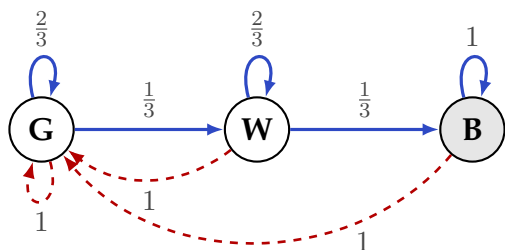


Lecture 12: Off-Policy Evaluation: Importance Sampling

How to use these notes. Board content in board order: figures, equations, tables, one-line statements. Today's task is *off-policy evaluation*: the value of a rule that was never run, computed from the log of a manager who ran another rule. The note stands on its own: every symbol is defined where it is first used. Reading: [Sutton and Barto \(2018, Sections 5.5, 5.9, and 7.3\)](#); optional, [Precup et al. \(2000\)](#). 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 Task: the Value of a Rule Never Run

A workshop runs one machine. Each week it is good (G), worn (W), or broken (B), and it is run or repaired. The random manager ran it for years, a coin flip every week, and kept the log. Wanted: what another rule would have earned, from his log alone; that rule is never run. The machine's numbers below are the truth, for checking; the learner never sees them.



The machine, both actions. Run (solid): a good machine earns 8 and stays good w.p. $\frac{2}{3}$, else wears; a worn one earns 4 and stays worn w.p. $\frac{2}{3}$, else breaks; a broken one earns 0 and stays broken. Repair (dashed): costs 10, leaves it good. Gray: hidden from the learner.

$V^{\pi}(s)$, the truth	G	W	B
π^* : repair when broken	44.30	32.41	29.87
never repair	27.5	10	0
b : the random manager	-15.20	-18.67	-21.53

The rule to evaluate, π^* ; the rule that acted, b (run or repair w.p. $\frac{1}{2}$ in every state); never repair, for comparison only. Known to the learner:

$\mathcal{S} = \{G, W, B\}$, $\mathcal{A} = \{\text{run}, \text{repair}\}$, $\gamma = 0.9$, the log with its actions, the coin's probabilities $\frac{1}{2}$, and π^* . Hidden: P (18 numbers), r (6), and every week π^* itself would have had.

Policy evaluation: what is wanted (stated) (Sutton and Barto, 2018, Section 3.5). The machine is a Markov decision process $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$: action a in state s pays a reward with mean $r(s, a)$ (on the machine, exactly $r(s, a)$) and moves to s' with probability $P(s' | s, a)$; $\gamma \in [0, 1)$ discounts; a policy $\pi(a | s)$ draws the action, written $\pi(s)$ when deterministic.

$$V^\pi(s) := \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t R_{t+1} \mid S_0 = s \right]$$

the value of π : the mean discounted return when π chooses every action

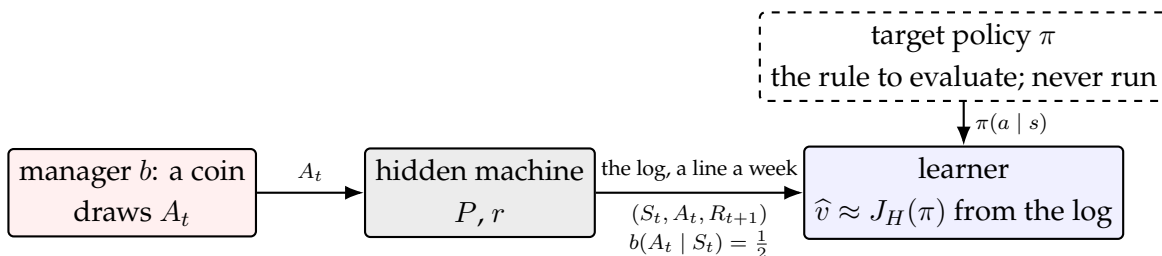
$$J_H(\pi) := \mathbb{E}_\pi \left[\sum_{t < H} \gamma^t R_{t+1} \mid S_0 = G \right]$$

the H -week value from a good machine: what today estimates

$$|V^\pi(G) - J_H(\pi)| \leq \gamma^H \cdot \frac{10}{1-\gamma} = 0.97 \text{ at } H = 44$$

the cut tail, $|R| \leq 10$: $J_{44}(\pi^*) = 43.94$ for 44.30

With P and r in hand, V^π is a linear solve and $J_H(\pi)$ is H backups. Today P and r are hidden, and the only weeks on record were chosen by another rule.



Off-policy evaluation. The behavior policy b drew the logged actions; the target policy π is the rule whose value is wanted, never run. The learner reads the log and the probabilities $\pi(a | s)$, and returns one number.

