lecture 6: Tabular Analysis

From Planning to Learning

Having developed planning algorithms (Lectures 1–4) and concentration tools (Lecture 5), we now analyze how well we can *learn* near-optimal policies from finite data in the **tabular setting**. We study the *certainty-equivalence* method, a natural model-based RL algorithm, and derive four increasingly sophisticated sample complexity bounds that illustrate fundamental tradeoffs between state-space size and effective horizon.

The Central Question: Given n samples per state-action pair, how close is the optimal policy of the *estimated* MDP to the true optimal policy? We measure closeness by $V_M^\star(s) - V_M^{\pi^\star_{\widehat{M}}}(s)$—the suboptimality when deploying the learned policy in the true environment. As we will see, different proof techniques yield bounds with different tradeoffs between the dependence on $|\mathcal{S}|$ and $1/(1-\gamma)$.

The Certainty-Equivalence Algorithm

Certainty-equivalence is a **model-based** RL algorithm. The idea is simple: (1) estimate an MDP model from data, then (2) compute the optimal policy in the estimated model as if it were the true model.

Data Collection

We assume access to a *generative model* (also called a simulator): for any state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$, we can query the environment and receive an independent sample (r, s') where $r \sim \mathcal{R}(s,a)$ and $s' \sim P(\cdot \mid s,a)$.

Given a dataset D (which could come from trajectories or from the generative model), we convert it into a bag of (s,a,r,s') tuples. For every $s \in \mathcal{S}$ and $a \in \mathcal{A}$, define:

$$D_{s,a} := \{(r,s') : (s,a,r,s') \in D\},$$

the subset of tuples where the state is s and the action is a. We write $n(s,a) = |D_{s,a}|$ for the number of samples at each state-action pair.

For simplicity, we assume a **balanced** data collection: $n(s,a) = n$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. The total number of samples is thus $n \cdot |\mathcal{S}| \cdot |\mathcal{A}|$.

Model Estimation

The certainty-equivalence model estimates the transition and reward functions by their empirical averages.

Definition 1 (Certainty-Equivalence Model)**.** *Given datasets $\{D_{s,a}\}_{s\in\mathcal{S},a\in\mathcal{A}}$, define the estimated MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{R}, \gamma)$ where:*

Estimated transition:

$$\widehat{P}(s' \mid s, a) = \frac{1}{|D_{s,a}|} \sum_{(r,s'')\in D_{s,a}} \mathbb{1}[s'' = s'], \qquad \forall s' \in \mathcal{S}.$$

Estimated reward:

$$\widehat{R}(s, a) = \frac{1}{|D_{s,a}|} \sum_{(r,s')\in D_{s,a}} r.$$

In words, $\widehat{P}(s' \mid s, a)$ is the empirical frequency of observing s' after taking action a in state s, and $\widehat{R}(s, a)$ is the average observed reward. These are the maximum likelihood estimates *of the transition and reward functions.*

Remark 1. *We need $n(s, a) > 0$ for every (s, a) to ensure $\widehat{P}(\cdot \mid s, a)$ is a well-defined probability distribution. The generative model assumption guarantees this by allowing us to query any (s, a) pair.*

The Algorithm

The complete certainty-equivalence algorithm is:

1. **Collect data:** For every $(s, a) \in \mathcal{S} \times \mathcal{A}$, collect n i.i.d. samples from $P(\cdot \mid s, a)$ and $R(s, a)$.
2. **Estimate model:** Compute $\widehat{P}$ and $\widehat{R}$ as above.
3. **Plan in estimated model:** Solve for the optimal policy $\pi^\star_{\widehat{M}}$ of $\widehat{M}$ (e.g., via Value Iteration or Policy Iteration).
4. **Deploy:** Execute $\pi^\star_{\widehat{M}}$ in the true MDP M.

Model-Based vs. Value-Based: Certainty-equivalence explicitly stores an estimated MDP model, which requires $O(|\mathcal{S}|^2|\mathcal{A}|)$ space (for the transition matrix) and has a batch nature. In contrast, value-based methods such as Q-learning and SARSA store only Q-value functions ($O(|\mathcal{S}||\mathcal{A}|)$ space) and can operate online. Value-based methods are typically less sample-efficient than model-based methods in the tabular setting, but they are easier to combine with function approximation (e.g., deep neural networks), which has led to many empirical successes.

Analysis Setup and Goals

We analyze the certainty-equivalence algorithm under the following assumptions:

- **Generative model:** For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, we collect n i.i.d. samples.
- **Bounded rewards:** $R(s, a) \in [0, R_{\max}]$ for all (s, a).
- **Discount factor:** $\gamma \in (0, 1)$.

Define $V_{\max} := R_{\max}/(1-\gamma)$, the maximum possible value. Our goal is to derive high-probability bounds on the *suboptimality* of $\pi^\star_{\widehat{M}}$:

$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s), \qquad \forall s \in \mathcal{S},$$

as a function of n, $|\mathcal{S}|$, $|\mathcal{A}|$, and $1/(1-\gamma)$.

We present **four different analyses**, each highlighting a different proof technique and achieving a different tradeoff:

Analysis	**Bound (ignoring $\widetilde{O}$)**	**Key Technique**
I: Naive	$\dfrac{\lvert\mathcal{S}\rvert \cdot V_{\max}}{\sqrt{n}(1-\gamma)}$	Entry-wise Hoeffding
II: Improved	$\dfrac{\sqrt{\lvert\mathcal{S}\rvert} \cdot V_{\max}}{\sqrt{n}(1-\gamma)}$	ℓ_1 concentration
III: No $\lvert\mathcal{S}\rvert$	$\dfrac{V_{\max}}{\sqrt{n}(1-\gamma)^2}$	Contraction
IV: Best of both	$\dfrac{V_{\max}}{\sqrt{n}(1-\gamma)}$	Bellman residual (large n)

Throughout this lecture, $\widetilde{O}(\cdot)$ suppresses poly-logarithmic factors in $|\mathcal{S}|$, $|\mathcal{A}|$, and $1/\delta$. We state bounds in terms of $V_{\max} = R_{\max}/(1-\gamma)$ and focus on the polynomial dependences on $|\mathcal{S}|$, n, and $1/(1-\gamma)$.

Key Analytical Tools

Before diving into the analyses, we develop three key lemmas that are used across all the analyses.

The Simulation Lemma

The Simulation Lemma bounds how much a policy's value changes when we evaluate it in a different MDP. This is the main workhorse for translating model estimation errors into value function errors.

