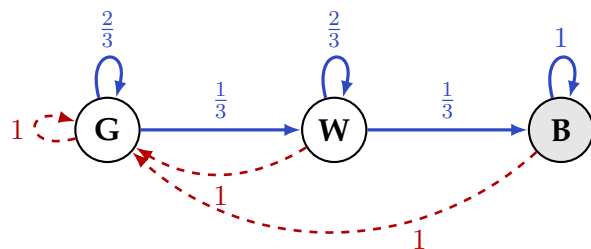


Lecture 7: Convergence Rates, Certificates, and Generalized Policy Iteration

How to use these notes. Board content in board order: figures, equations, tables, one-line statements. Today's models: Lecture 3's machine and, for two questions, Lecture 5's chain. Assumed known: Lecture 5's contraction, iteration count, stopping rule, and residual certificate; Lecture 6's policy iteration with the tie rule, the improvement theorem and its chain, "unchanged means optimal", the bound $|\mathcal{A}|^{|\mathcal{S}|}$, and the advantage table. Reading: [Sutton and Barto \(2018, Sections 4.4–4.7\)](#). Throughout the course, the notes draw on the public materials of Stanford CS 234 ([Brunskill, 2026](#)), UIUC CS 443 ([Jiang, 2026](#)), and Cornell CS 4789 ([Sun, 2025](#)); some examples are adapted from them.

Lecture 6 Left a Count

The machine again: good, worn, or broken; run or repair; $\gamma = 0.9$; it starts good.



$r(s, a)$	run	repair
G	8	-10
W	4	-10
B	0	-10

This week's profit, fixed once (s, a) is.

Where each action leads. Solid: run. Dashed: repair, always back to G.

	Value iteration (Lecture 5)	Policy iteration (Lecture 6)
One step costs	one lookahead per pair: $ \mathcal{S} ^2 \mathcal{A} $	one solve, $ \mathcal{S} ^3$, then the same lookaheads
Steps taken, machine	78 iterations to an error of 0.01	rounds 0, 1, 2; the stop at round 2
Steps guaranteed	$\frac{\log(R_{\max}/((1-\gamma)\epsilon))}{\log(1/\gamma)}$ iterations: 88	$ \mathcal{A} ^{ \mathcal{S} }$ rounds: 8 here, 262,144 on the Gridworld
At the stop	certifies $\ V_k - V^*\ _{\infty} \leq \gamma\theta/(1-\gamma)$	ends exactly: V^* and π^*
The run sees	the change $\ V_{k+1} - V_k\ _{\infty}$	the labels standing still

What Lectures 5 and 6 left. Two guarantees a world apart: 88 grows like $1/(1-\gamma)$, $|\mathcal{A}|^{|\mathcal{S}|}$ like a power of the state count. Today the second column gets the first column's count, and a certificate of its own.

Known from Lectures 3–6. For any table $V \in \mathbb{R}^{|S|}$ and any policy π : $Q_V(s, a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) V(s')$, $(\mathcal{T}^\pi V)(s) = \sum_a \pi(a | s) Q_V(s, a)$, $(\mathcal{T}V)(s) = \max_a Q_V(s, a)$.

- $\mathcal{T}^\pi V \leq \mathcal{T}V$, with equality iff π is greedy w.r.t. V ; $V^* \geq V^\pi$ for every π . [Lecture 4](#)
- Both are γ -contractions in $\|\cdot\|_\infty$; $V^* = \mathcal{T}V^*$ and $V^\pi = \mathcal{T}^\pi V^\pi$ are the only solutions. [Lectures 3, 5](#)
- Residual certificate: $\|V - V^*\|_\infty \leq \|\mathcal{T}V - V\|_\infty / (1 - \gamma)$ for any table V . [Lecture 5](#)
- Improvement theorem: $\mathcal{T}^{\pi'} V^\pi \geq V^\pi \Rightarrow V^\pi \leq \mathcal{T}^{\pi'} V^\pi \leq (\mathcal{T}^{\pi'})^2 V^\pi \leq \dots \rightarrow V^{\pi'}$. [Lecture 6](#)
- Advantages $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$: the step changes the action where an entry is positive; no positive entry, unchanged, optimal. [Lecture 6](#)

How many rounds does policy iteration need? How can a run certify that it may stop before the labels stand still? And what is a policy read off approximate values worth?

Today's targets.

1. Theorem: each round gains at least an iteration, $\|V^* - V^{\pi_{k+1}}\|_\infty \leq \gamma \|V^* - V^{\pi_k}\|_\infty$; three lines from Lecture 6's chain and Lecture 5's contraction.
2. Rounds: a loss of ε within Lecture 5's count; an optimal policy once $\gamma^k E_0$ falls below the gap Δ . Polynomial in the horizon, not $|\mathcal{A}|^{|S|}$.
3. A stopping rule for policy iteration: the largest advantage δ_k certifies a loss of at most $\delta_k / (1 - \gamma)$; $\delta_k = 0$ is Lecture 6's stop.
4. Evaluate less, then evaluate wrongly: the dial m , with two iterations beating the exact evaluation and five reading the worse policy; the bound $2\gamma\varepsilon / (1 - \gamma)$ (stated), attained on the machine.
5. Generalized policy iteration: the loop at any grain; the map of the second half.

Theorem: Each Round Gains at Least an Iteration

Theorem (the rate of policy iteration) ([Jiang, 2024](#), Theorem 3), ([Agarwal et al., 2022](#), Section 1.3.2). Let $\pi_0, \pi_1, \pi_2, \dots$ be the policies of Lecture 6's policy iteration (exact evaluation, greedy improvement). For every k ,

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq \gamma \|V^* - V^{\pi_k}\|_\infty, \quad \text{hence} \quad \|V^* - V^{\pi_k}\|_\infty \leq \gamma^k \|V^* - V^{\pi_0}\|_\infty.$$

The loss of the policy falls by at least the factor an iteration of value iteration takes off the error of a table.