The data, the target, and the question.

$$S_0 = G, A_0, R_1, S_1, A_1, R_2, \dots, S_{H-1}, A_{H-1}, R_H, \quad A_t \sim b(\cdot | S_t)$$

one log: a good machine run for H weeks by the behavior policy b

$$b(\text{run} | s) = b(\text{repair} | s) = \frac{1}{2} \text{ for every } s, \text{ written in the log}$$

the coin: the logged probability of each call is known

$$\pi(a | s) \text{ given; today } \pi = \pi^* : \text{ repair when broken}$$

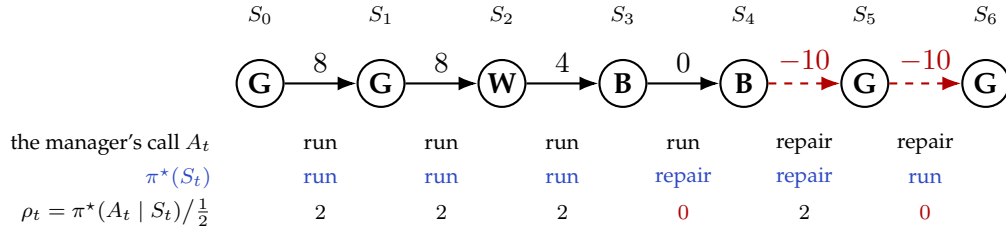
the target policy: the rule to evaluate, deterministic here, never run

$$\text{wanted: } \hat{v} \text{ from the } n \text{ logs with } \mathbb{E}[\hat{v}] = J_H(\pi)$$

off-policy evaluation: the rule evaluated is not the rule that acted

- n logs, each a good machine run for H weeks by the coin; log i gives one estimate $\hat{v}^{(i)}$; $\hat{v}$ is their average.

- Known to the learner: the logs, the probabilities of b (the coin, $\frac{1}{2}$), and π . Hidden: P , r , and the weeks π itself would have produced, which never happened.
- In the derivations π is any target policy. In every example and every number below, $\pi = \pi^*$.
- Truth, for checking: $J_8(\pi^*) = 28.31$; the H -week values $J_H(\pi^*)$ are tabulated below.



Six weeks of one log, used by hand below. The manager's calls are his coin; under them, what π^* would have called at the same state, and the ratio ρ_t : 2 where the two agree, 0 where they differ (weeks 3 and 5). The learner reads the log; it chose nothing.

From a log whose actions follow one rule, how do we price another rule that was never run: what weight makes the average right, why the weight is a product, how fast it blows up with the horizon, and what two repairs buy?

Today's targets.

1. One week: the ratio $\rho = \pi(A | S) / b(A | S)$ makes the manager's week count as a week of π , $\mathbb{E}_b[\rho R] = \mathbb{E}_\pi[R]$; on the coin ρ is 2 or 0.
2. H weeks: the two trajectory probabilities differ only in the action factors, so their ratio is $\prod_t \rho_t$; the hidden P cancels. By hand: 147.52 or 0 for a truth of 17.42.
3. The price: one log in 2^H counts, weighted 2^H ; one log's spread 8, 24.5, 82.6, 470.7 at $H = 1, 2, 4, 8$; at 44 weeks one log in 1.8×10^{13} .
4. Two repairs within the log: per-decision weights R_{t+1} by $\rho_{0:t}$ alone (70.72 by hand; spread 104.9 at $H = 8$); weighted IS averages the matched logs. Coverage: a log that never tried the call prices nothing.
5. The third repair: a table supplies the continuation, one ratio per backup is left; ratios in the weight = calls of b left inside the target: $H, n, 1, 0$. By computer 1.59 against 45.61.

One Week: the Weight That Fixes the Average

Change of measure in one step (Kahn and Marshall, 1953; Horvitz and Thompson, 1952). At a state s , for any function f of the action taken there:

$$\mathbb{E}_{A \sim \pi(\cdot | s)}[f(A)] = \sum_a \pi(a | s) f(a) \quad \text{what a week of } \pi \text{ at } s \text{ would average}$$

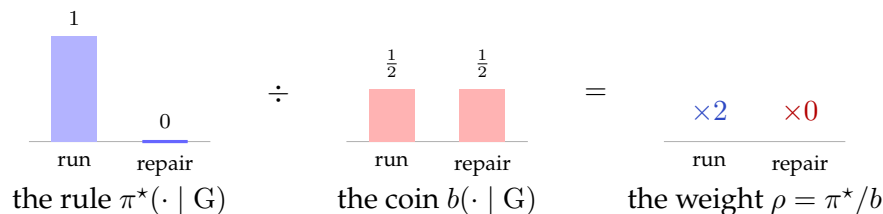
$$= \sum_a b(a | s) \frac{\pi(a | s)}{b(a | s)} f(a) = \mathbb{E}_{A \sim b(\cdot | s)}[\rho f(A)], \quad \rho := \frac{\pi(A | s)}{b(A | s)}$$

multiply and divide by $b(a | s)$ where $b(a | s) > 0$: a week of b , reweighted

$$\mathbb{E}_{A \sim b(\cdot | s)}[\rho] = \sum_a b(a | s) \frac{\pi(a | s)}{b(a | s)} = \sum_a \pi(a | s) = 1$$

the weights average to one: the log is re-counted, not inflated

ρ is *the ratio* (the importance weight): it turns a week drawn by b into a week that counts as a week of π . Needed: $b(a | s) > 0$ wherever $\pi(a | s) > 0$ (*coverage*, below). On the coin, $\rho = \pi(A | s) / \frac{1}{2}$: 2 when the logged call is the one π makes, 0 when it is not.