Lemma 1 (Simulation Lemma[1]). *Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ and $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{R}, \gamma)$ be two MDPs sharing the same state space, action space, and discount factor. Suppose*

$$\max_{s,a} |\widehat{R}(s,a) - R(s,a)| \le \varepsilon_R \qquad \textit{and} \qquad \max_{s,a} \|\widehat{P}(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \le \varepsilon_P.$$

Then for any policy $\pi : \mathcal{S} \to \mathcal{A}$,

$$\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \le \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma \varepsilon_P V_{\max}}{2(1-\gamma)},$$

where $V_{\max} := R_{\max}/(1-\gamma)$.

Proof. For any state $s \in \mathcal{S}$, by the Bellman equations in the two MDPs:

$$\begin{aligned}
&|V^\pi_{\widehat{M}}(s) - V^\pi_M(s)| \\
&= \left|\widehat{R}(s,\pi(s)) + \gamma\langle \widehat{P}(s,\pi(s)), V^\pi_{\widehat{M}}\rangle - R(s,\pi(s)) - \gamma\langle P(s,\pi(s)), V^\pi_M\rangle\right| \\
&= \left|\big(\widehat{R}(s,\pi(s)) - R(s,\pi(s))\big) + \gamma\langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), V^\pi_{\widehat{M}}\rangle + \gamma\langle P(s,\pi(s)), V^\pi_{\widehat{M}} - V^\pi_M\rangle\right| \\
&\le \left|\widehat{R}(s,\pi(s)) - R(s,\pi(s))\right| + \gamma\left|\langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), V^\pi_{\widehat{M}}\rangle\right| + \gamma\left|\langle P(s,\pi(s)), V^\pi_{\widehat{M}} - V^\pi_M\rangle\right| \\
&\le \varepsilon_R + \gamma\left|\langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), V^\pi_{\widehat{M}}\rangle\right| + \gamma\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty,
\end{aligned}$$

where $\langle P(s,a), V\rangle = \sum_{s'} P(s'|s,a)V(s')$. The second line adds and subtracts $\gamma\langle P(s,\pi(s)), V^\pi_{\widehat{M}}\rangle$, the third uses the triangle inequality, and the fourth uses $|\widehat{R}(s,\pi(s)) - R(s,\pi(s))| \le \varepsilon_R$ and $|\langle P(s,\pi(s)), V^\pi_{\widehat{M}} - V^\pi_M\rangle| \le \|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty$ (since $P(s,\pi(s))$ is a probability distribution).

The key step is the *centering trick*: since both $\widehat{P}(s,\pi(s))$ and $P(s,\pi(s))$ are valid probability distributions (each sums to 1), their difference sums to zero: $\langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), \mathbf{1}\rangle = 0$.

[1] Due to Kearns and Singh (2002).

Therefore, for any constant c,

$$\langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), V^\pi_{\widehat{M}} \rangle = \langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), V^\pi_{\widehat{M}} - c \cdot \mathbf{1} \rangle.$$

Choosing $c = V_{\max}/2$ to center the range of $V^\pi_{\widehat{M}}$ around the origin (since $V^\pi_{\widehat{M}} \in [0, V_{\max}]$), we get $\|V^\pi_{\widehat{M}} - \frac{V_{\max}}{2} \cdot \mathbf{1}\|_\infty \leq V_{\max}/2$. By Hölder's inequality:

$$\begin{aligned} \left|\langle \widehat{P}(s,\pi(s)) - P(s,\pi(s)), V^\pi_{\widehat{M}} - \tfrac{V_{\max}}{2} \cdot \mathbf{1} \rangle\right| &\leq \|\widehat{P}(s,\pi(s)) - P(s,\pi(s))\|_1 \cdot \|V^\pi_{\widehat{M}} - \tfrac{V_{\max}}{2} \cdot \mathbf{1}\|_\infty \\ &\leq \varepsilon_P \cdot \frac{V_{\max}}{2}. \end{aligned}$$

Putting it together:

$$|V^\pi_{\widehat{M}}(s) - V^\pi_M(s)| \leq \varepsilon_R + \frac{\gamma \varepsilon_P V_{\max}}{2} + \gamma \|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty.$$

Since this holds for all $s \in \mathcal{S}$, taking the ℓ_∞-norm on the left-hand side:

$$\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \leq \varepsilon_R + \frac{\gamma \varepsilon_P V_{\max}}{2} + \gamma \|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty.$$

Rearranging $(1-\gamma)\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \leq \varepsilon_R + \frac{\gamma \varepsilon_P V_{\max}}{2}$ gives the result. □

The Centering Trick: The trick of subtracting $\frac{V_{\max}}{2} \cdot \mathbf{1}$ is crucial. Without it, Hölder's inequality would give $\|\widehat{P} - P\|_1 \cdot \|V^\pi_{\widehat{M}}\|_\infty \leq \varepsilon_P \cdot V_{\max}$. The centering reduces the effective range by a factor of 2, exploiting the fact that the difference of two probability distributions sums to zero. This seemingly small improvement matters when the bound is propagated through the analysis.

Alternative Proof via Performance Difference

We now sketch an alternative and more "modern" proof of the Simulation Lemma using the *performance difference identity*, which is a powerful tool that will reappear later.

Lemma 2 (Performance Difference Identity)**.** *For any function $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, any initial state distribution $d_0 \in \Delta(\mathcal{S})$, and any policy π, let $d^\pi \in \Delta(\mathcal{S}\times\mathcal{A})$ be the discounted state-action occupancy induced by π from d_0:*

$$d^\pi(s,a) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr(s_t = s, a_t = a \mid \pi, d_0).$$

Then:

$$\mathbb{E}_{s\sim d_0,a\sim\pi}[f(s,a)] - J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}\left[f(s,a) - R(s,a) - \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[f(s',\pi(s'))]\right],$$

where $J(\pi) = \mathbb{E}_{s\sim d_0}[V^\pi(s)]$.

