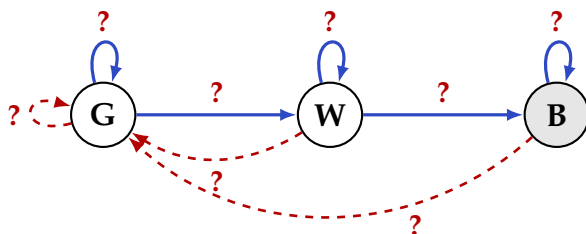


Lecture 8: Planning versus Learning, and Certainty Equivalence

How to use these notes. Board content in board order: figures, equations, tables, one-line statements. Today's model: Lecture 3's machine, with P and r hidden from the learner. Assumed known: Lecture 6's policy iteration with the tie rule; Lecture 5's contraction and residual certificate; Lecture 7's gap Δ , greedy-policy bound, and generalized policy iteration. Reading: [Sutton and Barto \(2018, Section 8.1, and the paragraph on the certainty-equivalence estimate in Section 6.3\)](#). Throughout the course, the notes draw on the public materials of Stanford CS 234 ([Brunskill, 2026](#)), UIUC CS 443 ([Jiang, 2026](#)), and Cornell CS 4789 ([Sun, 2025](#)); some examples are adapted from them.

The Model Taken Away

Lectures 3–7 read P and r off the picture. Same machine today; the labels are gone.



$r(s, a)$	run	repair
G	?	?
W	?	?
B	?	?

The arrows are still drawn; their probabilities are not. Solid: run.

Dashed: repair.

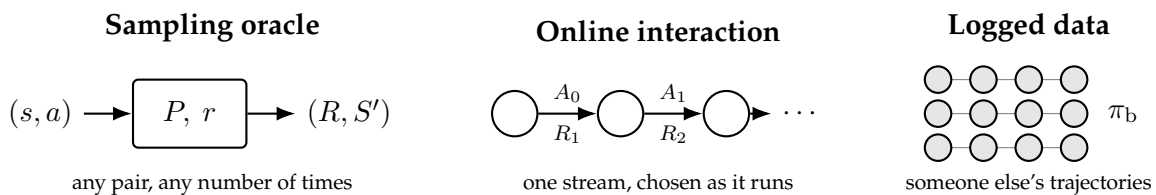
The profits too. Known: $\mathcal{S}, \mathcal{A}, \gamma = 0.9$, and that the machine starts good.

What the learner sees instead.

- One transition: choose (s, a) , observe (R, S') , R the week's profit and $S' \sim P(\cdot \mid s, a)$.
Lecture 2's step
- A dataset: a bag of four-tuples $D = \{(s_i, a_i, r_i, s'_i)\}_{i=1}^N$; $n(s, a)$ of them at the pair (s, a) .
- Hidden: the $|\mathcal{S}|^2|\mathcal{A}| + |\mathcal{S}||\mathcal{A}| = 18 + 6 = 24$ numbers of (P, r) . Lectures 3–7 used every one of them.

Access	The learner may	The data	Today
Sampling oracle (a simulator)	name any (s, a) and receive one (R, S') , as often as it likes	n samples at every pair, the learner's choice	the algorithm
Online interaction	act in the one running machine, see where it goes, choose again	one stream; the counts fall where the policy goes	later in the unit
Logged data	read trajectories that someone else's policy collected; no new experiments	fixed; counts depend on the collecting policy	later in the course

Lecture 1's table, with the model column gone. Three ways to hold a dataset; the same twenty-four unknowns behind each.



The three accesses as pictures. A simulator answers questions; the running machine answers only the question you are in; a log answers what its policy π_b once asked.

With P and r replaced by samples, what survives of Lectures 3–7, how wrong can its answer be, and how much data buys a right one?

Today's targets.