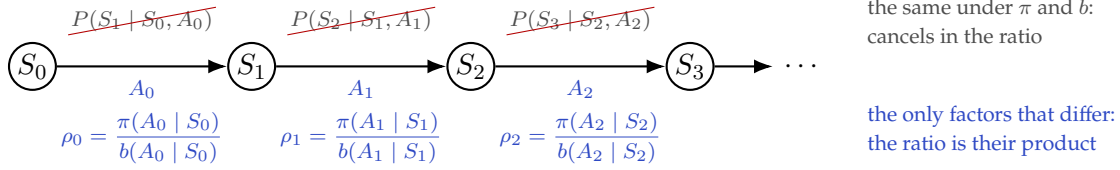
At G the coin puts $\frac{1}{2}$ on each call and π^* puts 1 on run. Every logged run week counts twice, every logged repair week counts zero: the reweighted log has the same mix of calls as π^* .

Example: week 0 of the log, $(G, \text{run}, 8)$. One week of π^* from G is worth $r(G, \text{run}) = 8$, a number the learner does not know. What does the manager's week say, ρR_1 ? And a week in which he repaired a good machine, $(G, \text{repair}, -10)$?

$2 \cdot 8 = 16$; $0 \cdot (-10) = 0$. Over his weeks at G , half say 16 and half say 0: mean 8, spread 8. A week run by π^* itself would say 8 every time, spread 0: the noise is in the weight, not in the reward.

- K actions, a uniform coin, a deterministic π : $\rho \in \{K, 0\}$, and a constant reward r has variance $r^2(K - 1)$ under the weights, 0 under π . One week in K counts.
- The weight uses the probability of the logged call under b and under π , never P or r . Here it works because the manager wrote the coin down.

H Weeks: the Weight Is a Product



One trajectory, two probabilities. The machine's moves cost the same under either rule; only the action factors differ.

The trajectory ratio: the hidden P cancels. For a trajectory $\tau = (s_0, a_0, \dots, s_{H-1}, a_{H-1}, s_H)$ from $s_0 = G$ with return $G_H(\tau) = \sum_{t < H} \gamma^t r(s_t, a_t)$:

$$P_\pi(\tau) = \pi(a_0 | s_0) P(s_1 | s_0, a_0) \cdots \pi(a_{H-1} | s_{H-1}) P(s_H | s_{H-1}, a_{H-1})$$

its probability when π chooses: an action factor and a move factor each week

$$P_b(\tau) = b(a_0 | s_0) P(s_1 | s_0, a_0) \cdots b(a_{H-1} | s_{H-1}) P(s_H | s_{H-1}, a_{H-1})$$

its probability when b chooses: the same move factors

$$\frac{P_\pi(\tau)}{P_b(\tau)} = \prod_{t=0}^{H-1} \frac{\pi(a_t | s_t)}{b(a_t | s_t)} = \prod_{t=0}^{H-1} \rho_t =: \rho_{0:H-1}$$

every P factor cancels: the weight needs the two rules and the log, not the machine

$$\mathbb{E}_b[\rho_{0:H-1} G_H] = \sum_{\tau} P_b(\tau) \frac{P_\pi(\tau)}{P_b(\tau)} G_H(\tau) = \sum_{\tau} P_\pi(\tau) G_H(\tau) = J_H(\pi)$$

the one-week identity on whole trajectories; coverage: $P_\pi > 0 \Rightarrow P_b > 0$ $\square$

On the coin with a deterministic π : $\rho_{0:H-1} = 2^H$ if every one of the H logged calls is the one π makes (a *matched log*), else 0.

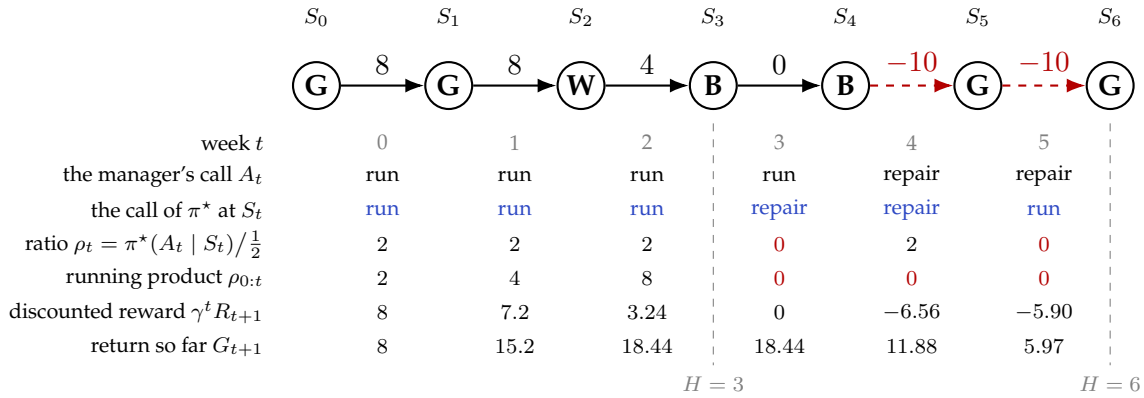
Ordinary importance sampling (Precup et al., 2000), (Sutton and Barto, 2018, Section 5.5).