Why, in three lines.

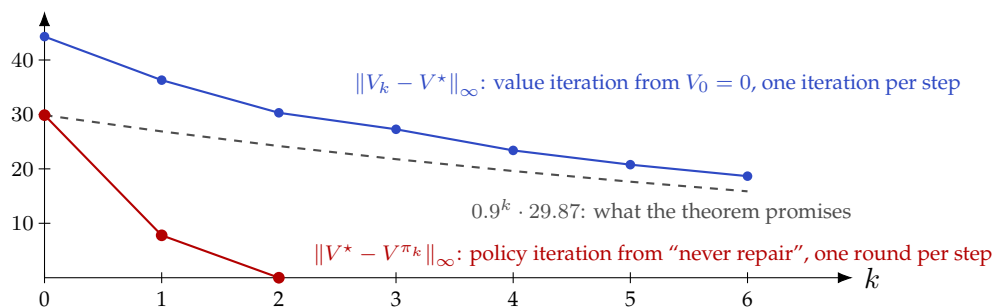
$$V^{\pi_{k+1}} \geq \mathcal{T}^{\pi_{k+1}} V^{\pi_k} = \mathcal{T} V^{\pi_k} \quad \text{Lecture 6's chain, second term; } \pi_{k+1} \text{ greedy w.r.t. } V^{\pi_k} \text{ (Lecture 4)}$$

$$0 \leq V^* - V^{\pi_{k+1}} \leq V^* - \mathcal{T} V^{\pi_k} = \mathcal{T} V^* - \mathcal{T} V^{\pi_k} \quad V^* \geq V^\pi \text{ and } V^* = \mathcal{T} V^*$$

$$\|V^* - V^{\pi_{k+1}}\|_\infty \leq \|\mathcal{T} V^* - \mathcal{T} V^{\pi_k}\|_\infty \leq \gamma \|V^* - V^{\pi_k}\|_\infty$$

a nonnegative vector; Lecture 5's contraction $\square$

Machine: $\mathcal{T} V^{\pi_0} = (27.5, 14.75, 14.75) \leq V^{\pi_1} = (38.46, 24.62, 24.62)$ and $\mathcal{T} V^{\pi_1} = (38.46, 26.15, 24.62) \leq V^{\pi_2} = V^*$; losses $29.87 \rightarrow 7.79 \rightarrow 0$, ratios 0.26 and 0 against the promised 0.9.



The two runs on the machine, step by step. Value iteration loses at most a factor γ per iteration and here nearly attains it (Lecture 5); policy iteration is promised the same factor per round (dashed) and takes 0.26, then 0.

- The rate is value iteration’s; only the start differs: $E_0 := \|V^* - V^{\pi_0}\|_\infty \leq 2R_{\max}/(1 - \gamma)$, 200 here, 29.87 from “never repair”.
- At least one iteration per round, often many: round 0 was worth nearly 13 iterations ($0.9^{12.8} = 0.26$), round 1 all of them.
- Needs V^{π_k} exact, in the first line. What a cheaper evaluation does to the round: two sections on.

Example. Lecture 6’s chain from “quit everywhere”: $V^{\pi_0} = (1, 1, 1, 1)$, $V^{\pi_1} = (1, 1, 1, 10)$, $V^{\pi_2} = (1, 1, 9, 10)$. Compare $\mathcal{T} V^{\pi_k}$ with $V^{\pi_{k+1}}$.

$\mathcal{T}(1, 1, 1, 1) = (1, 1, 1, 10)$ and $\mathcal{T}(1, 1, 1, 10) = (1, 1, 9, 10)$: equality at every round. There a round is exactly one iteration; losses 9, 8, 7.1, 6.29, 0.

Reading: a round is a Newton step. Policy iteration is Newton’s method on $V - \mathcal{T}V = 0$ (Puterman, 1994, Section 6.4): on the piece where the greedy policy is π_{k+1} the map is $V \mapsto (I - \gamma P^{\pi_{k+1}})V - r^{\pi_{k+1}}$, and the Newton step from V^{π_k} lands on $V^{\pi_{k+1}}$, the solve. The theorem is the guaranteed part, a fixed-point step’s factor γ ; the handful of rounds seen in practice is Newton’s fast convergence near the answer.

How Many Rounds

Corollary 1: rounds to a loss of ε .

$$\|V^* - V^{\pi_k}\|_\infty \leq \gamma^k E_0 \leq \varepsilon \quad \text{once} \quad k \geq \frac{\log(E_0/\varepsilon)}{\log(1/\gamma)} \quad \text{the theorem } k \text{ times; solve for } k$$

$$E_0 \leq \frac{2R_{\max}}{1-\gamma} \Rightarrow k = \frac{\log(2R_{\max}/((1-\gamma)\varepsilon))}{\log(1/\gamma)} \text{ rounds always suffice}$$

Lecture 5's count, with $2R_{\max}$ for $R_{\max}$

Machine, $\varepsilon = 0.01$: 94 rounds from $E_0 \leq 200$, 76 from $E_0 = 29.87$; Lecture 5 guaranteed 88 iterations. The run: 2 rounds.