1. Three kinds of access: a sampling oracle, the running machine, a log. Today's algorithm uses the sampling oracle.
2. Certainty equivalence: count, average, then Lecture 6's policy iteration on the estimate. By hand with $n = 3$ samples per pair: two datasets, two answers (π^* ; "repair when worn").
3. Derivation in three lines: the estimate's optimal policy loses at most $\varepsilon(\pi^*) + \varepsilon(\hat{\pi})$, two evaluation errors; under the gap, exact. Simulation lemma (stated): evaluation error as Lecture 5's certificate in the wrong model.
4. Hoeffding (stated): a fraction is within $\sqrt{\log(2/\delta)/(2n)}$ of its probability. The machine's decision rides on one fraction, wrong below 0.175; $n = 60$ certifies it; by computer the policy is wrong with probability 0.30 at $n = 3$ and 0.003 at $n = 60$.

Certainty Equivalence

Certainty equivalence (Sutton and Barto, 2018, Section 6.3), (Kearns and Singh, 2002).

- Input: for every (s, a) , n samples (R_i, S'_i) from the sampling oracle.
- Count and average:

$$\hat{P}(s' | s, a) = \frac{\#\{i : S'_i = s'\}}{n}, \quad \hat{r}(s, a) = \frac{1}{n} \sum_{i=1}^n R_i.$$

- Plan in the estimate: $\widehat{M} = (\mathcal{S}, \mathcal{A}, \hat{P}, \hat{r}, \gamma)$; run Lecture 6's policy iteration in $\widehat{M}$, returning its optimal policy $\hat{\pi} = \pi_{\widehat{M}}^*$ and values $\hat{V} = V_{\widehat{M}}^*$.
- Deploy $\hat{\pi}$ in the true machine M .

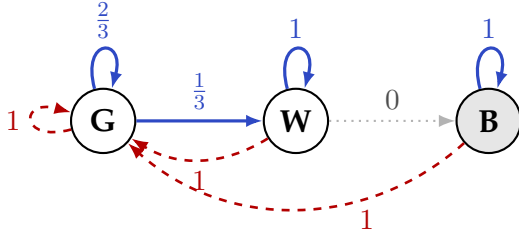
"Certainty": the estimate is treated as the truth. Nothing after the first step is new; the planner is Lecture 6's.

- 24 numbers from $6n$ samples: 18 probabilities as fractions, 6 rewards as averages. Names later: model-based learning.
- On the machine the rewards repeat, so $\hat{r} = r$ after one sample; four of the six rows of P land in one state every time, so $\hat{P} = P$ there.
- The noise sits in two fractions: $\hat{p}_G = \hat{P}(W | G, \text{run})$ and $\hat{p}_W = \hat{P}(B | W, \text{run})$, both $\frac{1}{3}$ in truth. The learner finds this out from the counts.

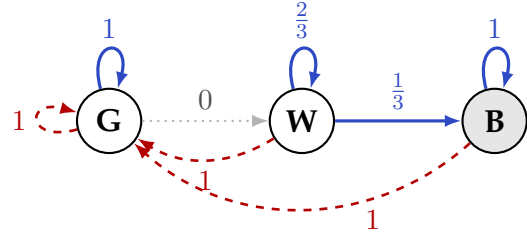
By hand, $n = 3$: one column per pair, one row per sample, an entry a profit and a next state. Two datasets of the same size:

		(G, run)	(W, run)	(B, run)	(G, repair)	(W, repair)	(B, repair)
A	$i = 1$	8, G	4, W	0, B	-10, G	-10, G	-10, G
	$i = 2$	8, W	4, W	0, B	-10, G	-10, G	-10, G
	$i = 3$	8, G	4, W	0, B	-10, G	-10, G	-10, G
	$\hat{P}(\cdot s, a)$	$(\frac{2}{3}, \frac{1}{3}, 0)$	$(0, 1, \boxed{0})$	$(0, 0, 1)$	$(1, 0, 0)$	$(1, 0, 0)$	$(1, 0, 0)$
B	$i = 1$	8, G	4, W	0, B	-10, G	-10, G	-10, G
	$i = 2$	8, G	4, B	0, B	-10, G	-10, G	-10, G
	$i = 3$	8, G	4, W	0, B	-10, G	-10, G	-10, G
	$\hat{P}(\cdot s, a)$	$(1, \boxed{0}, 0)$	$(0, \frac{2}{3}, \frac{1}{3})$	$(0, 0, 1)$	$(1, 0, 0)$	$(1, 0, 0)$	$(1, 0, 0)$
$\hat{r}(s, a), \text{ both}$		8	4	0	-10	-10	-10