Proof sketch. Let d_t^π denote the distribution of (s_t, a_t) under π starting from d_0. Note that $J(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}[R(s,a)]$, so the terms involving R can be canceled from both sides. The remaining terms are:

$$\frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi, s'\sim P(\cdot|s,a)}[f(s,a) - \gamma f(s',\pi(s'))] = \sum_{t=0}^{\infty}\gamma^t\mathbb{E}_{(s,a)\sim d_t^\pi, s'\sim P, a'\sim\pi}[f(s,a) - \gamma f(s',a')].$$

The key observation is a *telescoping*: the $\gamma f(s',a')$ term at time t cancels with the $f(s,a)$ term at time $t+1$, because (s',a') at time t has the same distribution d_{t+1}^π as (s,a) at time $t+1$, with the discount factor accounting for the γ multiplier. After cancellation, only the first term $\mathbb{E}_{(s,a)\sim d_0^\pi}[f(s,a)] = \mathbb{E}_{s\sim d_0,a\sim\pi}[f(s,a)]$ survives. □

To prove the Simulation Lemma using this identity, set $f = Q_{\widehat{M}}^\pi$ and choose d_0 as the point mass on any state s_0. For the left-hand side: $\mathbb{E}_{s\sim d_0,a\sim\pi}[f(s,a)] = Q_{\widehat{M}}^\pi(s_0,\pi(s_0)) = V_{\widehat{M}}^\pi(s_0)$, and $J(\pi) = V_M^\pi(s_0)$. For the right-hand side, we substitute $f = Q_{\widehat{M}}^\pi$ (noting that $f(s',\pi(s')) = Q_{\widehat{M}}^\pi(s',\pi(s')) = V_{\widehat{M}}^\pi(s')$) and then expand $Q_{\widehat{M}}^\pi(s,a)$ via its Bellman equation $Q_{\widehat{M}}^\pi(s,a) = \widehat{R}(s,a) + \gamma\langle\widehat{P}(s,a), V_{\widehat{M}}^\pi\rangle$ in $\widehat{M}$:

$$\begin{aligned} V_{\widehat{M}}^\pi(s_0) - V_M^\pi(s_0) &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}\left[Q_{\widehat{M}}^\pi(s,a) - R(s,a) - \gamma\langle P(s,a), V_{\widehat{M}}^\pi\rangle\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{(s,a)\sim d^\pi}\left[\widehat{R}(s,a) + \gamma\langle\widehat{P}(s,a), V_{\widehat{M}}^\pi\rangle - R(s,a) - \gamma\langle P(s,a), V_{\widehat{M}}^\pi\rangle\right]. \end{aligned}$$

The rest of the proof then bounds each term similarly to the original proof.

From Evaluation Error to Decision Loss

The Simulation Lemma bounds the *policy evaluation error* $\|V_{\widehat{M}}^\pi - V_M^\pi\|_\infty$, but we ultimately care about the *suboptimality* of the learned policy $\pi_{\widehat{M}}^\star$. The following lemma makes the connection.

Lemma 3 (Evaluation Error to Decision Loss)**.** *For any* $s \in \mathcal{S}$*:*

$$V_M^\star(s) - V_M^{\pi_{\widehat{M}}^\star}(s) \le 2\sup_{\pi:\mathcal{S}\to\mathcal{A}}\|V_{\widehat{M}}^\pi - V_M^\pi\|_\infty.$$

Proof. For any $s \in \mathcal{S}$:

$$\begin{aligned} V_M^\star(s) - V_M^{\pi_{\widehat{M}}^\star}(s) &= V_M^{\pi_M^\star}(s) - V_{\widehat{M}}^{\pi_M^\star}(s) + V_{\widehat{M}}^{\pi_M^\star}(s) - V_M^{\pi_{\widehat{M}}^\star}(s) \\ &\le V_M^{\pi_M^\star}(s) - V_{\widehat{M}}^{\pi_M^\star}(s) + V_{\widehat{M}}^{\pi_{\widehat{M}}^\star}(s) - V_M^{\pi_{\widehat{M}}^\star}(s) \\ &\le \|V_M^{\pi_M^\star} - V_{\widehat{M}}^{\pi_M^\star}\|_\infty + \|V_{\widehat{M}}^{\pi_{\widehat{M}}^\star} - V_M^{\pi_{\widehat{M}}^\star}\|_\infty. \end{aligned}$$

The second line uses the fact that $\pi_{\widehat{M}}^\star$ maximizes $V_{\widehat{M}}^\cdot$, so $V_{\widehat{M}}^{\pi_M^\star}(s) \le V_{\widehat{M}}^{\pi_{\widehat{M}}^\star}(s)$. □

Intuition: The suboptimality of $\pi_{\widehat{M}}^\star$ is bounded by how much the estimated model distorts the value of *any* policy. The factor of 2 arises because there are two sources of error: (1) $\pi_M^\star$ may appear worse in $\widehat{M}$ than it actually is, and (2) $\pi_{\widehat{M}}^\star$ may appear better in $\widehat{M}$ than it actually is.

Analysis I: Naive Bound

We now combine the Simulation Lemma with basic concentration inequalities (Lecture 5) to obtain our first sample complexity bound.

Concentration of the Estimated Model

By Hoeffding's inequality (Corollary 4 from Lecture 5) and the union bound:

Reward estimation. Each $\widehat{R}(s,a)$ is the average of n i.i.d. random variables in $[0, R_{\max}]$. By Hoeffding's inequality and a union bound over all $|\mathcal{S}| \cdot |\mathcal{A}|$ pairs, with probability at least $1 - \delta/2$:

$$\max_{s,a} |\widehat{R}(s,a) - R(s,a)| \le R_{\max} \sqrt{\frac{1}{2n} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}. \tag{1}$$

Transition estimation (entry-wise). For each fixed (s,a,s'), the quantity $\widehat{P}(s'|s,a)$ is the average of n i.i.d. Bernoulli random variables with mean $P(s'|s,a)$. By Hoeffding's inequality and a union bound over all $|\mathcal{S}| \cdot |\mathcal{A}| \cdot |\mathcal{S}|$ triples, with probability at least $1 - \delta/2$:

$$\max_{s,a,s'} |\widehat{P}(s'|s,a) - P(s'|s,a)| \le \sqrt{\frac{1}{2n} \log \frac{4|\mathcal{S}|^2|\mathcal{A}|}{\delta}}. \tag{2}$$

From ℓ_∞ to ℓ_1. To apply the Simulation Lemma, we need an ℓ_1 bound on $\widehat{P}(s,a) - P(s,a)$. The naive approach converts the entry-wise bound (2) to an ℓ_1 bound by summing over $|\mathcal{S}|$

entries:

$$\max_{s,a} \|\widehat{P}(s,a) - P(s,a)\|_1 \le |\mathcal{S}| \cdot \max_{s,a} \|\widehat{P}(s,a) - P(s,a)\|_\infty \le |\mathcal{S}| \cdot \sqrt{\frac{1}{2n} \log \frac{4|\mathcal{S}|^2|\mathcal{A}|}{\delta}}. \quad (3)$$

Where the $|\mathcal{S}|$ Factor Comes From: The conversion from ℓ_∞ to ℓ_1 introduces a factor of $|\mathcal{S}|$ (since $\|v\|_1 \le |\mathcal{S}| \cdot \|v\|_\infty$ for $v \in \mathbb{R}^{|\mathcal{S}|}$). This is the main source of looseness in the naive analysis.