Corollary 2: an optimal policy, exactly. Let the *gap* be the smallest loss of a policy that is not optimal:

$$\Delta := \min\{\|V^* - V^\pi\|_\infty : \pi \text{ deterministic, not optimal}\} > 0$$

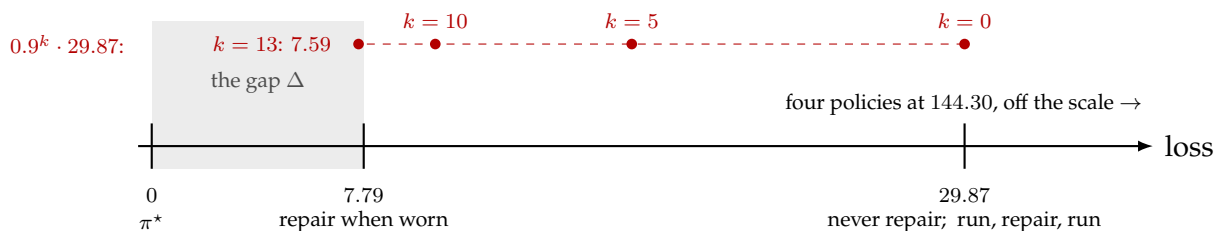
finitely many policies; each non-optimal one loses somewhere

$$\gamma^k E_0 < \Delta \Rightarrow \|V^* - V^{\pi_k}\|_\infty < \Delta \Rightarrow \pi_k \text{ optimal} \quad \text{the theorem; a loss below } \Delta \text{ can only be 0}$$

so $\pi_k = \pi^*$ for every $k > \frac{\log(E_0/\Delta)}{\log(1/\gamma)}$, and the run stops there at the latest

Lecture 6: the tie rule returns an optimal policy unchanged

Machine: the seven non-optimal policies lose 7.79 (repair when worn), 29.87 (two of them), 144.30 (the four that repair a good machine): $\Delta = 7.79$. $0.9^k \cdot 29.87 < 7.79$ from $k = 13$; the run had π^* at $k = 2$.



The eight policies' losses on one line and the theorem's envelope $\gamma^k E_0$ above it. Once the envelope enters the gap, the only loss left is 0: π_{13} is optimal whatever the path. The path took two rounds.

- Not $|A||S|$: about $\frac{1}{1-\gamma} \log \frac{E_0}{\Delta}$ rounds, polynomial in the horizon, logarithmic in $1/\Delta$. Machine: 13; Lecture 6 allowed 8 only because 2^3 is small.
- Δ is the model's, not the run's: near a tie (Lecture 6's γ^*) it shrinks to 0 and the count blows up while the run still stops in a handful. Bounds free of Δ : the reading box.
- Two ingredients: the rate turned Lecture 5's count into a round count; Lecture 6's

finiteness turned “within ε ” into “exactly”.

γ	$1/(1-\gamma)$	iterations to an error of 0.01		rounds to a loss of 0.01		round k with π_k optimal	
		taken	guaranteed	taken	guaranteed	taken	guaranteed (gap)
0.5	2	8	11	0	12	0	0
0.9	10	78	88	2	94	2	13
0.99	100	1047	1146	2	1215	2	262
0.999	1000	12,819	13,809	2	14,502	2	2828

Growth with the horizon, by computer: value iteration from $V_0 = 0$, policy iteration from “never repair”. Guaranteed iterations and rounds to 0.01 use $R_{\max}/(1-\gamma)$ and $2R_{\max}/(1-\gamma)$; the gap column uses the true E_0 and Δ (7.79, 26.26, 219.03). Every guarantee grows like $1/(1-\gamma)$; the machine’s rounds do not move.

Reading: round counts free of the gap. Ye (2011): at most $|\mathcal{S}|^2|\mathcal{A}| \log(|\mathcal{S}|^2/(1-\gamma))/(1-\gamma)$ rounds from any start, as stated by Agarwal et al. (2022, Section 1.3.2). Scherrer (2016), in one idea: after $k^* = \lceil \log(1/(1-\gamma))/\log(1/\gamma) \rceil + 1$ rounds the theorem has cut the loss by more than $1-\gamma$, which forces some non-optimal action of π_0 out for good; there are at most $|\mathcal{S}|(|\mathcal{A}|-1)$ such pairs, so at most $|\mathcal{S}|(|\mathcal{A}|-1)k^*$ rounds in all. Machine: 3 pairs, $k^* = 23$, at most 69 rounds, whatever the start and whatever Δ . Value iteration has no bound of this kind: the iterations it needs before its greedy policy settles on π^* can exceed any polynomial in $|\mathcal{S}|, |\mathcal{A}|, 1/(1-\gamma)$ (Feinberg et al., 2014).

A Stopping Rule for Policy Iteration

The run cannot see the loss. It can see Lecture 6’s advantage table.

The residual of a policy.

$$\delta_k := \|\mathcal{T}V^{\pi_k} - V^{\pi_k}\|_{\infty} = \max_s \max_a A^{\pi_k}(s, a) \geq 0$$

$(\mathcal{T}V^{\pi})(s) - V^{\pi}(s) = \max_a Q^{\pi}(s, a) - V^{\pi}(s)$; the entry at $\pi(s)$ is 0

The largest entry of the advantage table: the improvement step has already computed it. Machine: 14.75, 1.54, 0.