Datasets **A** and **B**, 18 transitions each. Gray columns: exact after one sample, identical in both. Bold: the one rare event each dataset saw. Boxed: the event it never saw, estimated impossible; the truth there is $\frac{1}{3}$.



$\widehat{M}_A$: the boxed 0 is the dotted arrow; a worn machine never breaks.



$\widehat{M}_B$: the boxed 0 is the dotted arrow; a good machine never wears.

Example, dataset A. Plan in $\widehat{M}_A$: $\widehat{V}$ and $\widehat{\pi}$?

W never breaks: run forever, $\widehat{V}(W) = 4/(1 - 0.9) = 40$. G: $\widehat{V}(G) = 8 + 0.9(\frac{2}{3}\widehat{V}(G) + \frac{1}{3} \cdot 40)$, so $0.4\widehat{V}(G) = 20$, $\widehat{V}(G) = 50$. B: repair $-10 + 0.9 \cdot 50 = 35$ beats run 0.

$\widehat{V} = (50, 40, 35)$, $\widehat{\pi} = \text{repair when broken} = \pi^*$. Values off by (5.70, 7.59, 5.13); the policy exactly right.

Example, dataset B. Plan in $\widehat{M}_B$.

G never wears: $\widehat{V}(G) = 8/(1 - 0.9) = 80$; repair pays $-10 + 72 = 62$ from W and from B.

W: run forever earns $(4 + 0.9 \cdot \frac{1}{3} \cdot 62)/(1 - 0.9 \cdot \frac{2}{3}) = 56.5 < 62$: repair.

$\widehat{V} = (80, 62, 62)$, $\widehat{\pi} = \text{repair when worn}$: worth (38.46, 24.62, 24.62) in the true machine, loss 7.79. Chance of this G row: $(2/3)^3 = 8/27 = 0.30$.

	$\widehat{p}_G$	$\widehat{p}_W$	$\widehat{V} = V_{\widehat{M}}^*$	$\widehat{\pi} = \pi_{\widehat{M}}^*$	$\ \widehat{V} - V^*\ _\infty$	loss $\ V^* - V_{\widehat{\pi}}\ _\infty$
dataset A	$\frac{1}{3}$	0	(50, 40, 35)	repair when broken	7.59	0
dataset B	0	$\frac{1}{3}$	(80, 62, 62)	repair when worn	35.70	7.79
the truth	$\frac{1}{3}$	$\frac{1}{3}$	(44.30, 32.41, 29.87)	repair when broken	0	0

Two datasets of the same size, one truth. Both estimates are wrong about the values; one is right about the policy.

The two answers differ at one state, W. Run Lecture 6's policy iteration from π^* in each of the three machines: evaluate it, then test the other action. At G and at B the test fails in all three; at W:

	the truth M	$\widehat{M}_A$	$\widehat{M}_B$
estimated impossible	nothing	a worn machine breaks	a good machine wears
a good machine, $V^{\pi^*}(G)$	44.30	50	80
keep running at W, $V^{\pi^*}(W)$	32.41	40	56.5
repair at W instead, $-10 + 0.9 V^{\pi^*}(G)$	29.87	35	62
policy iteration at W	keeps running	keeps running	switches to repair

The flip. Each column evaluates the same π^* in its named model. Repair at W buys a good machine, so it is worth $-10 + 0.9 V(G)$: the more a good machine seems worth, the better repair looks. B never saw one wear, prices it at 80, and buys.

- Both estimates put both numbers above the truth. A's missing breakdown makes W a safe place to keep running; B's missing wear makes G a machine worth buying.
- One entry off by $\frac{1}{3}$ in each; where it sits decides, not its size. Lecture 7 again: the policy rides on the order of two numbers, not on the numbers.