- Input: n logs of H weeks from G drawn by b , with the probabilities of b ; the target policy π . Output: one number $\hat{v}_{\text{IS}}$.
- For each log i : the weight $\rho^{(i)} = \prod_{t < H} \pi(A_t | S_t) / b(A_t | S_t)$ and the return $G^{(i)} = \sum_{t < H} \gamma^t R_{t+1}$; then

$$\hat{v}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^n \rho^{(i)} G^{(i)} \quad \text{unbiased for } J_H(\pi): \text{ the line above, averaged over independent logs}$$

- Uses P and r nowhere; needs coverage. On the coin: the matched logs count 2^H each, the others 0.

Example: the six weeks by hand. The target is π^* , repair when broken; the log is the six weeks drawn above. Under each week: the manager's call A_t ; the call π^* would make at S_t ; the ratio $\rho_t = \pi^*(A_t | S_t) / b(A_t | S_t)$, which is 2 when the two calls agree and 0 when they differ (since $b = \frac{1}{2}$); the running product $\rho_{0:t} = \rho_0 \rho_1 \cdots \rho_t$; the discounted reward $\gamma^t R_{t+1}$; and the return so far, $G_{t+1} = \sum_{k \leq t} \gamma^k R_{k+1}$. Then the estimate $\rho_{0:H-1} G_H$ at $H = 3$ and at $H = 6$.



The six weeks against π^* . Where the manager's call is the one π^* makes, the ratio is 2; where it is not (weeks 3 and 5, red), 0, and from the first 0 the running product stays 0. At $H = 3$: three matches, product 8, return $G_3 = 18.44$. At $H = 6$: product 0, whatever the return.

$$H = 3 : \quad \rho_{0:2} G_3 = 8 \times 18.44 = 147.52 \quad \text{truth } J_3(\pi^*) = 17.42$$

the first three calls all match: this log counts, weighted $2^3 = 8$; eight times too big

$$H = 6 : \quad \rho_{0:5} G_6 = 0 \times 5.97 = 0 \quad \text{truth } J_6(\pi^*) = 24.56$$

the call of week 3 did not match: this log is thrown away whole

- One log says 147.52 or 0, never anything near 17.42. Only the average over many logs is right: at $H = 3$ one log in 8 matches and counts 8 times.
- At $H = 6$ a log counts only if all six calls match π^* : run at G, run at W, repair at B, six weeks running. One log in $2^6 = 64$.

The Price of the Horizon

How many logs count, and how much each counts. The coin against a deterministic π , H weeks:

$$P_b(\rho_{0:H-1} \neq 0) = \left(\frac{1}{2}\right)^H \quad \text{\textit{H independent coin flips must all land on the call } \pi \text{ makes}}$$

$$\mathbb{E}_b[(\rho_{0:H-1} G_H)^2] = 2^H \mathbb{E}_\pi[G_H^2], \quad \text{Var}_b[\rho_{0:H-1} G_H] = 2^H \mathbb{E}_\pi[G_H^2] - J_H(\pi)^2$$

$$\rho^2 = 2^H \rho \text{ on the coin, and } \mathbb{E}_b[\rho G_H^2] = \mathbb{E}_\pi[G_H^2] \text{ by the identity above}$$

$$\mathbb{E}_b[\rho_t^2] = \frac{1}{2} \cdot 2^2 = 2$$

each week's ratio has mean 1 and mean square 2: the spread of the weight doubles every week

weeks H	1	2	4	6	8	10
logs of 100 that count (mean)	50	25	6.25	1.56	0.39	0.10
weight 2^H on each	2	4	16	64	256	1,024
truth $J_H(\pi^*)$	8.00	14.00	20.01	24.56	28.31	31.35
spread of one log's estimate $\rho_{0:H-1} G_H$	8.0	24.5	82.6	204.8	470.7	1,038.5
spread of G_H when π^* itself runs, one episode	0	1.70	7.13	7.84	8.18	8.43

Target π^* from G , exact. The spread (standard deviation) of the weighted return grows like $2^{H/2}$ from the weights alone, and more as the return itself grows; the return of an episode run by π^* itself never spreads past 8.5.

to pin $J_8(\pi^*)$ within ± 1 : $470.7^2 \approx 221,600$ logs, or $8.18^2 \approx 67$ episodes of π^*

one spread: the average of n estimates has spread $\text{spread}/\sqrt{n}$

$H = 44$ weeks (the tail under 1): $P_b(\text{matched}) = 2^{-44} \approx 6 \times 10^{-14}$

one log in 1.8×10^{13} counts, and it counts 1.8×10^{13} times

- The weight is a product of H factors with mean 1 each; its spread multiplies. Exponential in the horizon, whatever the machine: *the curse of horizon* (Liu et al., 2018).
- Unbiased is not enough: most logs say 0 and a rare one says $2^H G$. A short horizon, or a b close to π , keeps the weight tame.