What it certifies, in three lines.

$$\begin{aligned}
V^* - V^{\pi_k} &= (\mathcal{T}V^* - \mathcal{T}V^{\pi_k}) + (\mathcal{T}V^{\pi_k} - V^{\pi_k}) && \text{add and subtract } \mathcal{T}V^{\pi_k}; V^* = \mathcal{T}V^* \\
\|V^* - V^{\pi_k}\|_\infty &\leq \gamma \|V^* - V^{\pi_k}\|_\infty + \delta_k && \text{Lecture 5's contraction; the definition of } \delta_k \\
\|V^* - V^{\pi_k}\|_\infty &\leq \frac{\delta_k}{1-\gamma}, \quad \|V^* - V^{\pi_{k+1}}\|_\infty \leq \frac{\gamma \delta_k}{1-\gamma} && \text{solve for the norm; then the theorem}
\end{aligned}$$

Lecture 5's residual certificate at $V = V^{\pi_k}$, where the error of the table is the loss of the policy.

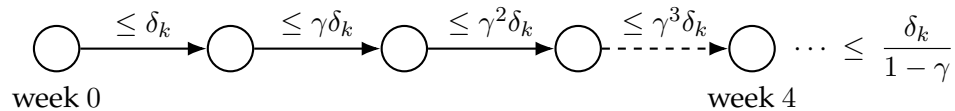
Stopping rule for policy iteration. Stop at the first round with

$$\delta_k \leq (1 - \gamma) \varepsilon, \quad \text{and return } \pi_k : \quad \|V^* - V^{\pi_k}\|_\infty \leq \varepsilon.$$

$\varepsilon = 0$: $\delta_k = 0$ iff no advantage is positive iff $\pi_{k+1} = \pi_k$ under the tie rule. Lecture 6's stop is this rule at $\varepsilon = 0$.

k	π_k	δ_k	certified for π_k	true loss of π_k	certified for π_{k+1}
0	never repair	14.75	147.5	29.87	132.75
1	repair when worn	1.54	15.38	7.79	13.85
2	repair when broken	0	0	0	—

The rule on the machine: certified $\delta_k/(1-\gamma)$ for π_k and $\gamma\delta_k/(1-\gamma)$ for π_{k+1} , against the truth; 4.9 and 2.0 times the loss. Lecture 5's chain from "quit everywhere": $\delta_k = 9, 8, 7.1, 6.29, 0$, certified 90, 80, 71, 62.9, 0, true loss = δ_k , exactly 10 times. Both runs jump to $\delta_k = 0$ rather than fading, so for any $\varepsilon < 15.38$ the rule stops where Lecture 6's did.



What the certificate charges: the largest advantage forgone every week, forever, discounted. The chain forgoes its one advantage once and the episode ends, so the certificate is ten times the loss; the machine under π^* keeps running, so the two are closer.

- The rule earns its keep when δ_k never reaches 0: the dial (next section) and Lecture 8's estimated model. Exact policy iteration on a finite MDP stops by itself.
- Corollary 1 becomes computable at round 0: $E_0 \leq \delta_0/(1-\gamma) = 147.5$, so 92 rounds certify 0.01 (94 blind, 76 with the true E_0).
- Lecture 5's certificate was nearly exact on the machine, this one is off by 2 to 10: what a certificate charges is a worst case over every model with this γ .

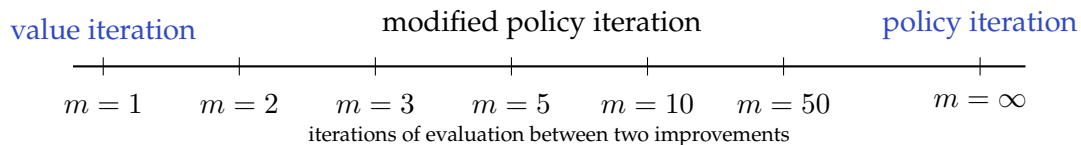
Example. Round 1 on the machine: the advantage table has one positive entry, 1.54 at W. What does the run know about π_1 , about π_2 , and what is true?
 π_1 within $1.54/0.1 = 15.38$ of V^* (true 7.79); π_2 within $0.9 \cdot 15.38 = 13.85$ (true 0). That π_2 is optimal is known only at round 2, when the table has no positive entry.

Evaluate Less: The Dial

Every line so far needed V^{π_k} exact. Lecture 5's run never was, and still read off π^* from $k = 2$: how much evaluation does a round need?

Modified policy iteration (Puterman and Shin, 1978), (Puterman, 1994, Section 6.5)

- Input: the model (P, r, γ) ; $m \in \{1, 2, 3, \dots\}$ or $m = \infty$; a tolerance θ ; V_0 (here 0).
- For $k = 0, 1, 2, \dots$:
 - *Improve*: $\pi_k(s) \leftarrow \operatorname{argmax}_a Q_{V_k}(s, a)$ for every s , keeping $\pi_{k-1}(s)$ when it is among the maximizers.
 - *Evaluate, m iterations*: $V_{k+1} \leftarrow (\mathcal{T}^{\pi_k})^m V_k$, m iterations of Lecture 3's $V \leftarrow r^{\pi_k} + \gamma P^{\pi_k} V$ from V_k ; for $m = \infty$ the solve, $V_{k+1} \leftarrow V^{\pi_k}$.
- Stop when $\|V_{k+1} - V_k\|_\infty \leq \theta$, or for $m = \infty$ when $\pi_k = \pi_{k-1}$; return π_k .



One dial. Every setting improves at every state and then evaluates the new policy for a while; only how long differs.

The two ends of the dial.

$$\begin{aligned}
 m = 1 : \quad V_{k+1} &= \mathcal{T}^{\pi_k} V_k = \mathcal{T} V_k && \pi_k \text{ greedy w.r.t. } V_k \text{ (Lecture 4): value iteration} \\
 m = \infty : \quad V_{k+1} &= V^{\pi_k}, \quad \pi_{k+1} = \text{greedy}(V^{\pi_k}) && \text{policy iteration, started from } \pi_0 = \text{greedy}(V_0)
 \end{aligned}$$

An iteration of value iteration is an improvement followed by one iteration of evaluation (Sutton and Barto, 2018, Section 4.4). Every m converges (Puterman, 1994, Section 6.5): the reading box below.

