Lecture 2b: Generative Models - Tabular Certainty-Equivalence RL

Qi Zhang

Last Updated: September 2025

1 Concentration Inequalities and Union Bound

Theorem 1 (The union bound). *Let $E_1, E_2, \ldots, E_n$ be a collection of events. Then,*

$$\Pr\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \Pr\left(E_i\right).$$

Additionally, if $E_1, E_2, \ldots$ is a countably infinite collection of events, then:

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \Pr\left(E_i\right).$$

Theorem 2 (Hoeffding's inequality). *Let n be a constant and $X_1, \ldots, X_n$ be independent random variables on $\mathbb{R}$ such that X_i is bounded in $[a_i, b_i]$. Let $S_n := \sum_{i=1}^{n} X_i$. Then for all $t > 0$,*

$$\Pr\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) \leq e^{-2t^2/\sum_{i=1}^{n}(b_i - a_i)^2}.$$

Remarks:

- Applying Theorem 2 to $\{-X_i\}_{i=1}^{n}$, we obtain the other one-sided inequality: $\Pr\left(S_n - \mathbb{E}\left[S_n\right] \leq -t\right) \leq e^{-2t^2/\sum_{i=1}^{n}(b_i - a_i)^2}$. Applying the union bound to both one-sided inequalities, we obtain the often used two-sided bound: $\Pr\left(\left|S_n - \mathbb{E}\left[S_n\right]\right| \geq t\right) \leq 2e^{-2t^2/\sum_{i=1}^{n}(b_i - a_i)^2}$.

- When all variables share the same support $[a, b]$ and we compare the empirical average with the true mean, the two-sided bound reduces to

$$\Pr\left(\left|\frac{S_n}{n} - \frac{\mathbb{E}\left[S_n\right]}{n}\right| \geq t\right) \leq 2e^{-2nt^2/(b-a)^2}.$$

  Setting $\delta := 2e^{-2nt^2/(b-a)^2}$ to be the probability of failure and solving it for t, we can rephrase the result as follows:

$$\text{With probability} \geq 1 - \delta, \quad \left|\frac{S_n}{n} - \frac{\mathbb{E}\left[S_n\right]}{n}\right| \leq (b - a)\sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}.$$

- The number of variables, n, is a constant in the theorem statement. When n is a random variable, Hoeffding's inequality still applies if n does not depend on the realization of $X_1, \ldots, X_n$. Otherwise, Hoeffding's inequality can be used with the union bound over possible realizations of n.

- We refer to section 6.3.4 of this note for an example of using Hoeffding's inequality with the union bound.

2 Tabular Certainty-Equivalence with Generative Models

Certainty-equivalence is a *model-based* method that estimates unknown quantities of the MDP of interest from data and performs policy optimization with the estimation as if it were true. Certainty-equivalence explicitly stores an estimated MDP and performs planning after all the data are collected.

This note focuses on the setting where (only) the transition function P of an finite-horizon MDP $M = (\mathcal{S}, \mathcal{A}, P, R, H)$ is unknown but can be queried as a generative model to draw samples $s' \sim P_h(s,a)$ for any (s,a,h). As the problem is still non-trivial even when reward function R is known, we therefore assume it is known for simplicity. To identify an (near-)optimal policy for M, we can estimate P from samples from it. If we query the (s,a,h) tuple $n_{s,a,h}$ times and get next-state samples $\{s'_i\}_{i=1}^{n_{s,a,h}}$, *tabular* certainty-equivalence estimates $P_h(s,a)$ as

$$\widehat{P}_h\left(s'|s,a\right) = \frac{1}{n_{s,a,h}} \sum_{i=1}^{n_{s,a,h}} \mathbf{1}\left[s'_i = s'\right].$$

We assume every (s,a,h) gets the same number of next-state samples and write $n \equiv n_{s,a,h}$. This way, we obtain the estimated MDP $\widehat{M} := (\mathcal{S}, \mathcal{A}, \widehat{P}, R, H)$ and its optimal policy $\widehat{\pi}$ as a function of n. We are interested in providing high probability guarantees for the quality of $\widehat{\pi}$ in the original MDP. Specifically, we aim to show, when n is large, $V_1^{M,\pi^*}(s_1) - V_1^{M,\widehat{\pi}}(s)$ is small with high probability for any initial state s, where π^* is an optimal policy for M.