Reading (optional): infinite variance on the machine (Sutton and Barto, 2018, Example 5.5). Nothing below depends on this box. Here the target is never repair, run everywhere, chosen because its return is complete once the machine breaks: evaluate it from the coin's log to the week the machine breaks, T , after which every reward is 0; $T = a + b$ with $a, b \sim \text{Geometric}(\frac{1}{3})$. A matched log has $\rho = 2^T$ and $G \geq 8$, so $\mathbb{E}_b[(\rho G)^2] = \mathbb{E}_\pi[2^T G^2] \geq 64 \mathbb{E}_\pi[2^T]$, and $\mathbb{E}_\pi[2^a] = \sum_{k \geq 1} (\frac{2}{3})^{k-1} \frac{1}{3} 2^k = \infty$ since $2 \cdot \frac{2}{3} > 1$. Unbiased with infinite variance: the average over logs never settles. Per-decision IS (below) fares no better here: the week- t term has mean square at least $32 \cdot (0.81 \cdot 2 \cdot \frac{2}{3})^t = 32 \cdot 1.08^t$. Weighted IS stays in $[0, 80]$.

Two Repairs

ordinary: $\rho_{0:5}$ on every reward	$\times 64$	$\times 64$	$\times 64$	$\times 64$	$\times 64$	$\times 64$
	$\textcircled{R_1}$	$\textcircled{R_2}$	$\textcircled{R_3}$	$\textcircled{R_4}$	$\textcircled{R_5}$	$\textcircled{R_6}$
per-decision: $\rho_{0:t}$ on R_{t+1}	$\times 2$	$\times 4$	$\times 8$	$\times 16$	$\times 32$	$\times 64$
one ratio: ρ_t on the backup of week t	$\times 2$	$\times 2$	$\times 2$	$\times 2$	$\times 2$	$\times 2$

A six-week log whose calls all match π . The trajectory weight $2^6 = 64$ sits on every reward; the per-decision weight on R_{t+1} counts only the $t + 1$ calls that could have changed it; with a table for the continuation, the backup of week t carries ρ_t alone: the third repair, below.

Repair 1: per-decision importance sampling (Precup et al., 2000), (Sutton and Barto, 2018, Section 5.9). The reward of week $t + 1$ is decided by the calls up to week t ; the later calls only add noise to its weight.

$$\mathbb{E}_b[\rho_{t+1} \mid S_0, A_0, \dots, S_{t+1}] = \sum_a b(a \mid S_{t+1}) \frac{\pi(a \mid S_{t+1})}{b(a \mid S_{t+1})} = 1$$

the one-week identity at week $t + 1$: a later ratio averages to one, given the past

$$\mathbb{E}_b[\rho_{0:H-1} R_{t+1}] = \mathbb{E}_b[\rho_{0:t} R_{t+1} \rho_{t+1} \cdots \rho_{H-1}] = \mathbb{E}_b[\rho_{0:t} R_{t+1}]$$

peel the later ratios off one at a time, last first; each averages to one

$$\hat{v}_{\text{PD}} := \sum_{t < H} \gamma^t \rho_{0:t} R_{t+1}, \quad \mathbb{E}_b[\hat{v}_{\text{PD}}] = J_H(\pi)$$

sum over t : unbiased, with 2^{t+1} in place of 2^H on the week- t reward

On the six weeks: $2 \cdot 8 + 4 \cdot 7.2 + 8 \cdot 3.24 = 16 + 28.8 + 25.92 = 70.72$ at every $H \geq 3$: the mismatch at week 3 zeroes the later terms; the three earlier ones survive.

Repair 2: weighted importance sampling (Sutton and Barto, 2018, Section 5.5), (Precup et al., 2001). Over the n logs, with $\rho^{(i)} = \rho_{0:H-1}$ of log i :

$$\hat{v}_{\text{IS}} = \frac{1}{n} \sum_{i=1}^n \rho^{(i)} G^{(i)} = \frac{\#\{\text{matched}\}}{n/2^H} \cdot \frac{\sum_{\text{matched}} G^{(i)}}{\#\{\text{matched}\}}$$

on the coin: the matched logs' average return, times their count over its expected count $n/2^H$

$$\hat{v}_{\text{W}} := \frac{\sum_i \rho^{(i)} G^{(i)}}{\sum_i \rho^{(i)}} = \frac{\sum_{\text{matched}} G^{(i)}}{\#\{\text{matched}\}} \quad (0 \text{ if no log matched})$$

drop the count factor: normalize by the weights actually seen

- A matched log is an episode of π itself: the coin called what π calls, and the machine did the rest.
- So $\hat{v}_{\text{W}}$ is Monte Carlo on the matched logs: right on average whenever one log matched, 0 when none did.
- Bounded, $\min_i G^{(i)} \leq \hat{v}_{\text{W}} \leq \max_i G^{(i)}$, whatever the weights. In general (a random π , an uneven b) biased at finite n , consistent as $n \rightarrow \infty$.