Example. Read Lecture 5's $V_1 = (8, 4, 0) \rightarrow V_2 \rightarrow V_3(B)$ as improve, then evaluate once.
 $\pi_1 = \text{greedy}(V_1) = \text{never repair}$ (repair column -2.8); $\mathcal{T}^{\pi_1} V_1 = (14, 6.4, 0) = V_2$. $\pi_2 = \text{greedy}(V_2) = \text{repair when broken}$ (repair column 2.6); at B, $-10 + 0.9 \cdot 14 = 2.6 = V_3(B)$.

From $V_0 = 0$, $\pi_0 = \text{greedy}(0) = \text{"never repair"}$ (repair column -10). Round 0 for several settings of the dial:

m	$V_1 = (\mathcal{T}^{\pi_0})^m V_0$	repair, any s	run at W	$\pi_1 = \text{greedy}(V_1)$	$V^{\pi_1}(G)$
1	(8, 4, 0)	-2.8	6.4	never repair	27.5
2	(14, 6.4, 0)	2.6	7.84	repair when broken	44.30
3	(18.32, 7.84, 0)	6.49	8.70	repair when broken	44.30
4	(21.34, 8.70, 0)	9.21	9.22	repair when broken	44.30
5	(23.42, 9.22, 0)	11.08	9.53	repair when worn	38.46
10	(27.03, 9.94, 0)	14.33	9.96	repair when worn	38.46
∞	(27.5, 10, 0)	14.75	10	repair when worn	38.46

Round 0 from $V_0 = 0$ for each m : the evaluation is Lecture 3's iterates of "never repair". Repair pays $-10 + 0.9 V_1(G)$; run at W pays $4 + 0.9(\frac{2}{3} V_1(W) + \frac{1}{3} V_1(B))$; at G run wins throughout, at B repair wins from $m = 2$ on.

Example. Two iterations of evaluation, then the exact evaluation: π_1 in each case, and its worth?

$m = 2$: $V_1 = (14, 6.4, 0)$; repair 2.6; W run $7.84 > 2.6$, B repair $2.6 > 0$: $\pi_1 = \text{repair when broken} = \pi^*$, worth 44.30.

$m = \infty$: $V_1 = (27.5, 10, 0)$; repair 14.75; W repair $14.75 > 10$: $\pi_1 = \text{repair when worn}$, worth 38.46. The exact evaluation of π_0 gave the worse π_1 .

- Decided at W: the repair column climbs with $V_1(G)$, 8 toward 27.5; the run entry with $V_1(W)$, 4 toward 10; they cross between $m = 4$ and $m = 5$.
- Lecture 6's theorem at $m = \infty$: $V^{\pi_1} \geq V^{\pi_0}$, and it holds ($38.46 > 27.5$). It ranks π_1 against π_0 only, never one π_1 against another.
- For finite m it says nothing: V_1 is not V^{π_0} . That $m = 2, 3, 4$ landed on π^* is this machine's luck.
- Whole runs to 0.01, by computer: $m = 1, 2, 3$ take 78 iterations (78, 39, 26 rounds); $m = 5, 10, 50$ take 80, 90, 200 through "repair when worn"; the solve, 3 rounds. Stopped after the $m = 5$ round, the policy loses 5.84 at G.

Reading: why every m converges, from a V_0 with $\mathcal{T}V_0 \geq V_0$ (here $\mathcal{T}0 = (8, 4, 0) \geq 0$).
Write $V_k^{\text{VI}} = \mathcal{T}^k V_0$ for Lecture 5's iterates.

$$V_k \leq \mathcal{T}V_k = \mathcal{T}^{\pi_k} V_k \leq (\mathcal{T}^{\pi_k})^2 V_k \leq \dots \leq (\mathcal{T}^{\pi_k})^m V_k = V_{k+1} \leq V^{\pi_k} \leq V^*$$

Lecture 6's chain from V_k , then its limit

$$V_{k+1} = (\mathcal{T}^{\pi_k})^m V_k \leq (\mathcal{T}^{\pi_k})^{m+1} V_k = \mathcal{T}^{\pi_k} V_{k+1} \leq \mathcal{T}V_{k+1}$$

one more link; $\mathcal{T}^{\pi} \leq \mathcal{T}$: the first line holds at $k+1$

$$V_{k+1} \geq \mathcal{T}V_k \geq \mathcal{T}V_k^{\text{VI}} = V_{k+1}^{\text{VI}} \quad \text{the first line; } \mathcal{T} \text{ monotone and } V_k \geq V_k^{\text{VI}}, \text{ by induction}$$

$$V_k^{\text{VI}} \leq V_k \leq V^* \Rightarrow \|V^* - V_k\|_{\infty} \leq \|V^* - V_k^{\text{VI}}\|_{\infty} \leq \gamma^k \|V^* - V_0\|_{\infty}$$

no m is slower per round than $m = 1$

Machine: every V_1 in the table lies between $(8, 4, 0)$ and $(27.5, 10, 0)$; the rounds to 0.01 fall from 78 as m grows. The per-round rate of the theorem, however, is $m = \infty$'s alone: the tables are not V^{π_k} .

What a Policy Read Off Any Table Is Worth

The $m = 5$ round read the worse policy off an approximate table, and Lecture 5's stop hands over greedy w.r.t. V_k : what is a policy read off an approximate table worth?