What a Wrong Model Costs

Dataset B's values were off by 35.70 and its policy lost 7.79. The general rule, in three lines.

Lemma: evaluation error to decision loss (Kearns and Singh, 2002), (Agarwal et al., 2022, Chapter 2). Write $\widehat{V}^\pi := V_{\widehat{M}}^\pi$ for a fixed policy's value in the estimated model; V^π is its value in the true model. The optimal estimated value is still $\widehat{V} = V_{\widehat{M}}^*$. Let $\varepsilon(\pi) := \|\widehat{V}^\pi - V^\pi\|_\infty$ be the evaluation error, and $\widehat{\pi} = \pi_{\widehat{M}}^*$. For every s ,

$$\begin{aligned}
 V^{\widehat{\pi}}(s) &\geq \widehat{V}^{\widehat{\pi}}(s) - \varepsilon(\widehat{\pi}) && \widehat{\pi}'s \text{ true value, off from its value in } \widehat{M} \text{ by at most } \varepsilon(\widehat{\pi}) \\
 &\geq \widehat{V}^{\pi^*}(s) - \varepsilon(\widehat{\pi}) && \widehat{\pi} \text{ is optimal in } \widehat{M}, \text{ so in } \widehat{M} \text{ it beats } \pi^* \\
 &\geq V^{\pi^*}(s) - \varepsilon(\pi^*) - \varepsilon(\widehat{\pi}) && \pi^*'s \text{ value in } \widehat{M}, \text{ off from the truth by at most } \varepsilon(\pi^*) \quad \square
 \end{aligned}$$

$$\|V^* - V^{\widehat{\pi}}\|_\infty \leq \varepsilon(\pi^*) + \varepsilon(\widehat{\pi}) \leq 2 \max_\pi \varepsilon(\pi); \quad \varepsilon(\pi^*) + \varepsilon(\widehat{\pi}) < \Delta \Rightarrow \widehat{\pi} = \pi^*.$$

The last step is Lecture 7's: a loss below the gap $\Delta = 7.79$ can only be 0.

	$\widehat{\pi}$	$\varepsilon(\pi^*)$	$\varepsilon(\widehat{\pi})$	bound $\varepsilon(\pi^*) + \varepsilon(\widehat{\pi})$	true loss
dataset A	repair when broken	7.59	7.59	15.19	0
dataset B	repair when worn	35.70	41.54	77.23	7.79

The lemma on the two datasets. In $\widehat{M}_B$, π^* is worth (80, 56.5, 62) and repair when worn (80, 62, 62), against (44.30, 32.41, 29.87) and (38.46, 24.62, 24.62) in truth: both look better than they are, the wrong one by more. The two errors the proof pays for.

- Two errors, not one: π^* may look worse in $\widehat{M}$ than it is, $\widehat{\pi}$ may look better. Lecture 7's greedy bound had the same 2 for the same reason.
- A worst case over every machine with these errors: 15.19 against 0, 77.23 against 7.79. What it does promise: once every evaluation error is under $\Delta/2 = 3.9$, the policy is exact.
- $\max_{\pi}$ runs over the eight policies: 105 on dataset A (a policy that repairs a good machine and runs a worn one), 52.5 on B. The two-error form is the sharper board line.

Simulation lemma (stated): an evaluation error is Lecture 5's certificate in the wrong model (Kearns and Singh, 2002). Fix the same policy π in both models. As in Lecture 3, average over its actions, then define the Bellman backups for any value table U :

$$\begin{aligned}\widehat{r}^{\pi}(s) &:= \sum_a \pi(a | s) \widehat{r}(s, a), & \widehat{P}^{\pi}(s' | s) &:= \sum_a \pi(a | s) \widehat{P}(s' | s, a), \\ \mathcal{T}_M^{\pi} U &:= r^{\pi} + \gamma P^{\pi} U, & \mathcal{T}_{\widehat{M}}^{\pi} U &:= \widehat{r}^{\pi} + \gamma \widehat{P}^{\pi} U.\end{aligned}$$