Estimator	One log's estimate	Mean	Spread, $H = 8$	Needs
Monte Carlo, on-policy	G_H of an episode run by π	$J_H(\pi)$	8.18	π run for H weeks
ordinary IS	$\rho_{0:H-1} G_H$	$J_H(\pi)$	470.7; $2^{H/2}$ and worse	the probabilities of b ; coverage
per-decision IS	$\sum_{t < H} \gamma^t \rho_{0:t} R_{t+1}$	$J_H(\pi)$	104.9	the same
weighted IS	$\sum_i \rho^{(i)} G^{(i)} / \sum_i \rho^{(i)}$ over the n logs	$J_H(\pi)$ as $n \rightarrow \infty$; here exact given a match	within $[\min G, \max G]$	the same, and a rule for no match

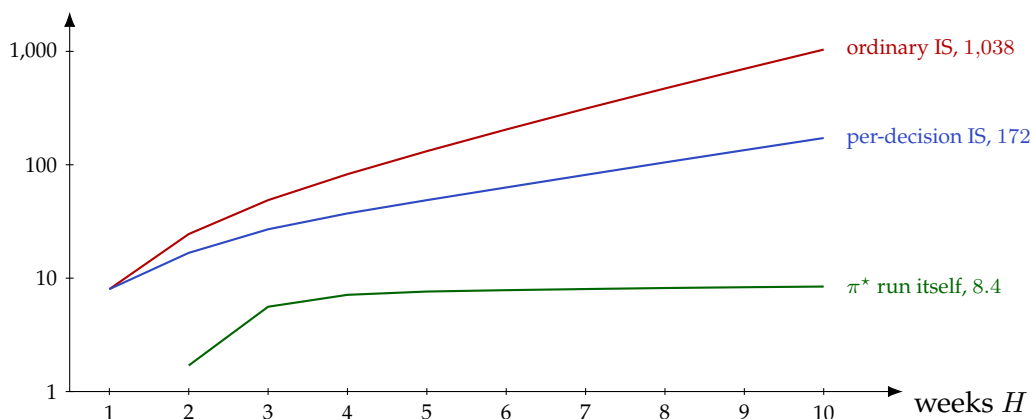
Four ways to price π^* over 8 weeks. The spreads are exact, one log or one episode each; the weighted estimator's spread is over the n logs together and cannot leave the range of the returns.

The experiment: three estimators, one coin, the horizon as the dial.

1. $n = 100$ logs of H weeks from G by the coin; the target π^* ; $H = 2, 4, 6, 8, 10$. From the same 100 logs, all three estimates.
2. Repeat on 1,000 fresh sets of 100 logs; report the root-mean-square error against $J_H(\pi^*)$ and the mean number of matched logs. Reference: Monte Carlo on 100 episodes of π^* itself.

weeks H	matched logs of 100 (mean)	error, 100 logs of the coin			error, 100 episodes of π^* itself
		ordinary IS	weighted IS	per-decision IS	
2	25.0	2.50	0.34	1.69	0.17
4	6.2	8.10	3.29	3.63	0.71
6	1.5	19.84	12.32	6.26	0.78
8	0.4	45.61	23.64	10.30	0.82
10	0.1	103.69	29.90	16.75	0.84
8, 10,000 logs	38.8	4.86	1.40	1.08	—

Root-mean-square error against $J_H(\pi^*)$; seed 639; 1,000 sets of logs each row. Weighted IS wins while logs match ($H = 2$: Monte Carlo on 25 matched logs, $1.70/\sqrt{25} = 0.34$) and returns 0 when none does, two sets in three at $H = 8$; per-decision IS wins once matches are rare. A hundred times the logs divides every error by about ten.



The spread of one log's estimate against the horizon, exact, the vertical axis logarithmic. Two straight lines of different slope: a factor about 1.7 a week for ordinary IS, 1.3 for per-decision IS; the return of π^* itself levels off.

- At $H = 2$ a quarter of the logs match: weighted IS is Monte Carlo on 25 episodes and beats everything but running π^* .
- From $H = 6$ on, most sets have one matched log or none (67% none at $H = 8$): weighted IS returns one return or 0; per-decision IS still uses every log's first weeks.

- Ordinary IS at $H = 10$: error 103.69 on a truth of 31.35, from one log in a thousand weighted 1,024.
- Running π^* for 100 episodes is within 0.84. The gap to every off-policy column is the price of never running it.

The Third Repair: One Ratio, Not a Product

Importance sampling takes the continuation after week t from the log and pays a ratio for every call in it. A table V in place of that continuation leaves one call to pay for.

