

Lecture 11: Off-Policy Control: Q-Learning

CS 639: Introduction to Reinforcement Learning

Tengyang Xie

Thursday, October 8, 2026

Housekeeping

- HW2 is out on Gradescope (Lectures 5–9); due Wed 10/21; hard close 48 hours later.
 - Four slip days shared by HW1–HW4, at most two on one assignment; applied automatically.
- Quiz 3 is Tue 10/13, the first 15 minutes: Lectures 8–11. Closed book; the formulas you need are printed on the sheet.
- Reading for today: Sutton and Barto, Sections 6.5 and 6.7, Example 6.6 (Szepesvári, Section 4.3.1, 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.

Two learners, three behaviors, by computer

learner, data	greedy policy of Q_t is π^* : share of 200			mean max $ Q_t - Q^* $			$ Q_t - Q^b $ at 100,000	b earns per week
	1,000	10,000	100,000	1,000	10,000	100,000		
week t =	1,000	10,000	100,000	1,000	10,000	100,000		
Q-learning, the random manager's log	0.570	1.000	1.000	16.75	2.38	2.33	52.87	-1.62
SARSA, the random manager's log	0.000	0.025	0.005	51.38	54.88	54.93	3.65	-1.62
Q-learning, the log, $\alpha_t = 10/(10 + N_t(s,a))$	1.000	1.000	1.000	4.23	0.63	0.15	53.51	-1.62
Q-learning, its own coin, $\varepsilon = 0.1$	1.000	0.995	1.000	20.25	2.55	2.65	—	3.35
SARSA, its own coin, $\varepsilon = 0.1$	1.000	0.990	1.000	22.62	6.85	6.85	—	3.34
Q-learning, the repair-when-worn log	0.000	0.000	0.000	32.41	32.41	32.41	—	3.50

The experiment: 200 logs of the random manager (run or repair with probability $\frac{1}{2}$ each, 100,000 weeks from G), and 200 streams in which the learner acts ε -greedily on its own table ($\varepsilon = 0.1$, ties at random); $Q_0 = 0$, $\alpha = 0.1$ unless stated; $N_t(s,a)$ counts the visits to the pair. After t weeks the greedy policy is read off Q_t and compared with $\pi^* = \text{repair when broken}$, and the max-norm distances to Q^* and to the random manager's own Q^b are recorded, averaged over the 200 logs or streams. The last column is what the behavior itself earned per week. The repair-when-worn log never visits B and never runs a worn machine, so four entries stay at 0.

Cliff Walking, by computer

episodes 401–500	return per episode	cliff falls per episode	greedy path read off the final table (of 100 walkers)
SARSA, $\varepsilon = 0.1$	-27.3	0.050	the top row, -17: 75; a detour, -19: 5; a rule that loops: 20
Q-learning, $\varepsilon = 0.1$	-51.0	0.343	the edge, -13: 100
SARSA, $\varepsilon = 0.01$	-16.7	0.006	the middle row, -15: 59; the top row, -17: 28; loops: 12
Q-learning, $\varepsilon = 0.01$	-16.0	0.026	the edge, -13: 100

The task (Sutton and Barto, Example 6.6): a 4×12 grid, start S bottom left, goal G bottom right; every step pays -1; the ten bottom cells between S and G are the cliff: stepping in pays -100 and puts the walker back at S; $\gamma = 1$. The best path hugs the edge, 13 steps (-13); the top row takes 17 (-17). 100 walkers, $\alpha = 0.5$, $Q_0 = 0$, 500 episodes, ties at random. The return column is the sum of rewards per episode over the last 100 episodes, averaged over the walkers: what the exploring walker earns. The greedy path is read off each walker's final table; a table whose greedy arrows loop between two cells never arrives.

Before you leave

- Reading: Sutton and Barto, Sections 6.5 and 6.7, Example 6.6; Szepesvári, Section 4.3.1, optional.
- HW2 is out (Lectures 5–9); due Wed 10/21 on Gradescope; hard close 48 hours later.
- Quiz 3: Tue 10/13, first 15 minutes, Lectures 8–11.
- Lecture 12 (Tue 10/13): off-policy evaluation. Importance sampling: the value of one policy from the data of another.