2.1 Coarse analysis

Intuitively, $V^{M,\pi^*} - V^{M,\widehat{\pi}}$ is small because, when n is large, $\widehat{P} \approx P$ and therefore $\widehat{M} \approx M$, so $\widehat{\pi}$ that is optimal for $\widehat{M}$ should be near-optimal for M. This reasoning can be formalized using the following error decomposition:

$$\begin{aligned}
&V_1^{M,\pi^*}(s) - V_1^{M,\widehat{\pi}}(s) \\
=&V_1^{M,\pi^*}(s) - V_1^{\widehat{M},\pi^*}(s) + \underbrace{V_1^{\widehat{M},\pi^*}(s) - V_1^{\widehat{M},\widehat{\pi}}(s)}_{\leq 0 \text{ as } \widehat{\pi} \text{ is optimal for } \widehat{M}} + V_1^{\widehat{M},\widehat{\pi}}(s) - V_1^{M,\widehat{\pi}}(s) \\
\leq& \underbrace{V_1^{M,\pi^*}(s) - V_1^{\widehat{M},\pi^*}(s)}_{\text{(i)}} + \underbrace{V_1^{\widehat{M},\widehat{\pi}}(s) - V_1^{M,\widehat{\pi}}(s)}_{\text{(ii)}}.
\end{aligned}$$

The simulation lemma. Terms (i) and (ii) are small because of the same reason: whenever the two MDPs M and $\widehat{M}$ close (in terms of their transition functions), their value functions are also close. This statement is made precise by Lemma 3 known as the simulation lemma, where we use the following notation: P_h is treated as a matrix of shape $|\mathcal{S} \times \mathcal{A}| \times \mathcal{S}$ with entries $P_h(s'|s,a)$ where (s,a) indexing rows and s' indexing columns; value function V_h is treated as a column vector of shape $|\mathcal{S}|$ with its s-th entry being $V_h(s)$. Therefore, $P_h V_{h+1}$ is a matrix-vector product yielding a vector of size $|\mathcal{S} \times \mathcal{A}|$ with the (s,a)-th entry

$$[P_h V_{h+1}]_{(s,a)} = \sum_{s' \in \mathcal{S}} P_h(s'|s,a) V_{h+1}(s') = \mathbb{E}_{s' \sim P_h(s,a)}[V_{h+1}(s')].$$

Lemma 3 (Simulation lemma)**.** *Let MDPs M and $\widehat{M}$ differ only by their transition functions P and $\widehat{P}$. For any policy π and (s,h), we have*

$$V_h^{M,\pi}(s) - V_h^{\widehat{M},\pi}(s) = \sum_{i=h}^{H} \mathbb{E}_{(s_i,a_i)\sim M,\pi|s_h=s}\left[\left[\left(P_i - \widehat{P}_i\right) V_{i+1}^{\widehat{M},\pi}\right]_{(s_i,a_i)}\right]$$
$$= \sum_{i=h}^{H} \mathbb{E}_{(s_i,a_i)\sim \widehat{M},\pi|s_h=s}\left[\left[\left(P_i - \widehat{P}_i\right) V_{i+1}^{M,\pi}\right]_{(s_i,a_i)}\right].$$

A proof of Lemma 3 is deferred to Section 3 as an optional reading. Lemma 3 quantifies the discrepancy between the value functions by the discrepancy between the transition functions of the two MDPs, which is precisely what we need.