Putting It Together

Setting $\varepsilon_R = R_{\max}\sqrt{\frac{1}{2n} \log \frac{4|\mathcal{S}||\mathcal{A}|}{\delta}}$ and $\varepsilon_P = |\mathcal{S}| \cdot \sqrt{\frac{1}{2n} \log \frac{4|\mathcal{S}|^2|\mathcal{A}|}{\delta}}$, the Simulation Lemma (Lemma 1) gives:

$$\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty \le \frac{\varepsilon_R}{1-\gamma} + \frac{\gamma \varepsilon_P V_{\max}}{2(1-\gamma)}.$$

The dominant term (for large $|\mathcal{S}|$) is the transition estimation error:

$$\frac{\gamma \varepsilon_P V_{\max}}{2(1-\gamma)} = \frac{\gamma |\mathcal{S}| V_{\max}}{2(1-\gamma)} \cdot \sqrt{\frac{1}{2n} \log \frac{4|\mathcal{S}|^2|\mathcal{A}|}{\delta}}.$$

Combining with Lemma 3:

Analysis I Result:

$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \widetilde{O}\left(\frac{|\mathcal{S}| \cdot V_{\max}}{\sqrt{n}(1-\gamma)}\right), \qquad \forall s \in \mathcal{S}.$$

Sample complexity. To achieve ε-suboptimality, we need

$$n = \widetilde{O}\left(\frac{|\mathcal{S}|^2 \cdot V^2_{\max}}{\varepsilon^2(1-\gamma)^2}\right) = \widetilde{O}\left(\frac{|\mathcal{S}|^2 R^2_{\max}}{\varepsilon^2(1-\gamma)^4}\right)$$

samples per state-action pair.

Analysis II: Improved State-Space Dependence

The previous analysis bounded the ℓ_1 error of $\widehat{P}(s,a) - P(s,a)$ by first bounding each entry in ℓ_∞ and then multiplying by $|\mathcal{S}|$. This is loose because it ignores the *correlations* between

the entries (they must sum to zero since both $\widehat{P}$ and P are probability distributions). We now prove a tighter *direct* ℓ_1 concentration bound for multinomial distributions.

ℓ_1 Concentration via Signed Vectors

The key observation is the following variational characterization of the ℓ_1 norm:

Lemma 4 (ℓ_1 Norm as Supremum)**.** *For any vector $v \in \mathbb{R}^{|\mathcal{S}|}$:*

$$\|v\|_1 = \sup_{u \in \{-1,1\}^{|\mathcal{S}|}} u^\top v.$$

Proof. By choosing $u_i = \operatorname{sign}(v_i)$ for each i, we achieve $u^\top v = \sum_i |v_i| = \|v\|_1$. Conversely, $u^\top v = \sum_i u_i v_i \le \sum_i |v_i| = \|v\|_1$ for any $u \in \{-1,1\}^{|\mathcal{S}|}$. □

Strategy. For each fixed (s,a), the empirical transition $\widehat{P}(s,a)$ is the average of n i.i.d. random vectors $\mathbf{e}_{s'_1}, \ldots, \mathbf{e}_{s'_n}$ (unit vectors), each with mean $P(s,a)$. For any fixed $u \in \{-1,1\}^{|\mathcal{S}|}$, the projection $u^\top \widehat{P}(s,a)$ is the average of i.i.d. scalar random variables $u^\top \mathbf{e}_{s'_i}$ taking values in $\{-1,1\}$, and thus has bounded range $[-1,1]$.

By Hoeffding's inequality, for each fixed (s,a) and $u \in \{-1,1\}^{|\mathcal{S}|}$, with probability at least $1-\delta'$:

$$u^\top(\widehat{P}(s,a) - P(s,a)) \le \sqrt{\frac{2}{n}\log\frac{1}{\delta'}}. \tag{4}$$

Here we use the one-sided Hoeffding bound: $\Pr(u^\top(\widehat{P} - P) \ge t) \le \exp(-2nt^2/4)$, where the range is $R = 1 - (-1) = 2$. Setting the RHS to δ' and solving gives $t = \sqrt{\frac{2}{n}\log\frac{1}{\delta'}}$.

Taking a union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and all $u \in \{-1,1\}^{|\mathcal{S}|}$, and choosing $\delta' = \delta/(2|\mathcal{S}||\mathcal{A}| \cdot 2^{|\mathcal{S}|})$, we obtain that with probability at least $1 - \delta/2$:

$$\max_{s,a} \|\widehat{P}(s,a) - P(s,a)\|_1 = \max_{s,a} \max_{u \in \{-1,1\}^{|\mathcal{S}|}} u^\top(\widehat{P}(s,a) - P(s,a)) \le \sqrt{\frac{2}{n}\log\frac{2|\mathcal{S}||\mathcal{A}| \cdot 2^{|\mathcal{S}|}}{\delta}}. \tag{5}$$

Improvement: The naive bound (3) gives $\|\widehat{P}(s,a) - P(s,a)\|_1 = \widetilde{O}(|\mathcal{S}| \cdot n^{-1/2})$, while the direct ℓ_1 bound (5) gives $\widetilde{O}(\sqrt{|\mathcal{S}|/n})$. The improvement from $|\mathcal{S}|$ to $\sqrt{|\mathcal{S}|}$ comes from two sources: (1) the projection $u^\top \widehat{P}(s,a)$ concentrates at the scalar rate $O(1/\sqrt{n})$ regardless of the dimension, and (2) the union bound over $2^{|\mathcal{S}|}$ choices of u only contributes an additive $|\mathcal{S}| \cdot \log 2$ inside the logarithm.

Remark 2. *A tiny improvement: for $u = \pm\mathbf{1}$ (the all-ones or all-negative-ones vector), $u^\top(\widehat{P}(s,a) - P(s,a)) = \pm(\sum_{s'} \widehat{P}(s'|s,a) - \sum_{s'} P(s'|s,a)) = 0$ exactly (since both are probability distributions). So we can exclude $u = \pm\mathbf{1}$ from the union bound, replacing $2^{|\mathcal{S}|}$ with $2^{|\mathcal{S}|} - 2$.*

Propagating the Improvement

Substituting the improved ℓ_1 bound into the Simulation Lemma and following the same steps as Analysis I:

Analysis II Result:

$$V_M^\star(s) - V_M^{\pi^\star_{\widehat{M}}}(s) = \widetilde{O}\left(\frac{\sqrt{|\mathcal{S}|} \cdot V_{\max}}{\sqrt{n}(1-\gamma)}\right), \qquad \forall s \in \mathcal{S}.$$

The dependence on state-space size improves from $|\mathcal{S}|$ to $\sqrt{|\mathcal{S}|}$, with no change in the horizon dependence $1/(1-\gamma)$.

