

Lecture 10: On-Policy Control: SARSA

CS 639: Introduction to Reinforcement Learning

Tengyang Xie

Tuesday, October 6, 2026

Housekeeping

- HW2 is out today 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.
- HW1 closed Sun 10/04; solutions and the common-error note are on Canvas.
- 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 5.4, 6.4, 6.6 (Szepesvári, Section 4.4.1, 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.

Five learners on the machine, by computer

	greedy policy of Q_t is π^* : share of 200 streams				mean error max $ Q_t - Q^* $				earned per week
week $t =$	100	1,000	10,000	100,000	100	1,000	10,000	100,000	
SARSA, $\varepsilon = 0$	0.005	0.005	0.005	0.005	38.79	38.79	38.79	38.79	0.02
SARSA, $\varepsilon = 0.1$	0.055	1.000	0.990	1.000	34.06	22.62	6.85	6.85	3.34
SARSA, $\varepsilon_t = 1/t$	0.010	0.030	0.065	0.110	37.13	35.65	34.41	33.60	0.35
SARSA, $\varepsilon_t = 1/\sqrt{t}$	0.075	1.000	1.000	1.000	33.41	25.56	11.26	3.11	3.69
Expected SARSA, $\varepsilon = 0.1$	0.015	1.000	1.000	1.000	34.03	22.48	6.13	6.17	3.35
π^* in the long run									3.71
softened π^* ($\varepsilon = 0.1$) in the long run									3.35

The experiment: start the machine at G with $Q_0 = 0$, $\alpha = 0.1$, and ties broken uniformly at random; run 100,000 weeks: one stream. After t weeks the greedy policy is read off Q_t and compared with $\pi^* = \text{repair when broken}$, and the error $\max_{\{s,a\}} |Q_t(s,a) - Q^*(s,a)|$ is recorded; both are averaged over 200 independent streams. The last column is the average reward per week over the whole stream. The softened π^* is ε -greedy with respect to its own values; its table $Q^{\{\pi_\varepsilon\}}$ sits 4.04 from Q^* .

Before you leave

- Reading: Sutton and Barto, Sections 5.4, 6.4, 6.6; Szepesvári, Section 4.4.1, 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 11 (Thu 10/08): off-policy control. Q-learning: the optimality backup turned into a learning rule.