Concentration of $\widehat{P} \approx P$. We will assume reward is bounded and, without loss of generality, bounded in $[0,1]$, i.e., $R_h(s,a) \in [0,1]$ for any (s,a,h). Therefore, the value function is bounded as $V_h^\pi(s) \in [0,H]$ for any π, h and therefore $\|V_h^\pi\|_\infty := \max_s V_h^\pi(s) \le H$. Noting $\left[\left(P_h - \widehat{P}_h\right) V\right]_{(s,a)} \le \left\|P_h(s,a) - \widehat{P}_h(s,a)\right\|_1 \cdot \|V\|_\infty$ by Cauchy–Schwarz (the dual norm form), it therefore suffices to provide a high probability guarantee of the ℓ_1 norm when n is large, which we give here using Hoeffding's inequality with the union bound:

(1) Fix any (s,a,h,s') and define random variables $X_i := \mathbf{1}[s_i' = s'], i = 1,\ldots,n$ for the n next-state samples. Show in your **Homework 2's 3a** that, by applying Hoeffding's inequality, we have: With probability $\ge 1-\delta$, $\left|\widehat{P}_h\left(s'|s,a\right) - P_h\left(s'|s,a\right)\right| \le \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)}$.

(2) Fix any (s,a,h). Show in your **Homework 2's 3b** that, by applying the union bound over all $s' \in \mathcal{S}$ on top of step (1), we have: With probability $\ge 1-\delta$,

$$\left\|P_h(s,a) - \widehat{P}_h(s,a)\right\|_1 = \sum_{s'\in\mathcal{S}} \left|\widehat{P}_h\left(s'|s,a\right) - P_h\left(s'|s,a\right)\right| \le |\mathcal{S}|\sqrt{\frac{1}{2n}\ln\left(\frac{2|\mathcal{S}|}{\delta}\right)}.$$

Hint: To achieve the failure probability of δ, we split the δ in step (1) evenly among all s'.

(3) We then apply the union bound over all (s,a,h) on top of step (2). To achieve failure probability of δ, we split the δ in step (2) evenly among all (s,a,h): With probability $\ge 1-\delta$,

$$\left\|P_h(s,a) - \widehat{P}_h(s,a)\right\|_1 \le |\mathcal{S}|\sqrt{\frac{1}{2n}\ln\left(\frac{2|\mathcal{S}|}{\delta/(|\mathcal{S}||\mathcal{A}|H)}\right)} \quad \textit{for all} \quad (s,a,h).$$

Putting it together. We are ready to make the following statement for terms (i) and (ii) : With probability $\geq 1-\delta$,

$$
\begin{aligned}
\text{(i)} :=& V_1^{M,\pi^*}(s) - V_1^{\widehat{M},\pi^*}(s) \\
=& \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim M,\pi^*|s_1=s}\left[\left[\left(P_h - \widehat{P}_h\right)V_{h+1}^{\widehat{M},\pi^*}\right]_{(s_h,a_h)}\right] && \textbf{(3c)} \\
\leq& \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim M,\pi^*|s_1=s}\left[\left\|P_h(s_h,a_h) - \widehat{P}_h(s_h,a_h)\right\|_1 \cdot \left\|V_{h+1}^{\widehat{M},\pi^*}\right\|_\infty\right] && \textbf{(3d)} \\
\leq& \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim M,\pi^*|s_1=s}\left[|\mathcal{S}|\sqrt{\frac{1}{2n}\ln\left(\frac{2|\mathcal{S}|}{\delta/(|\mathcal{S}||\mathcal{A}|H)}\right)} \cdot H\right] && \textbf{(3e)} \\
=& |\mathcal{S}|H^2\sqrt{\frac{1}{2n}\ln\left(\frac{2|\mathcal{S}|^2|\mathcal{A}|H}{\delta}\right)} =: \epsilon(n,\delta)
\end{aligned}
$$

and (ii) $\leq \epsilon(n,\delta)$ for the same reason.

To achieve an total error for ϵ, we can choose n large enough such that $\epsilon(n,\delta) \leq \epsilon/2$, which leads to Proposition 1.