The subscript names the model. $\widehat{V}^{\pi}$ is exact in $\widehat{M}$, a guess for M :

$$\|V^{\pi} - \widehat{V}^{\pi}\|_{\infty} \leq \frac{\|\mathcal{T}_M^{\pi} \widehat{V}^{\pi} - \widehat{V}^{\pi}\|_{\infty}}{1 - \gamma}$$

Lecture 5's certificate: $\mathcal{T}_M^{\pi}$ is a γ -contraction with fixed point V^{π}

$$\begin{aligned}\mathcal{T}_M^{\pi} \widehat{V}^{\pi} - \widehat{V}^{\pi} &= \mathcal{T}_M^{\pi} \widehat{V}^{\pi} - \mathcal{T}_{\widehat{M}}^{\pi} \widehat{V}^{\pi} = (r^{\pi} - \widehat{r}^{\pi}) + \gamma (P^{\pi} - \widehat{P}^{\pi}) \widehat{V}^{\pi} \\ \widehat{V}^{\pi} &= \mathcal{T}_{\widehat{M}}^{\pi} \widehat{V}^{\pi}: \text{the residual is the model error applied to } \widehat{V}^{\pi}\end{aligned}$$

Machine: $\widehat{r} = r$; only an incorrectly estimated run row contributes. Dataset A, π^* : at W, $0.9 \cdot \frac{1}{3} (35 - 40) = -1.5$; certificate 15.0, truth 7.59. Dataset B, π^* : at G, $0.9 \cdot \frac{1}{3} (56.5 - 80) = -7.05$; certificate 70.5, truth 35.70. Dataset B, repair when worn: $0.9 \cdot \frac{1}{3} (62 - 80) = -5.4$; 54.0 against 41.54.

- The residual lives only where the estimate is wrong, and the true backup of the estimated table is what a sampled transition can check.
- The model error multiplies value differences, e.g. $\widehat{V}^{\pi}(W) - \widehat{V}^{\pi}(G)$ in dataset B. A constant shift of $\widehat{V}^{\pi}$ costs nothing, as in Lecture 7.

How Much Data: Hoeffding

How accurate must the two fractions be? By computer, over every pair $(\hat{p}_G, \hat{p}_W)$: the plan's only possible mistake is "repair when worn", and it makes it exactly when $\hat{p}_G < \frac{11}{63} = 0.175$, the *boundary*; the W row never changes the answer. The truth $\frac{1}{3}$ sits 0.159 above it, the *margin*: keep $\hat{p}_G$ within 0.159 of the truth and the policy is exact. How many samples do that?

Hoeffding's inequality (stated) (Hoeffding, 1963). $X_1, \dots, X_n$ independent, each in $[a, b]$, mean μ , average $\bar{X}_n$. Choose a failure-probability bound $\delta \in (0, 1)$; $\delta = 0.1$ means at least 90% confidence. For every $t > 0$,

$$\Pr(|\bar{X}_n - \mu| \geq t) \leq 2 \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

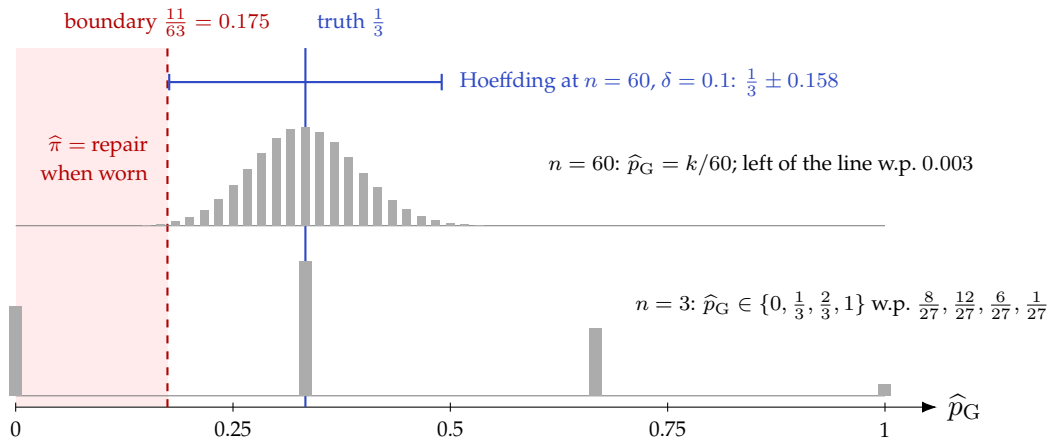
the average of n bounded draws concentrates; proof not given