Repair 3: off-policy TD(0), the template with one ratio on the step (Sutton and Barto, 2018, Sections 7.3 and 11.1), (Jiang, 2023, Section 1.3).

$$V(S_t) \leftarrow V(S_t) + \alpha \rho_t (R_{t+1} + \gamma V(S_{t+1}) - V(S_t)), \quad \rho_t = \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

TD(0) on the logged week t ; the error is multiplied by the ratio of that one week

$\mathbb{E}_b[\rho_t (R_{t+1} + \gamma V(S_{t+1})) | S_t = s] = \mathbb{E}_\pi[R_{t+1} + \gamma V(S_{t+1}) | S_t = s] = (\mathcal{T}^\pi V)(s)$
the one-week identity at s with $f(a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) V(s')$: the target counts as a week of π

$$\mathbb{E}_b[\rho_t V(S_t) | S_t = s] = V(s) \Rightarrow \text{mean step at } s = \alpha((\mathcal{T}^\pi V)(s) - V(s))$$

$\mathbb{E}_b[\rho_t] = 1$; zero at $V = V^\pi$: the mean step of TD(0) run under π itself

Why one ratio: the target $R_{t+1} + \gamma V(S_{t+1})$ holds one call of b , the call A_t at S_t . The weeks after t are the table's business, so no later ratio appears. Ordinary IS holds H calls of b in G_H and pays $\rho_{0:H-1}$.

the whole log, G_H (ordinary IS)	R_1 R_2 $\dots$ R_H	H calls of b inside: H ratios, $\rho_{0:H-1}$
n weeks, then the table	R_{t+1} $\dots$ R_{t+n} $\gamma^n V(S_{t+n})$	n calls: n ratios, $\rho_{t:t+n-1}$
one week, then the table (Repair 3)	R_{t+1} $\gamma V(S_{t+1})$	one call, A_t : one ratio, ρ_t
a Q table, priced at (S_t, A_t)	R_{t+1} $\gamma \sum_a \pi(a S_{t+1}) Q(S_{t+1}, a)$	no call of b inside: no ratio

What each target holds. Red: a reward decided by a call of b at a state being priced, one ratio each. Blue: a table entry, the continuation without the log. Gray: the reward of the very pair being priced, no ratio. Ratios in the weight = calls of b left inside the target: $H, n, 1, 0$; the Q table reads the next call of π off the table (stated).

Example: the six weeks, one ratio each. Target π^* ; step $\alpha = 0.1$; the table V starts at 0 in every state. Week t moves $V(S_t)$ with step $\alpha\rho_t$: 0.2 when the manager's call is the one π^* makes, nothing when it is not. Each line: the week's record (S_t, A_t, R_{t+1}) , then S_{t+1} .

week 0, (G, run, 8) $\rightarrow$ G : $V(G) \leftarrow 0 + 0.2(8 + 0.9 \cdot 0 - 0) = 1.6$

week 1, (G, run, 8) $\rightarrow$ W : $V(G) \leftarrow 1.6 + 0.2(8 + 0.9 \cdot 0 - 1.6) = 2.88$

week 2, (W, run, 4) $\rightarrow$ B : $V(W) \leftarrow 0 + 0.2(4 + 0.9 \cdot 0 - 0) = 0.8$

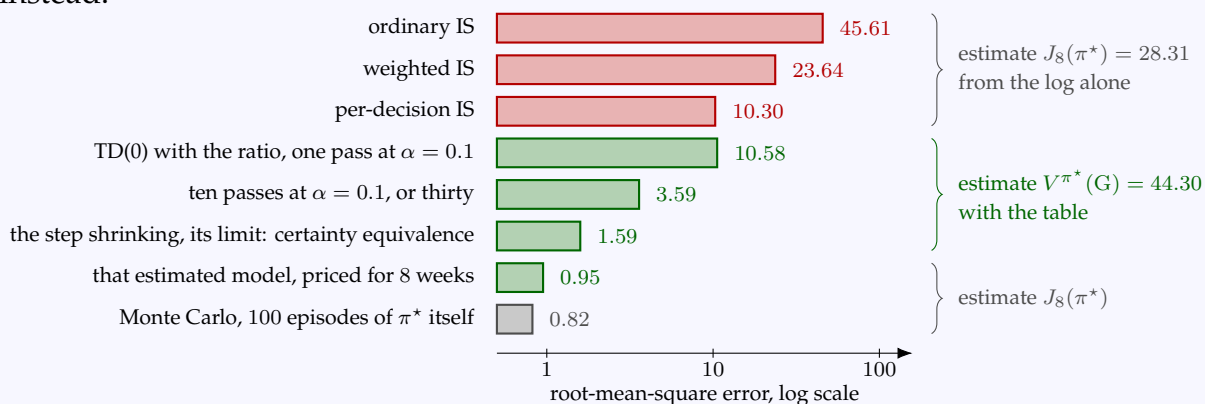
week 3, (B, run, 0) : $\rho_3 = 0$, nothing

week 4, (B, repair, -10) $\rightarrow$ G : $V(B) \leftarrow 0 + 0.2(-10 + 0.9 \cdot 2.88 - 0) = -1.48$

week 5, (G, repair, -10) : $\rho_5 = 0$, nothing

Half the weeks count, each on its own coin flip: not one log in 2^H .

By computer: the same 100 logs, $H = 8$, the record replayed. Replayed with a shrinking step, the table settles at certainty equivalence in the model estimated from the matched steps (stated): here 400 of the 800 steps, 319 at G, 66 at W, 16 at B. A fixed step circles instead.



Within ± 1 of J_8 : 221,600 logs for ordinary IS, about 90 with the ratio and the table, 67 episodes of π^* itself.