Theorem (stated): the loss of a greedy policy (Singh and Yee, 1994). If $\|V - V^*\|_{\infty} \leq \varepsilon$ and π is greedy with respect to V , then

$$V^{\pi}(s) \geq V^*(s) - \frac{2\gamma\varepsilon}{1-\gamma} \quad \text{for every } s.$$

The reason, in three lines (the proof is a recursion on the third):

$$|Q_V(s, a) - Q^*(s, a)| = \gamma \left| \sum_{s'} P(s' | s, a) (V(s') - V^*(s')) \right| \leq \gamma\varepsilon$$

an average of V 's errors, times γ

$$Q^*(s, \pi(s)) \geq Q_V(s, \pi(s)) - \gamma\varepsilon \geq Q_V(s, \pi^*(s)) - \gamma\varepsilon \geq Q^*(s, \pi^*(s)) - 2\gamma\varepsilon$$

the line above; π greedy w.r.t. V ; the line above again

$$\text{one week's choice costs at most } 2\gamma\varepsilon; \text{ forever: } 2\gamma\varepsilon(1 + \gamma + \gamma^2 + \dots) = \frac{2\gamma\varepsilon}{1-\gamma}$$

π may repeat the mistake every week

Example (the bound is attained). $V = V^* + 1.66(-1, 0, +1)$, so $\|V - V^*\|_\infty = 1.66$: the greedy action at B?

run $0.9(29.87 + 1.66) = 28.38$, repair $-10 + 0.9(44.30 - 1.66) = 28.38$: a tie, at $\varepsilon = 2.99/1.8$ exactly. G, W: run, as at V^* .

Broken toward run: never repair, worth $(27.5, 10, 0)$, loss $29.87 = 18 \cdot 1.66$, the bound with equality (Williams and Baird, 1993). The wrong action at B forgoes $(1 - \gamma)V^*(B) = 2.99$ every week.

The same certificate through the residual. For any table V with $\|\mathcal{T}V - V\|_\infty \leq \theta$ and π greedy w.r.t. V , so that $\mathcal{T}^\pi V = \mathcal{T}V$:

$$\|V - V^*\|_\infty \leq \frac{\theta}{1 - \gamma}, \quad \|V - V^\pi\|_\infty \leq \frac{\|\mathcal{T}^\pi V - V\|_\infty}{1 - \gamma} = \frac{\theta}{1 - \gamma}$$

Lecture 5's certificate, for $\mathcal{T}$ and for $\mathcal{T}^\pi$

$$V^* - V^\pi = (\mathcal{T}V^* - \mathcal{T}V) + (\mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi) \leq \gamma \|V^* - V\|_\infty + \gamma \|V - V^\pi\|_\infty \leq \frac{2\gamma\theta}{1 - \gamma}$$

$\mathcal{T}V = \mathcal{T}^\pi V$; two contractions; the line above twice

18θ at $\gamma = 0.9$. Through the error instead, Lecture 5's rule gives $\varepsilon = \gamma\theta/(1 - \gamma)$ and the theorem $2\gamma^2\theta/(1 - \gamma)^2 = 162\theta$: the factor $1/(1 - \gamma)$ paid twice.

The table	Its residual $\ \mathcal{T}V - V\ _\infty$	Certifies the table	Certifies the policy
any V	θ	$\ V - V^*\ _\infty \leq \theta/(1 - \gamma)$ (Lecture 5)	greedy(V) loses at most $2\gamma\theta/(1 - \gamma) = 18\theta$
V_k at value iteration's stop	the change, $\leq \theta$	the same; V_{k+1} within $\gamma\theta/(1 - \gamma)$	the policy read off V_k : 18θ ; read off V_{k+1} through its error: 162θ
V^{π_k} of policy iteration	δ_k , the largest advantage	$\ V^{\pi_k} - V^*\ _\infty \leq$ $\delta_k/(1 - \gamma)$	the same number: the table's error is π_k 's loss; π_{k+1} within $\gamma\delta_k/(1 - \gamma)$

One residual, three readings. Certifying a loss of 0.01 on the machine: 85 iterations through the residual, 105 through the error; policy iteration certifies 0 at round 2.

k	$\varepsilon = \ V_k - V^*\ _\infty$	greedy w.r.t. V_k	true loss $\ V^* - V^\pi\ _\infty$	bound 18ε
0	44.30	never repair	29.87	797
2	30.30	repair when broken	0	545
40	0.52	repair when broken	0	9.3
78	0.0095	repair when broken	0	0.17

The bound against the truth on Lecture 5's iterates: useless early (larger than any value in the problem), loose late (the policy has been π^* since $k = 2$).

- Certificates are worst cases over every model with this γ : at $\theta = 0.001$ the run knows 0.018 (or 0.162), not 0.
- Why $2\gamma\varepsilon$: greedy sees comparisons only; $Q_{V+c} = Q_V + \gamma c$, a constant shift is free. An error that differs between states, $+\gamma\varepsilon$ on one entry and $-\gamma\varepsilon$ on the other, flips one: the Example spends that at B.
- In Q form: $\|Q - Q^*\|_\infty \leq \varepsilon$ and $\pi = \text{greedy}(Q)$ lose at most $2\varepsilon/(1 - \gamma)$, the lookahead being inside Q (Agarwal et al., 2022, Lemma 1.11).
- One-sided errors ($V \geq V^*$ everywhere, or $V \leq V^*$) cost half, $\gamma\varepsilon/(1 - \gamma) = 9\varepsilon$: one swap in the second line is free. Attained at $V^* + (0, 0, 3.32)$.