so, with probability at least $1 - \delta$: $|\bar{X}_n - \mu| \leq (b-a) \sqrt{\frac{\log(2/\delta)}{2n}}$
set the right side to δ and solve for t

$\hat{P}(s' | s, a)$ is the average of n indicators, $b - a = 1$; $\hat{r}(s, a)$ the average of n profits, $b - a \leq 2R_{\max}$.

n	3	10	30	60	100	300	600
radius $\sqrt{\log 20/(2n)}$, $\delta = 0.1$	0.707	0.387	0.223	0.158	0.122	0.071	0.050
$\hat{p}_G \geq 0.175$ certified?	no	no	no	just	yes	yes	yes

Hoeffding's radius for one fraction. The margin from the truth $\frac{1}{3}$ to the boundary is 0.159: $n \geq \log 20 / (2 \cdot 0.159^2) = 59.5$, so $n = 60$ per pair, 360 samples in all. A radius of 0.1 needs 150, of 0.05 needs 600.



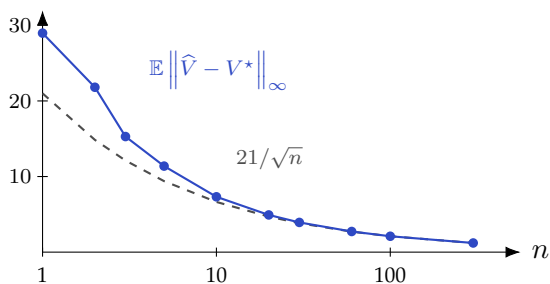
The fraction that decides the policy, at $n = 3$ and $n = 60$: its distribution (bars, exact), the boundary, the truth, and Hoeffding's interval, which just clears the boundary at $n = 60$. Shaded: the estimate repairs a worn machine.

- To halve the radius, quadruple n ; the radius never sees $|\mathcal{S}|$: a fraction is a fraction. Rewards: the same with $b - a = 2R_{\max}$.
- All 18 entries at once: a union bound adds $\log 18 = 2.9$ to $\log 20 = 3.0$, about twice the n ; the size of the table enters through a logarithm.
- A worst case over every distribution on $[0, 1]$: at $n = 60$ Hoeffding allows 0.097 for leaving the margin, the fair-ish coin does it with probability 0.003; at $n = 3$ the radius 0.707 certifies nothing.

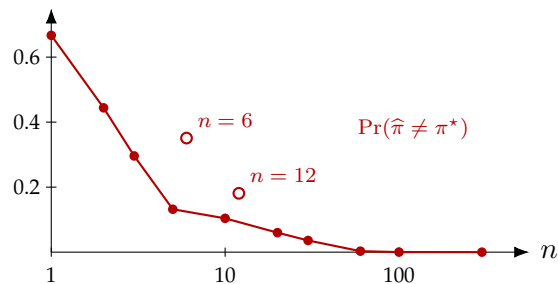
n per pair	$\Pr(\hat{\pi} \neq \pi^*)$	expected loss $\mathbb{E} \ V^* - V^{\hat{\pi}}\ _{\infty}$	$\mathbb{E} \ \hat{V} - V^*\ _{\infty}$	radius, $\delta = 0.1$
1	0.667	5.19	28.97	1.224
2	0.444	3.46	21.81	0.865
3	0.296	2.31	15.29	0.707
5	0.132	1.03	11.38	0.547
10	0.104	0.81	7.32	0.387
20	0.060	0.47	4.92	0.274
30	0.036	0.28	3.93	0.223
60	0.003	0.025	2.73	0.158
100	0.0002	0.002	2.10	0.122
300	0.0000	0.000	1.21	0.071

Loss against sample size, by computer: exact expectations over the two binomial counts.

