

Lecture 5: Value Iteration

CS 639: Introduction to Reinforcement Learning

Tengyang Xie

Thursday, September 17, 2026

Housekeeping

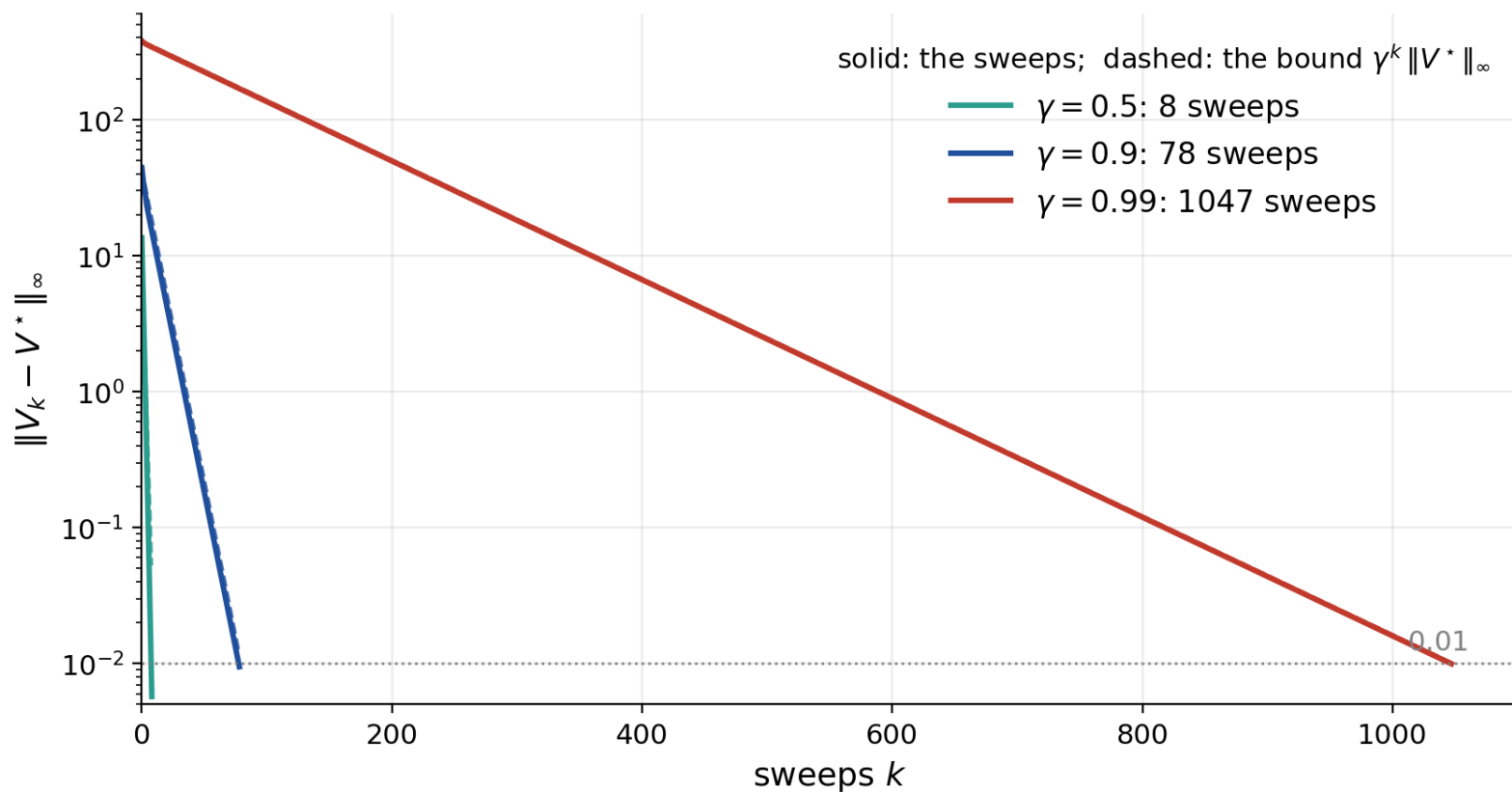
- HW1 is out later today (latest tomorrow): due 2 weeks from releasing on Gradescope; hard close 48 hours later.
 - Four slip days shared by HW1–HW4, at most two on one assignment; applied automatically.
- HW0 and Quiz 1 solutions are on Canvas.
- Office hours
 - Tengyang Xie: Wed 1–2pm, MH 5530
 - Sarina Xi: Mon 9–11am, MH 5683
 - Yurun Yuan: Fri 9–11am, MH 5683
- Questions: Piazza + Feedback form

Eighty iterations, by computer

k	$V_k(G)$	$V_k(W)$	$V_k(B)$	$\ V_k - V^*\ $	ratio	greedy w.r.t. V_k
0	0	0	0	44.30		never repair
1	8	4	0	36.30	0.82	never repair
2	14	6.4	0	30.30	0.83	repair when broken
3	18.32	7.84	2.6	27.27	0.90	repair when broken
4	21.34	9.48	6.49	23.39	0.86	repair when broken
5	23.65	11.64	9.21	20.77	0.89	repair when broken
10	32.08	20.18	17.65	12.23	0.90	repair when broken
20	40.04	28.14	25.61	4.26	0.90	repair when broken
40	43.79	31.89	29.36	0.52	0.90	repair when broken
78	44.29	32.40	29.86	0.0095	0.90	repair when broken
∞	44.30	32.41	29.87	0		repair when broken

From $V_0 = 0$, $\gamma = 0.9$, every entry of V_{k+1} computed from the whole of V_k . The policy is optimal from iteration 2; the values are within 0.01 after 78 iterations (88 guaranteed). The ratio settles at exactly 0.9: under the optimal policy the machine never stops, so a reward k weeks ahead is missing from V_k and is worth 0.9^k .

The price of the horizon



γ	iterations to 0.01	guaranteed	$1/(1-\gamma)$
0.5	8	11	2
0.9	78	88	10
0.99	1047	1146	100

The policy read off V_k is “repair when broken” from $k = 2$ on at all three discounts. The values take 8, 78, 1047 iterations: the horizon $1/(1 - \gamma)$ is paid for in iterations. One state, reward 1 every week: the error is exactly $\gamma^k/(1 - \gamma)$. At $\gamma = 0.99$: 459 iterations to come within 1 of $V^* = 100$.