- On this machine the bootstrapped log is nearly as good as episodes of π^* itself: 90 logs against 67 episodes. The bill moved from 2^H to the states b seldom visits, and to the step.
- One pass at $\alpha = 0.1$ sits 10 low: B sees 16 matched steps per 100 logs, and a backup crosses the cycle G, W, B one step a visit.
- A fixed step never lands: 3.59 after ten passes and after thirty. Halving the step every five passes lands within 0.34 of the table's 1.59 after forty; more logs shrink 1.59 itself, 0.72 at 400.
- The ratio on the error, not on the target: one update at G has spread 7.1 against 44.9; on

one record in eight the second form runs away under repeated passes.

- What the table trusts: the state. Steps arrive as b visits; a table settles anyway (stated) (Sutton and Barto, 2018, Section 11.1); features in its place can diverge (stated) (Baird, 1995). IS trusts the log alone and pays the horizon.

Coverage: What the Log Must Hold

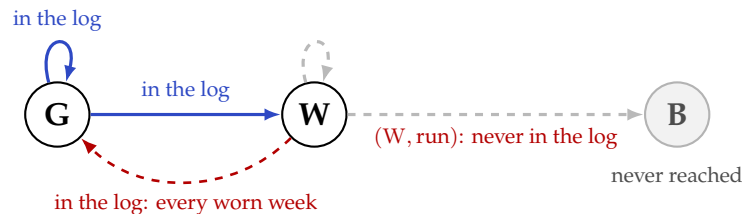
Coverage (Sutton and Barto, 2018, Section 5.5).

$\pi(a | s) > 0 \Rightarrow b(a | s) > 0$ for every pair π can reach

the one step that multiplies and divides by $b(a | s)$ needs $b(a | s) > 0$

the coin: $b(a | s) = \frac{1}{2} > 0$ at every pair: every rule is covered *expensive (2^H), but possible*

Where it fails, no weight exists: the log holds no week of the kind π would produce, and no re-counting of other weeks makes one.



The log of a manager who repairs whenever the machine is worn. It holds (G, run) and (W, repair) and nothing else; π^* needs (W, run) .

Example: can the repair-when-worn manager's log price π^* ? His rule: run at G, repair at W; under him the machine never breaks. π^* runs at W.

No. At W, $\pi^*(\text{run} | W) = 1$ and $b(\text{run} | W) = 0$: no ratio exists. Every log reaches W within a few weeks and is unusable from that week on. His log prices rules that repair at W, and nothing else: a call never tried in the log cannot be priced from it.

Summary

Off-policy evaluation on one board

- Task: $J_H(\pi)$, the H -week value of a rule never run, from n logs of the random manager's coin ($b = \frac{1}{2}$, written down). Truth: $J_8(\pi^*) = 28.31$; $V^{\pi^*}(G) = 44.30$.
- One week: $\mathbb{E}_b[\rho f(A)] = \mathbb{E}_\pi[f(A)]$ with $\rho = \pi(A | s)/b(A | s)$, $\mathbb{E}_b[\rho] = 1$; on the coin 2 or 0. Needs $b > 0$ where $\pi > 0$: coverage.
- H weeks: $P_\pi(\tau)/P_b(\tau) = \prod_t \rho_t$, the hidden P cancels; $\mathbb{E}_b[\rho_{0:H-1} G_H] = J_H(\pi)$. By hand: 147.52 or 0 for 17.42.
- The price: one log in 2^H matches and counts 2^H ; spread 470.7 at $H = 8$ against 8.18 for an episode of π^* itself; 2^{-44} at 44 weeks.
- Per-decision: $\sum_t \gamma^t \rho_{0:t} R_{t+1}$, unbiased, later ratios average to one; 70.72 by hand; spread 104.9 at $H = 8$. Weighted: average the matched logs; bounded; 0 with no match.
- Bootstrap: $V(S_t) \leftarrow V(S_t) + \alpha \rho_t (R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$, one ratio per backup; ratios = calls of b left inside the target: $H, n, 1, 0$. Half the steps count. It trusts the state.
- By computer, 100 logs at $H = 8$: errors 45.61, 23.64, 10.30; with the table, the step shrinking, 1.59 on $V^{\pi^*}(G)$, against 0.82 on-policy. A log that never tried the call prices nothing: coverage, not the estimator.

References

- Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 30–37. Morgan Kaufmann, 1995.
- Emma Brunskill. CS 234: Reinforcement learning. Stanford University course materials, 2026. URL <https://web.stanford.edu/class/cs234/>. Winter 2026 offering.
- Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260): 663–685, 1952.
- Nan Jiang. Notes on importance sampling and policy gradient. Lecture notes, University of Illinois Urbana-Champaign, 2023. URL <https://nanjiang.cs.illinois.edu/files/cs542f23/note6.pdf>. November 22, 2023.
- Nan Jiang. CS 443: Reinforcement learning. University of Illinois Urbana-Champaign course materials, 2026. URL <https://nanjiang.cs.illinois.edu/cs443/>. Spring 2026 offering.

- Herman Kahn and Andy W. Marshall. Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.
- Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 5356–5366, 2018.
- Doina Precup, Richard S. Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 759–766, 2000.
- Doina Precup, Richard S. Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 417–424, 2001.
- 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>.