Analysis III: Removing State-Space Dependence

The previous two analyses worked with the Simulation Lemma, which bounds the *policy evaluation error* $\|V^\pi_{\widehat{M}} - V^\pi_M\|_\infty$ for *every* policy π, and then uses Lemma 3 to convert this to suboptimality. We now take a fundamentally different approach: instead of comparing V^π across models, we directly compare the *optimal Q-value functions* $Q^\star_{\widehat{M}}$ and $Q^\star_M$.

Contraction Bound

The starting point is the following observation, which uses the contraction property of the Bellman optimality operator.

Lemma 5 (Contraction Bound)**.** *Let $\mathcal{T}_{\widehat{M}}$ denote the Bellman optimality operator in $\widehat{M}$. Then:*

$$\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty \le \frac{1}{1-\gamma}\|Q^\star_M - \mathcal{T}_{\widehat{M}} Q^\star_M\|_\infty. \tag{6}$$

Proof. Since $Q^\star_{\widehat{M}}$ is the fixed point of $\mathcal{T}_{\widehat{M}}$:

$$\begin{aligned}
\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty &= \|\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}} - \mathcal{T}_{\widehat{M}} Q^\star_M + \mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty \\
&\le \|\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}} - \mathcal{T}_{\widehat{M}} Q^\star_M\|_\infty + \|\mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty \\
&\le \gamma\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty + \|\mathcal{T}_{\widehat{M}} Q^\star_M - Q^\star_M\|_\infty,
\end{aligned}$$

where the last step uses the γ-contraction property of $\mathcal{T}_{\widehat{M}}$. Rearranging gives the result. □

Why $Q^\star_M$ and $\mathcal{T}_{\widehat{M}}$? A natural alternative is to swap the roles: bound $\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty \le \frac{1}{1-\gamma}\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty$. The RHS is the *Bellman residual* of $Q^\star_{\widehat{M}}$ in the true MDP M—a standard notion. However, this version has a subtle problem: $Q^\star_{\widehat{M}}$ is *data-dependent*, so we cannot directly apply Hoeffding's inequality to bound it. In contrast, in Eq. (6), the term $\mathcal{T}_{\widehat{M}} Q^\star_M$ involves $Q^\star_M$ which is *deterministic* (it depends only on the true MDP). This makes the concentration analysis straightforward, as we show next.

Concentration of $\mathcal{T}_{\widehat{M}} Q^\star_M$

The key lemma shows that each entry of $\mathcal{T}_{\widehat{M}} Q^\star_M$ concentrates around the corresponding entry of $Q^\star_M = \mathcal{T}_M Q^\star_M$.

Lemma 6 (Concentration via Combined Samples)**.** *For any fixed $s \in \mathcal{S}$, $a \in \mathcal{A}$, with probability at least $1-\delta$:*

$$\left|Q^\star_M(s,a) - \left(\widehat{R}(s,a) + \gamma \langle \widehat{P}(s,a), V^\star_M \rangle\right)\right| \le \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}.$$

Proof. The crucial observation is:

$$\widehat{R}(s,a) + \gamma \langle \widehat{P}(s,a), V^\star_M \rangle = \frac{1}{n} \sum_{(r,s') \in D_{s,a}} \left(r + \gamma V^\star_M(s')\right).$$

The right-hand side is the average of n i.i.d. random variables $r_i + \gamma V^\star_M(s'_i)$, each taking values in $[0, R_{\max} + \gamma V_{\max}] = [0, R_{\max}/(1-\gamma)]$, with expectation:

$$\mathbb{E}[r + \gamma V^\star_M(s')] = R(s,a) + \gamma \sum_{s'} P(s'|s,a) V^\star_M(s') = Q^\star_M(s,a).$$

The result follows directly from Hoeffding's inequality. □

The Key Insight: Instead of separately concentrating $\widehat{R}(s,a)$ and $\widehat{P}(s,a)$, we treat the combined quantity $r + \gamma V^\star_M(s')$ as a single random variable. This avoids the ℓ_∞-to-ℓ_1 conversion entirely—there is no sum over $|\mathcal{S}|$ entries! The "price" is that the range of $r + \gamma V^\star_M(s')$ is $R_{\max}/(1-\gamma) = V_{\max}$ (larger than the range of r alone), which introduces a $1/(1-\gamma)$ factor.

From Q-Function Error to Policy Loss

To translate $\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty$ into suboptimality, we use the following standard result.

Lemma 7 (Approximate Greedy Policy Loss). *If $\|Q_M^\star - \widehat{Q}\|_\infty \le \varepsilon_Q$ for some $\widehat{Q} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$, and $\pi_{\widehat{Q}}(s) = \operatorname{argmax}_a \widehat{Q}(s,a)$ is greedy with respect to $\widehat{Q}$, then:*

$$V_M^\star(s) - V_M^{\pi_{\widehat{Q}}}(s) \le \frac{2\varepsilon_Q}{1-\gamma}, \qquad \forall s \in \mathcal{S}.$$

Proof. For any state s:

$$\begin{aligned} V_M^\star(s) - V_M^{\pi_{\widehat{Q}}}(s) &= Q_M^\star(s,\pi_M^\star(s)) - Q_M^\star(s,\pi_{\widehat{Q}}(s)) + Q_M^\star(s,\pi_{\widehat{Q}}(s)) - V_M^{\pi_{\widehat{Q}}}(s) \\ &\le \big(Q_M^\star(s,\pi_M^\star(s)) - \widehat{Q}(s,\pi_M^\star(s))\big) + \big(\widehat{Q}(s,\pi_{\widehat{Q}}(s)) - Q_M^\star(s,\pi_{\widehat{Q}}(s))\big) \\ &\quad + Q_M^\star(s,\pi_{\widehat{Q}}(s)) - V_M^{\pi_{\widehat{Q}}}(s) \\ &\le 2\varepsilon_Q + Q_M^\star(s,\pi_{\widehat{Q}}(s)) - V_M^{\pi_{\widehat{Q}}}(s), \end{aligned}$$

where the second line uses $\widehat{Q}(s,\pi_{\widehat{Q}}(s)) \ge \widehat{Q}(s,\pi_M^\star(s))$ (greedy maximization). For the remaining term, let $a = \pi_{\widehat{Q}}(s)$:

$$\begin{aligned} Q_M^\star(s,a) - V_M^{\pi_{\widehat{Q}}}(s) &= R(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_M^\star(s')] - R(s,a) - \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_M^{\pi_{\widehat{Q}}}(s')] \\ &= \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V_M^\star(s') - V_M^{\pi_{\widehat{Q}}}(s')] \\ &\le \gamma \|V_M^\star - V_M^{\pi_{\widehat{Q}}}\|_\infty. \end{aligned}$$