$\Pr(\hat{\pi} \neq \pi^*) = \Pr(\hat{p}_G < \frac{11}{63})$; the expected loss is 7.79 times it, the only wrong policy losing 7.79. The value error falls like $21/\sqrt{n}$ from $n = 20$ on.



The values: wrong forever, less and less.



The policy: wrong or exactly right, and right for good.

- Two curves, two shapes: the values improve like $1/\sqrt{n}$ without end; the policy's loss is 0 or at least Δ , so its curve dies. Lecture 7's gap, with samples in place of rounds.
- Not monotone (hollow points): one W in six draws puts $\hat{p}_G = 0.167$ under 0.175, one in

five does not; $n = 6$ is wrong with probability 0.351, $n = 3$ with 0.296.

- Hoeffding's $n = 60$ against the exact 0.003: the certificate is twenty times too cautious here, and right about the shape.

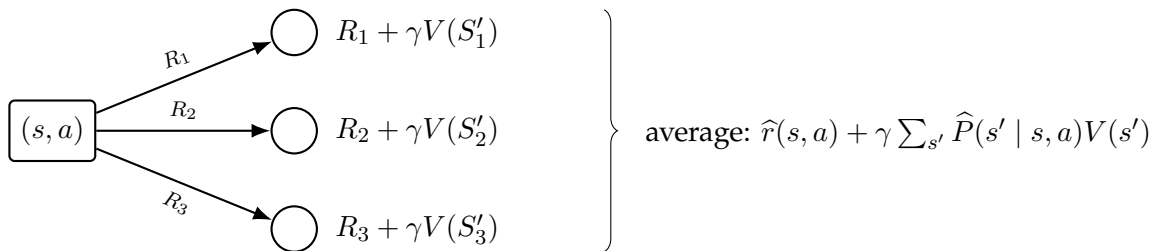
Without Building the Model

The estimate's backup is a sample average. For any table V and any pair with samples (R_i, S'_i) , $i = 1, \dots, n$:

$$\hat{r}(s, a) + \gamma \sum_{s'} \hat{P}(s' | s, a) V(s') = \frac{1}{n} \sum_{i=1}^n (R_i + \gamma V(S'_i))$$

$\hat{P}(s' | s, a) = \frac{1}{n} \# \{i : S'_i = s'\}$; regroup the sum by sample

Policy iteration and value iteration in $\widehat{M}$ touch $\widehat{P}$ only through this backup: keep the samples, never build the table.



Dataset A at (G, run) : three sampled backups, $8 + 0.9V(G)$, $8 + 0.9V(W)$, $8 + 0.9V(G)$, whose average is the estimate's lookahead.

	Work at one (s, a)	Work over all (s, a)
Dense $\widehat{P}$ table	$O(\mathcal{S})$	$O(\mathcal{S} ^2 \mathcal{A})$
n stored samples per pair	$O(n)$	$O(n \mathcal{S} \mathcal{A})$

One full Bellman backup, holding V fixed. Same data, same empirical backup; fewer terms to sum when $n \ll |\mathcal{S}|$.

Prediction or control: the same estimated model.

- Prediction: given π , solve $\widehat{V}^\pi = \widehat{r}^\pi + \gamma \widehat{P}^\pi \widehat{V}^\pi$ by Lecture 3's policy evaluation.
 - Control: choose $\widehat{\pi}$ by Lecture 6's policy iteration in $\widehat{M}$; return its optimal value $\widehat{V}$.
- Dataset B at W: fixing "repair when broken" gives $\widehat{V}^{\pi^*}(W) = 56.5$; optimizing gives $\widehat{V}(W) = 62$. Next meeting: the same prediction question, answered by averaging sampled returns.

