

Lecture 12: Off-Policy Evaluation: Importance Sampling

CS 639: Introduction to Reinforcement Learning

Tengyang Xie

Tuesday, October 13, 2026

Quiz 3

- 15 minutes, 11:00–11:15. Closed book; the formulas you need are printed on the sheet.
- Name, UW NetID, and Campus ID at the top.
- Final answers go in the table at the top: numbers for 1(a)–(b), one letter for 2(a)–(c).
- Exact fractions or decimals are fine. No partial credit.
- Work alone. No notes, phones, or laptops.
- Finished early? Turn the sheet face down and wait. Sheets are collected at 11:15.

Housekeeping

- Quiz 3 is in. Solutions and scores follow on Canvas.
 - Best 5 of 6 quizzes count; no make-ups.
- HW2 is due Wed 10/21 on Gradescope; hard close 48 hours later.
 - Four slip days shared by HW1–HW4, at most two on one assignment; applied automatically.
- Midterm: Tue 10/27 in class, 75 minutes, Lectures 1–14; more details to come.
- Reading for today: Sutton and Barto, Sections 5.5, 5.9, and 7.3 (Precup, Sutton, and Singh 2000, optional).
- Office hours
 - Tengyang Xie: Wed 1–2pm, MH 5530
 - Sarina Xi: Mon 9–11am, MH 5683
 - Yurun Yuan: Fri 9–11am, MH 5683
- Questions: Piazza first; private posts are fine.

Three estimators against the horizon

weeks H	$J_H(\pi^*)$	matched logs of 100 (mean)	ordinary IS	weighted IS	per-decision IS	sets with no match	Monte Carlo, 100 episodes of π^*
2	14.00	25.0	2.50	0.34	1.69	0.000	0.17
4	20.01	6.2	8.10	3.29	3.63	0.000	0.71
6	24.56	1.5	19.84	12.32	6.26	0.199	0.78
8	28.31	0.4	45.61	23.64	10.30	0.673	0.82
10	31.35	0.1	103.69	29.90	16.75	0.902	0.84
8, with 10,000 logs	28.31	38.8	4.86	1.40	1.08	—	—

Target π^* = repair when broken, from a good machine; $n = 100$ logs of H weeks by the random manager's coin (run or repair with probability $\frac{1}{2}$, written down); the three estimates from the same logs; root-mean-square error against $J_H(\pi^*)$ over 1,000 sets of logs (seed 639). A matched log is one whose every call is the one π^* makes; it carries weight 2^H , the others 0. Weighted IS returns 0 when no log matched. The last column is Monte Carlo on 100 episodes of π^* itself (its spread over 10): the price of never running it. Same 100 logs at $H = 8$, TD(0) with one ratio per backup, the 800 steps replayed with a shrinking step: error 1.59 on $V^{\pi^*}(G) = 44.30$ and 0.95 on J_8 (a fixed step $\alpha = 0.1$ circles at 3.59); one step in two matches, not one log in 256.

Before you leave

- Reading: Sutton and Barto, Sections 5.5, 5.9, and 7.3; Precup, Sutton, and Singh (2000), optional.
- HW2 is due Wed 10/21 on Gradescope; hard close 48 hours later.
- Midterm Tue 10/27 in class, Lectures 1–14; more details to come.
- Lecture 13 (Thu 10/15): function approximation. When the table is too big: features and linear values.