Combining and taking the supremum over s: $\|V_M^\star - V_M^{\pi_{\widehat{Q}}}\|_\infty \le 2\varepsilon_Q + \gamma\|V_M^\star - V_M^{\pi_{\widehat{Q}}}\|_\infty$, which yields the result after rearranging. □

Combining the Pieces

Applying a union bound over all (s,a) in Lemma 6 with failure probability $\delta/(|\mathcal{S}||\mathcal{A}|)$ per pair, we get that with probability at least $1-\delta$:

$$\|Q_M^\star - \mathcal{T}_{\widehat{M}} Q_M^\star\|_\infty \le \frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}}.$$

By the contraction bound (Lemma 5):

$$\|Q_{\widehat{M}}^\star - Q_M^\star\|_\infty \le \frac{1}{1-\gamma}\cdot\frac{R_{\max}}{1-\gamma}\sqrt{\frac{1}{2n}\log\frac{2|\mathcal{S}||\mathcal{A}|}{\delta}}.$$

Applying the approximate greedy policy loss (Lemma 7) with $\widehat{Q} = Q_{\widehat{M}}^\star$ and $\varepsilon_Q = \|Q_{\widehat{M}}^\star - Q_M^\star\|_\infty$:

Analysis III Result:

$$V_M^\star(s) - V_M^{\pi_{\widehat{M}}^\star}(s) = \widetilde{O}\left(\frac{V_{\max}}{\sqrt{n}(1-\gamma)^2}\right), \qquad \forall s \in \mathcal{S}.$$

Three Sources of Horizon Dependence: The $1/(1-\gamma)^3$ factor (equivalently, $V_{\max}/(1-\gamma)^2$) comes from three distinct sources:

1. **Range of value:** The random variables $r + \gamma V_M^\star(s')$ have range $V_{\max} = R_{\max}/(1-\gamma)$, contributing one $1/(1-\gamma)$ factor.
2. **Contraction:** Translating the Bellman error $\|Q_M^\star - \mathcal{T}_{\widehat{M}} Q_M^\star\|_\infty$ to $\|Q_{\widehat{M}}^\star - Q_M^\star\|_\infty$ via contraction contributes another $1/(1-\gamma)$.
3. **Greedy policy loss:** Translating the Q-function error to suboptimality via Lemma 7 contributes the third $1/(1-\gamma)$.

In contrast, Analyses I and II only pay $1/(1-\gamma)^2$ (quadratic in horizon) because they directly bound policy evaluation errors without steps 2 and 3.

Discussion: Data Dependence

An important subtlety deserves emphasis. In Eq. (6), we compare $\mathcal{T}_{\widehat{M}}$ with $Q_M^\star$, a *deterministic* function. This is essential for applying Hoeffding's inequality.

If we had instead used the swapped version $\|Q_{\widehat{M}}^\star - Q_M^\star\|_\infty \le \frac{1}{1-\gamma}\|Q_{\widehat{M}}^\star - \mathcal{T}_M Q_{\widehat{M}}^\star\|_\infty$, we would need to concentrate:

$$(\mathcal{T}_{\widehat{M}} Q_{\widehat{M}}^\star)(s,a) - (\mathcal{T}_M Q_{\widehat{M}}^\star)(s,a) = \widehat{R}(s,a) + \gamma\langle \widehat{P}(s,a), V_{\widehat{M}}^\star\rangle - R(s,a) - \gamma\langle P(s,a), V_{\widehat{M}}^\star\rangle.$$

It would be tempting to claim that $r + \gamma V_{\widehat{M}}^\star(s')$ are i.i.d. random variables for $(r,s') \in D_{s,a}$, with expected value $R(s,a) + \gamma\langle P(s,a), V_{\widehat{M}}^\star\rangle$. But this is *not* true: $V_{\widehat{M}}^\star(s')$ is itself random and depends on the data in $D_{s,a}$! Hence Hoeffding's inequality does not apply directly.

When Can the Argument with $V_{\widehat{M}}^\star$ Still Work? In cases where the MDP's state space forms a directed acyclic graph (DAG)—for example, in finite-horizon MDPs where earlier states cannot revisit later states—the argument with $V_{\widehat{M}}^\star$ can still work, because $V_{\widehat{M}}^\star(s')$ only depends on the datasets for *later* state-action pairs, which do not include the current (s,a) under consideration. This argument is straightforward here because we have a very simple and clean data collection procedure. However, one must be extremely careful when using this argument in more realistic settings. For example, in the exploration setting, even if $V_{\widehat{M}}^\star(s')$ is estimated from datasets not including $D_{s,a}$, the outcomes in $D_{s,a}$ might have determined which later states have sufficient samples and which do not, introducing very subtle interdependence with $V_{\widehat{M}}^\star$.

Connection to Monte-Carlo Tree Search

The independence of n on $|\mathcal{S}|$ in Analysis III is the core idea that leads to *Sparse Sampling*[2], a prototype algorithm for the family of Monte-Carlo tree search (MCTS) algorithms that played a crucial role in the success of AlphaGo.

Conceptually, we could run the certainty-equivalence method with n set according to Analysis III (no dependence on $|\mathcal{S}|$). When $|\mathcal{S}|$ is large, collecting n samples for *every* (s, a) is impractical. But if we only need to know $\pi^\star(s_0)$ for some particular state s_0 (the setting of online planning), we can perform *lazy evaluation*: only generate the datasets for state-action pairs that contribute to the calculation of $V^\star_{\widehat{M}}(s_0)$, and truncate at the effective horizon $H_{\text{eff}} = O(1/(1-\gamma))$.

Roughly speaking, this requires a total of $(n \cdot |\mathcal{A}|)^{O(1/(1-\gamma))}$ samples to compute $\pi^\star(s_0)$, with *no dependence on* $|\mathcal{S}|$. This is remarkable for MDPs with enormous (or even continuous) state spaces.

Analysis IV: Best of Both Worlds

We have seen two types of analyses:

- Analyses I and II: pay a factor of $|\mathcal{S}|$ or $\sqrt{|\mathcal{S}|}$ in the bound, but only $1/(1-\gamma)^2$ horizon dependence.
- Analysis III: no $|\mathcal{S}|$ factor, but $1/(1-\gamma)^3$ horizon dependence.