	What is stored, what is averaged	Where
Certainty equivalence	the table $\widehat{M}$, $ \mathcal{S} ^2 \mathcal{A} $ numbers; then plan; the backup above at every pair, every iteration	today
Monte Carlo	no model; the sampled returns that followed (s, a) , one number per visit	the next meeting
Temporal difference	no model; one sampled backup $R + \gamma V(S')$ at a time, moved toward rather than averaged	later in the unit

Three ways to use the same transitions. [Sutton and Barto \(2018, Section 6.3\)](#): for a fixed policy π , batch temporal-difference prediction converges to the empirical-model value $\widehat{V}^\pi$, without forming $\widehat{M}$.

Summary

Lecture 8 on one board

- Access: a sampling oracle (any pair, n times), the running machine (one stream), a log (someone else's stream). Hidden: the 24 numbers of (P, r) .
- Certainty equivalence: $\widehat{P} = \text{counts}/n$, $\widehat{r} = \text{average}$, then Lecture 6's policy iteration in $\widehat{M}$. Machine, $n = 3$: dataset A $\widehat{V} = (50, 40, 35)$, $\widehat{\pi} = \pi^*$; dataset B $\widehat{V} = (80, 62, 62)$, $\widehat{\pi} = \text{repair when worn}$, loss 7.79, w.p. 8/27.
- The flip, at W: repair buys a good machine, worth $-10 + 0.9 V(G)$; $V(G) = 44.30, 50, 80$ in $M, \widehat{M}_A, \widehat{M}_B$, so repair at W is worth 29.87, 35, 62 against running on, 32.41, 40, 56.5.
- Lemma, three lines: $\|V^* - V^{\widehat{\pi}}\|_\infty \leq \varepsilon(\pi^*) + \varepsilon(\widehat{\pi}) \leq 2 \max_\pi \varepsilon(\pi)$; under Δ , exact. Simulation lemma (stated): $\varepsilon(\pi) \leq \|\mathcal{T}_M^\pi \widehat{V}^\pi - \widehat{V}^\pi\|_\infty / (1 - \gamma)$, the residual at the wrong row: $-7.05, 70.5$ against 35.70.
- By computer, wrong iff $\widehat{p}_G < \frac{11}{63} = 0.175$, margin 0.159. Hoeffding (stated): $|\bar{X}_n - \mu| \leq (b - a) \sqrt{\log(2/\delta)/(2n)}$; $n = 60$ certifies the margin; the policy is wrong w.p. 0.30, 0.10, 0.003 at $n = 3, 10, 60$; values $\approx 21/\sqrt{n}$.
- Bridge: $\widehat{r} + \gamma \sum \widehat{P}V = \frac{1}{n} \sum (R_i + \gamma V(S'_i))$: the estimate's backup is a sample average. One full backup: dense table $O(|\mathcal{S}|^2|\mathcal{A}|)$, stored samples $O(n|\mathcal{S}||\mathcal{A}|)$.
- Prediction: fix π , evaluate $\widehat{V}^\pi$ in $\widehat{M}$; control: optimize the policy. Next meeting, Monte Carlo answers the prediction question by averaging returns.

References

- Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. Book draft, 2022. URL <https://rltheorybook.github.io/>.
- Emma Brunskill. CS 234: Reinforcement learning. Stanford University course materials, 2026. URL <https://web.stanford.edu/class/cs234/>. Winter 2026 offering.
- Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
- Nan Jiang. CS 443: Reinforcement learning. University of Illinois Urbana-Champaign course materials, 2026. URL <https://nanjiang.cs.illinois.edu/cs443/>. Spring 2026 offering.
- Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002.
- Wen Sun. CS 4/5789: Introduction to reinforcement learning. Cornell University course materials, 2025. URL https://wensun.github.io/CS4789_spring_2025.html. Spring 2025 offering.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. URL <http://incompleteideas.net/book/the-book-2nd.html>.