Generalized Policy Iteration

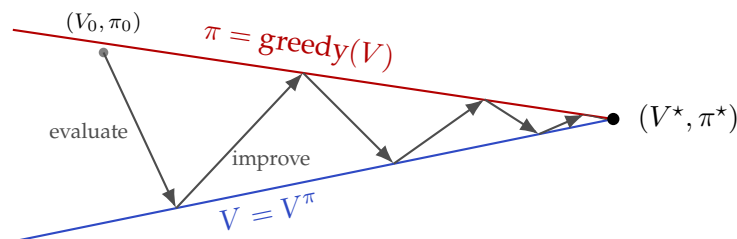
Value iteration, policy iteration, and every setting of the dial are one loop.

Two processes, one loop (Sutton and Barto, 2018, Section 4.6).

- *Evaluation* moves the table toward the policy's values, $V \rightarrow V^\pi$: a solve; m iterations of $\mathcal{T}^\pi$; one iteration; later, averages of sampled returns.
- *Improvement* moves the policy toward greedy w.r.t. the table, $\pi \rightarrow \text{greedy}(V)$: at every state or at some; exactly or nearly (ε -greedy, below).
- Generalized policy iteration: any interleaving of the two that keeps updating every state.

$$V = V^\pi \text{ and } \pi = \text{greedy}(V) \Rightarrow V^\pi = \mathcal{T}^\pi V^\pi = \mathcal{T} V^\pi \Rightarrow (V, \pi) = (V^*, \pi^*)$$

Lecture 6's "unchanged means optimal": both stand still only there



Sutton and Barto's picture of Section 4.6, redrawn. Each process reaches its own line and leaves the other; the only point on both is (V^*, π^*) . Policy iteration takes full steps to each line; a finite m takes partial ones toward the blue line.

Improvement need not be greedy: ε -greedy improves too (Sutton and Barto, 2018, Section 5.4). Let π give every action probability at least $\varepsilon/|\mathcal{A}|$ (ε -soft), and let π' be ε -greedy w.r.t. Q^π : the greedy action with probability $1 - \varepsilon + \varepsilon/|\mathcal{A}|$, each other action with $\varepsilon/|\mathcal{A}|$.

$$(\mathcal{T}^{\pi'} V^\pi)(s) = \frac{\varepsilon}{|\mathcal{A}|} \sum_a Q^\pi(s, a) + (1 - \varepsilon) \max_a Q^\pi(s, a) \geq \sum_a \pi(a | s) Q^\pi(s, a) = V^\pi(s)$$

the weights $\pi(a | s) - \varepsilon/|\mathcal{A}| \geq 0$ sum to $1 - \varepsilon$; a max beats their average

$$\Rightarrow V^{\pi'} \geq V^\pi$$

Lecture 6's proof, unchanged for a stochastic π' : it needs only $P^{\pi'} \geq 0$

Machine, $\varepsilon = 0.1$, from “never repair” softened (run w.p. 0.95): (29.12, 14.43, 5.59) $\rightarrow$ (33.30, 20.07, 19.86) $\rightarrow$ (39.82, 28.87, 25.70), then unchanged: Lecture 6's path, softened, ending 4.49 below V^* at G, the price of exploring.

With the model	Evaluation of π	Improvement
Policy iteration	exact: one solve	greedy, every state
Value iteration	one iteration of $\mathcal{T}^\pi$	greedy, every state
Modified policy iteration	m iterations of $\mathcal{T}^\pi$	greedy, every state
Asynchronous DP (Sutton and Barto, 2018, Section 4.5)	one state at a time, in any order that keeps visiting every state	greedy at the states updated

The loop with the model: four settings of the same two steps, every one of them convergent.

Reading: what the model's absence costs. Without P and r , Q^π is estimated from sampled transitions: the evaluation is noisy, so Lecture 6's theorem is no longer a guarantee, and the dial's $m = 5$ round, which read the worse policy off an approximate table, is the standing warning; a deterministic policy learns nothing about its other actions, so the improvement stays ε -greedy; the two processes run one transition at a time, the policy changing under the evaluation. Convergence is proved for some of these methods and only hoped for in others: the loop is the map, not the guarantee (Sutton and Barto, 2018, Section 4.8). Lecture 8 keeps the model instead: estimate P and r , then run policy iteration on the estimate.

Without the model	Evaluation of π	Improvement	Data
Monte Carlo control	average the sampled returns that followed (s, a)	ε -greedy: mostly greedy, every action kept alive	whole episodes
SARSA	one step toward $R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})$ per transition	ε -greedy at every step	transitions
Q-learning	one step toward $R_{t+1} + \gamma \max_a Q(S_{t+1}, a)$: value iteration's update, one sampled pair at a time	greedy off the table; ε -greedy to collect data	transitions
Actor-critic	a critic tracks V^π or Q^π	a gradient step on a parameterized π toward higher Q	trajectories
Search (AlphaZero)	self-play returns fit a value network	a search tree sharpens the move distribution	a simulator

The loop without the model: each row is an evaluation and an improvement with samples in place of P and r . The rows are named by what they are, all to come.

Example. Control always has a max somewhere. Q-learning's is explicit, $\max_a Q(S_{t+1}, a)$. Where is SARSA's?
 In the choice of A_{t+1} : ε -greedy w.r.t. Q , so $Q(S_{t+1}, A_{t+1})$ is the max with probability $1 - \varepsilon + \varepsilon/|\mathcal{A}|$. Improvement at every step, evaluating a policy that keeps changing.