Proposition 1. *Given any (ϵ,δ) choosing n large enough such that $\epsilon(n,\delta) \leq \epsilon/2$, with probability $\geq 1-\delta$, we have $V_1^{M,\pi^*}(s) - V_1^{M,\widehat{\pi}}(s) \leq \epsilon$ for all initial state s.*

3 Proof of Lemma 3 (The Simulation Lemma)

We will prove the first equality below; the second can be obtained by the relabeling of $(M,\widehat{M}) \to (\widehat{M},M)$. The key idea is to unroll the timesteps using the Bellman equations.

Without loss of generality, we will show the case of $h=1$. Writing $s_1 \equiv s$, we have

$$
\begin{aligned}
& V_1^{M,\pi}(s_1) - V_1^{\widehat{M},\pi}(s_1) \\
=& \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[Q_1^{M,\pi}(s_1,a_1)\right] - \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[Q_1^{\widehat{M},\pi}(s_1,a_1)\right] \\
=& \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[R_1(s_1,a_1) + \left[P_1V_2^{M,\pi}\right]_{(s_1,a_1)}\right] - \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[R_1(s_1,a_1) + \left[\widehat{P}_1V_2^{\widehat{M},\pi}\right]_{(s_1,a_1)}\right] \\
=& \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[\left[P_1V_2^{M,\pi} - \widehat{P}_1V_2^{\widehat{M},\pi}\right]_{(s_1,a_1)}\right] \\
=& \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[\left[P_1V_2^{M,\pi} - P_1V_2^{\widehat{M},\pi} + P_1V_2^{\widehat{M},\pi} - \widehat{P}_1V_2^{\widehat{M},\pi}\right]_{(s_1,a_1)}\right] \\
=& \underbrace{\mathbb{E}_{a_1\sim\pi_1(s_1)}\left[\left[P_1\left(V_2^{M,\pi} - V_2^{\widehat{M},\pi}\right)\right]_{(s_1,a_1)}\right]}_{\text{(i)}} + \underbrace{\mathbb{E}_{a_1\sim\pi_1(s_1)}\left[\left[\left(P_1 - \widehat{P}_1\right)V_2^{\widehat{M},\pi}\right]_{(s_1,a_1)}\right]}_{\text{(ii)}}
\end{aligned}
$$

At this point, note term (ii) is the very first summand of the RHS in the lemma's first equality (i.e., our goal). For term (i), we have

$$
\begin{aligned}
\text{(i)} =& \mathbb{E}_{a_1\sim\pi_1(s_1)}\left[\mathbb{E}_{s_2\sim P_1(s_1,a_1)}\left[V_2^{M,\pi}(s_2) - V_2^{\widehat{M},\pi}(s_2)\right]\right] \\
=& \mathbb{E}_{s_2\sim M,\pi|s_1}\left[V_2^{M,\pi}(s_2) - V_2^{\widehat{M},\pi}(s_2)\right]
\end{aligned}
$$

where $V_2^{M,\pi}(s_2) - V_2^{\widehat{M},\pi}(s_2)$ can be expanded the same way as $V_1^{M,\pi}(s_1) - V_1^{\widehat{M},\pi}(s_1)$ above. Recursively expanding all the way down to the last timestep H, we obtain

$$\begin{aligned}
&V_1^{M,\pi}(s_1) - V_1^{\widehat{M},\pi}(s_1)\\
=&\underbrace{\mathbb{E}_{s_H,a_H\sim M,\pi|s_1}\left[\left[P_H\left(V_{H+1}^{M,\pi} - V_{H+1}^{\widehat{M},\pi}\right)\right]_{(s_H,a_H)}\right]}_{=0 \text{ as } V_{H+1}^{M,\pi}=V_{H+1}^{\widehat{M},\pi}=0} + \sum_{h=1}^{H}\mathbb{E}_{(s_h,a_h)\sim M,\pi|s_1}\left[\left[\left(P_h - \widehat{P}_h\right)V_{h+1}^{\widehat{M},\pi}\right]_{(s_h,a_h)}\right]
\end{aligned}$$

which completes the proof.