Can we get the best of both worlds—neither the extra $|\mathcal{S}|$ factor nor the extra $1/(1-\gamma)$ factor?

The answer is *yes*, in the **large-sample regime**. The key insight is to use a more refined bound on the suboptimality that avoids the intermediate step of bounding $\|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty$.

Bellman Error to Decision Loss

The following lemma, due to Xie and Jiang (2020), provides a tighter translation from Bellman residuals to policy suboptimality.

Lemma 8 (Bellman Error to Decision Loss)**.** *Let $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be an MDP. For any $f \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, define $\pi_f(s) = \operatorname{argmax}_a f(s,a)$ (the greedy policy w.r.t. f). Then for any initial distribution $d_0 \in \Delta(\mathcal{S})$ and any policy π:*

[2] Kearns et al. (2002).

$$J(\pi) - J(\pi_f) \le \frac{1}{1-\gamma}\big(\mathbb{E}_{d^\pi}[\mathcal{T}f - f] + \mathbb{E}_{d^{\pi_f}}[f - \mathcal{T}f]\big),$$

where $J(\pi) = \mathbb{E}_{s\sim d_0}[V^\pi(s)]$, $\mathcal{T}$ *is the Bellman optimality operator, and* $\mathbb{E}_\mu[g]$ *is shorthand for* $\mathbb{E}_{(s,a)\sim\mu}[g(s,a)]$.

Proof. The first step uses the optimality of π_f over f: for any state s,

$$f(s,\pi_f(s)) = \max_a f(s,a) \ge f(s,\pi(s)).$$

Therefore $\mathbb{E}_{s\sim d_0,a\sim\pi_f}[f(s,a)] \ge \mathbb{E}_{s\sim d_0,a\sim\pi}[f(s,a)]$, which gives:

$$J(\pi) - J(\pi_f) \le \big(J(\pi) - \mathbb{E}_{s\sim d_0,a\sim\pi}[f(s,a)]\big) + \big(\mathbb{E}_{s\sim d_0,a\sim\pi_f}[f(s,a)] - J(\pi_f)\big).$$

For the two differences on the right-hand side, we invoke the Performance Difference Identity (Lemma 2) twice—once with policy π, and once with policy π_f:

$$J(\pi) - \mathbb{E}_{s\sim d_0,a\sim\pi}[f(s,a)] = \frac{1}{1-\gamma}\mathbb{E}_{d^\pi}[\mathcal{T}^\pi f - f],$$
$$\mathbb{E}_{s\sim d_0,a\sim\pi_f}[f(s,a)] - J(\pi_f) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_f}}[f - \mathcal{T}^{\pi_f} f].$$

The proof is completed by recognizing that $\mathcal{T}^\pi f \le \mathcal{T}f$ (the max over actions is at least as large as any specific action) and $\mathcal{T}^{\pi_f} f = \mathcal{T}f$ (since π_f is greedy w.r.t. f, it achieves the maximum). □

Applying Lemma 8

Setting $f = Q^\star_{\widehat{M}}$, $\pi = \pi^\star_M$, and noting that $\pi_f = \pi^\star_{\widehat{M}}$ (since the greedy policy w.r.t. $Q^\star_{\widehat{M}}$ is the optimal policy in $\widehat{M}$), we get for any initial distribution d_0:

$$J(\pi^\star_M) - J(\pi^\star_{\widehat{M}}) \le \frac{1}{1-\gamma}\big(\mathbb{E}_{d^{\pi^\star_M}}[\mathcal{T}_M Q^\star_{\widehat{M}} - Q^\star_{\widehat{M}}] + \mathbb{E}_{d^{\pi^\star_{\widehat{M}}}}[Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}]\big).$$

Since the expectations are bounded by the ℓ_∞-norm (both $d^{\pi^\star_M}$ and $d^{\pi^\star_{\widehat{M}}}$ are distributions), we immediately obtain:

$$\|V^\star_M - V^{\pi^\star_{\widehat{M}}}_M\|_\infty \le \frac{2}{1-\gamma}\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty, \tag{7}$$

by choosing d_0 as the point mass on any state s_0.

Comparison with Analysis III: In Analysis III, the path was: $\|Q^\star_M - \mathcal{T}_{\widehat{M}} Q^\star_M\|_\infty \xrightarrow{\times 1/(1-\gamma)} \|Q^\star_{\widehat{M}} - Q^\star_M\|_\infty \xrightarrow{\times 2/(1-\gamma)} \|V^\star_M - V^{\pi^\star_{\widehat{M}}}_M\|_\infty$, paying a total factor of $2/(1-\gamma)^2$. In Eq. (7), we go directly from the Bellman residual to the suboptimality, paying only $2/(1-\gamma)$. This saves one horizon factor!
The catch is that the Bellman residual in Eq. (7) involves $Q^\star_{\widehat{M}}$, which is *data-dependent*.

Bounding $\|Q^\star_{\widehat{M}} - \mathcal{T}_M Q^\star_{\widehat{M}}\|_\infty$

To handle the data dependence of $Q^\star_{\widehat{M}}$, we use a decomposition trick. For any (s,a):

$$\begin{aligned}
|Q^\star_{\widehat{M}}(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)| &= |(\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}})(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)| \\
&\le \underbrace{|(\mathcal{T}_{\widehat{M}} Q^\star_M)(s,a) - (\mathcal{T}_M Q^\star_M)(s,a)|}_{\text{(A): ``easy'' term}} \\
&\quad + \underbrace{|(\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}})(s,a) - (\mathcal{T}_{\widehat{M}} Q^\star_M)(s,a) + (\mathcal{T}_M Q^\star_M)(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)|}_{\text{(B): ``burn-in'' term}}.
\end{aligned}$$

The idea is to replace $Q^\star_{\widehat{M}}$ with the deterministic $Q^\star_M$: term (A) is the same object we analyzed in Analysis III and does not depend on $|\mathcal{S}|$; term (B) is the error incurred by this replacement.

