

Lecture 6: Policy Iteration

CS 639: Introduction to Reinforcement Learning

Tengyang Xie

Tuesday, September 22, 2026

Housekeeping

- HW1 is due Fri 10/02 on Gradescope; hard close 48 hours later.
 - Four slip days shared by HW1–HW4, at most two on one assignment; applied automatically.
- Quiz 2: Tue 09/29, first 15 minutes of class, closed book, Lectures 4–7.
 - Best 5 of 6 quizzes count; no make-ups.
- 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.

From each of the eight, by computer

$\pi_0(\mathbf{G}), \pi_0(\mathbf{W}), \pi_0(\mathbf{B})$		policies visited after π_0	evaluations
run, run, run	never repair	repair when worn → repair when broken	3
run, run, repair	repair when broken	unchanged: stops at once	1
run, repair, run		repair when worn → repair when broken	3
run, repair, repair	repair when worn	repair when broken	2
repair, run, run		never repair → repair when worn → repair when broken	4
repair, run, repair		never repair → repair when worn → repair when broken	4
repair, repair, run		never repair → repair when worn → repair when broken	4
repair, repair, repair	always repair	never repair → repair when worn → repair when broken	4

Exact evaluation, greedy improvement keeping the current action at ties, stop when unchanged. Every run ends at “repair when broken” with V^* exactly; every run but one passes through “repair when worn”; the longest takes four evaluations of the eight possible.

The chain of the proof, by computer

k	G	W	B	$(\mathcal{J}^{\pi'})^k V^{\pi}: \pi' \text{ for } k \text{ weeks, then } \pi$
0	27.5	10	0	$\pi = \text{never repair, forever: } V^{\pi}$
1	27.5	14.75	14.75	one week of π': the hypothesis, $\mathcal{J}^{\pi'} V^{\pi} = \mathcal{J} V^{\pi}$
2	28.93	14.75	14.75	two weeks
3	29.78	16.03	16.03	three weeks
4	30.68	16.80	16.80	four weeks
∞	38.46	24.62	24.62	$\pi' = \text{repair when worn, forever: } V^{\pi'}$

Each row is at least the one above, entry by entry: the chain of the proof, round 0 of the run. G moves only from $k = 2$, when the improvement at W reaches it; the limit is the value of “repair when worn”, $5/0.13 = 38.46$ at G.

Before you leave, and next time

- Next time, Thu 09/24: what a round costs against an iteration; how much to evaluate between improvements (m iterations of evaluation: value iteration at $m = 1$, policy iteration at $m = \infty$); what a policy read off approximate values is worth; generalized policy iteration, the loop behind every method to come.
- Reading: Sutton & Barto, Sections 4.1–4.3 (today); Sections 4.4–4.7 for next time.
- Quiz 2: Tue 09/29, first 15 minutes of class, closed book, Lectures 4–7. HW1 is due Fri 10/02 on Gradescope.