Bounding term (A). This is identical to the analysis in Section "Analysis III". By Lemma 6 and a union bound:

$$\max_{s,a} |(\mathcal{T}_{\widehat{M}} Q^\star_M)(s,a) - (\mathcal{T}_M Q^\star_M)(s,a)| = \widetilde{O}\left(\frac{V_{\max}}{\sqrt{n}}\right).$$

Bounding term (B). Expanding the Bellman operators:

$$\begin{aligned}
&|(\mathcal{T}_{\widehat{M}} Q^\star_{\widehat{M}})(s,a) - (\mathcal{T}_{\widehat{M}} Q^\star_M)(s,a) + (\mathcal{T}_M Q^\star_M)(s,a) - (\mathcal{T}_M Q^\star_{\widehat{M}})(s,a)| \\
&= \gamma \left| \langle \widehat{P}(s,a), V^\star_{\widehat{M}} - V^\star_M \rangle - \langle P(s,a), V^\star_{\widehat{M}} - V^\star_M \rangle \right| \\
&= \gamma \left| \langle \widehat{P}(s,a) - P(s,a), V^\star_{\widehat{M}} - V^\star_M \rangle \right| \\
&\le \gamma \|\widehat{P}(s,a) - P(s,a)\|_1 \cdot \|V^\star_{\widehat{M}} - V^\star_M\|_\infty.
\end{aligned}$$

Now we bound each factor separately:

- $\|\widehat{P}(s,a) - P(s,a)\|_1$ can be bounded using the ℓ_1 concentration from Analysis II: $\widetilde{O}(\sqrt{|\mathcal{S}|/n})$.

- $\|V^\star_{\widehat{M}} - V^\star_M\|_\infty$ can be bounded using Analysis III: $\widetilde{O}(V_{\max}/(\sqrt{n}(1-\gamma)))$.

Therefore, term (B) scales as:

$$\text{(B)} = \widetilde{O}\left(\sqrt{\frac{|\mathcal{S}|}{n}} \cdot \frac{V_{\max}}{\sqrt{n}(1-\gamma)}\right) = \widetilde{O}\left(\frac{\sqrt{|\mathcal{S}|} \cdot V_{\max}}{n(1-\gamma)}\right).$$

The Burn-In Term: Term (A) scales as $O(1/\sqrt{n})$, while term (B) is the product of two $O(1/\sqrt{n})$ quantities and hence scales as $O(1/n)$. When n is sufficiently large, the $O(1/n)$ burn-in term is dominated by the $O(1/\sqrt{n})$ main term. The burn-in term has worse dependencies on $|\mathcal{S}|$ and $1/(1-\gamma)$ compared to the main term, so "sufficiently large n" means n must exceed some threshold that depends on $|\mathcal{S}|$ and $1/(1-\gamma)$.

The Final Bound

Combining terms (A) and (B) with Eq. (7):

$$\|V^\star_M - V^{\pi^\star_{\widehat{M}}}_M\|_\infty \le \frac{2}{1-\gamma}\left[\widetilde{O}\left(\frac{V_{\max}}{\sqrt{n}}\right) + \widetilde{O}\left(\frac{\sqrt{|\mathcal{S}|} \cdot V_{\max}}{n(1-\gamma)}\right)\right].$$

In the *large-sample regime* where n is sufficiently large relative to $|\mathcal{S}|$ and $1/(1-\gamma)$ (so that the $1/n$ burn-in term is dominated by the $1/\sqrt{n}$ term):

Analysis IV Result (Large-Sample Regime):

$$V^\star_M(s) - V^{\pi^\star_{\widehat{M}}}_M(s) = \widetilde{O}\left(\frac{V_{\max}}{\sqrt{n}(1-\gamma)}\right), \qquad \forall s \in \mathcal{S}.$$

This achieves neither the extra $|\mathcal{S}|$ factor from Analyses I/II nor the extra $1/(1-\gamma)$ factor from Analysis III.

Further Improvements

The bounds can be further improved by replacing Hoeffding's inequality with **Bernstein's inequality** (Theorem 2 from Lecture 5), which provides sharper concentration bounds when the variance of the random variables is substantially smaller than their range squared.

In our setting, the range of random variables in the concentration of $\langle \widehat{P}(s,a) - P(s,a), V\rangle$ is $V_{\max}$, so the worst-case variance is $O(V^2_{\max})$. However, for certain V (e.g., $V = V^\pi_M$), such variance cannot be large for all (s,a) simultaneously as it adds up to $O(V^2_{\max})$ along the

occupancy of π. Leveraging this property leads to further improved sample complexities.[3]

Summary

We analyzed the tabular certainty-equivalence algorithm for reinforcement learning and derived four progressively refined sample complexity bounds. The following table summarizes the results:

Summary of Suboptimality Bounds:

Analysis	Suboptimality Bound	Key Idea
I: Naive	$\widetilde{O}\left(\frac{\lvert\mathcal{S}\rvert \cdot V_{\max}}{\sqrt{n}(1-\gamma)}\right)$	Entry-wise Hoeffding + union bound
II: Improved $\lvert\mathcal{S}\rvert$	$\widetilde{O}\left(\frac{\sqrt{\lvert\mathcal{S}\rvert} \cdot V_{\max}}{\sqrt{n}(1-\gamma)}\right)$	ℓ_1 concentration via signed vectors
III: No $\lvert\mathcal{S}\rvert$	$\widetilde{O}\left(\frac{V_{\max}}{\sqrt{n}(1-\gamma)^2}\right)$	Contraction + direct Hoeffding
IV: Best of both	$\widetilde{O}\left(\frac{V_{\max}}{\sqrt{n}(1-\gamma)}\right)$	Bellman residual bound (large n)

Key Takeaways:

- **Analyses I and II** work through the Simulation Lemma (model error $\to$ value error). The $|\mathcal{S}|$ factor arises from the ℓ_∞-to-ℓ_1 conversion in bounding $\|\widehat{P} - P\|_1$. Analysis II improves this via a direct ℓ_1 concentration bound.
- **Analysis III** bypasses the Simulation Lemma entirely. By working with $Q^\star_M$ and the contraction property, it avoids any dependence on $|\mathcal{S}|$, at the cost of an extra $1/(1-\gamma)$ factor.
- **Analysis IV** combines both approaches. In the large-sample regime, it achieves no extra $|\mathcal{S}|$ and no extra $1/(1-\gamma)$, at the cost of a $O(1/n)$ burn-in term.
- The tradeoff between $|\mathcal{S}|$ and $1/(1-\gamma)$ is a recurring theme in RL theory, arising from the tension between bounding model errors (which scale with $|\mathcal{S}|$) and bounding value function errors (which scale with $1/(1-\gamma)$).

[3] See Azar et al. (2013) and Section 2.3 of Agarwal et al. (2022).

References

Alekh Agarwal, Nan Jiang, Sham Kakade, and Wen Sun. *Reinforcement Learning: Theory and Algorithms*. 2022. https://rltheorybook.github.io/.

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.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002.

Michael Kearns, Yishay Mansour, and Andrew Y Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Machine Learning*, 49(2-3): 193–208, 2002.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.