The Planning Unit, Closed

	Brute force (Lecture 4)	Guess and test (Lecture 4)	Value iteration (Lecture 5)	Policy iteration (Lecture 6)
Computes	V^π for every table π	V^π for one guess, then $\mathcal{T}V^\pi$	$V_k = \mathcal{T}^k V_0$, no policy fixed	V^{π_k} for a chain of guesses, each the greedy step from the last
On the machine	8 solves	1 solve and 6 lookaheads per guess	78 iterations, 468 lookaheads	3 solves and 18 lookaheads
Steps guaranteed	$ \mathcal{A} ^{ \mathcal{S} }$ solves	one solve; no guarantee the guess passes	$\frac{1}{1-\gamma} \log \frac{R_{\max}}{(1-\gamma)\epsilon}$ iterations	the same count to a loss of ϵ ; $\frac{1}{1-\gamma} \log \frac{E_0}{\Delta}$ rounds to π^*
Answer	exact	exact if the guess passes, else a failed guess	within $\gamma\theta/(1-\gamma)$; a policy within 18θ	exact; a loss within $\delta_k/(1-\gamma)$ at every round

Example. Which would you run? (a) A million states, each row of P with a handful of nonzero entries. (b) Twenty states, $\gamma = 0.999$. (c) The exact optimal policy is required. (d) P and r unknown; only sampled transitions. Answers: the table.

Case	Run	Why
(a) 10^6 states, sparse P	value iteration, or the dial at a small m	a solve is $ \mathcal{S} ^3$, out of reach; an iteration is one lookahead per pair, cheap when P is sparse
(b) 20 states, $\gamma = 0.999$	policy iteration	the guarantees grow alike, but a round is worth at least an iteration and usually many: 12,819 iterations against 2 rounds on the machine
(c) exact π^* required	policy iteration	it ends exactly, within $\frac{1}{1-\gamma} \log \frac{E_0}{\Delta}$ rounds; value iteration certifies within
(d) no model, only samples	neither as written	the loop with samples for the model: the table above, and the rest of the course

Summary

Lecture 7 on one board

- Theorem: $V^{\pi_{k+1}} \geq \mathcal{T}V^{\pi_k}$, so $\|V^* - V^{\pi_{k+1}}\|_\infty \leq \gamma \|V^* - V^{\pi_k}\|_\infty$: each round gains at least an iteration. Machine 0.26, then 0; chain exactly one iteration per round.
- Rounds: a loss of ε within $\log(E_0/\varepsilon)/\log(1/\gamma)$, $E_0 \leq 2R_{\max}/(1-\gamma)$ (94 against Lecture 5's 88); π^* exactly once $\gamma^k E_0 < \Delta$ (13; the run took 2). Polynomial in $1/(1-\gamma)$, not $|\mathcal{A}|^{|\mathcal{S}|}$.
- Stopping rule: $\delta_k = \|\mathcal{T}V^{\pi_k} - V^{\pi_k}\|_\infty$, the largest advantage, certifies $\|V^* - V^{\pi_k}\|_\infty \leq \delta_k/(1-\gamma)$; stop at $\delta_k \leq (1-\gamma)\varepsilon$; $\varepsilon = 0$ is Lecture 6's stop. Machine 147.5, 15.38, 0 against 29.87, 7.79, 0.
- The dial: $\pi_k = \text{greedy}(V_k)$, $V_{k+1} = (\mathcal{T}^{\pi_k})^m V_k$; $m = 1$ value iteration, $m = \infty$ policy iteration. Round 0 from zero: $m = 2$ gives π^* , $m = \infty$ "repair when worn"; the theorem ranks π_1 against π_0 only.
- Bound (stated): $\|V - V^*\|_\infty \leq \varepsilon$ gives a loss of at most $2\gamma\varepsilon/(1-\gamma)$; attained at $\varepsilon = 1.66$ (greedy sees comparisons, not distances); through the residual, 18θ . One residual certifies any table, value iteration's stop, and policy iteration's round.
- Generalized policy iteration: $V \rightarrow V^\pi$, $\pi \rightarrow \text{greedy}(V)$, any interleaving, greedy or ε -greedy; both still only at (V^*, π^*) . Every method to come is the loop with samples for the model.

References

- Alekh Agarwal, Nan Jiang, Sham M. Kakade, and Wen Sun. Reinforcement learning: Theory and algorithms. Book draft, 2022. URL <https://rltheorybook.github.io/>.
- Emma Brunskill. CS 234: Reinforcement learning. Stanford University course materials, 2026. URL <https://web.stanford.edu/class/cs234/>. Winter 2026 offering.
- Eugene A. Feinberg, Jefferson Huang, and Bruno Scherrer. Modified policy iteration algorithms are not strongly polynomial for discounted dynamic programming. *Operations Research Letters*, 42(6–7):429–431, 2014.
- Nan Jiang. MDP preliminaries. Lecture notes, University of Illinois Urbana-Champaign, 2024. URL <https://nanjiang.cs.illinois.edu/files/cs542f22/note1.pdf>. Dated October 5, 2019; updated September 27, 2024. Accessed 2026-09-13.
- Nan Jiang. CS 443: Reinforcement learning. University of Illinois Urbana-Champaign course materials, 2026. URL <https://nanjiang.cs.illinois.edu/cs443/>. Spring 2026 offering.
- Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.
- Martin L. Puterman and Moon Chirl Shin. Modified policy iteration algorithms for discounted Markov decision problems. *Management Science*, 24(11):1127–1137, 1978.
- Bruno Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41(3):758–774, 2016.
- Satinder P. Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.
- Wen Sun. CS 4/5789: Introduction to reinforcement learning. Cornell University course materials, 2025. URL https://wensun.github.io/CS4789_spring_2025.html. Spring 2025 offering.
- Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. URL <http://incompleteideas.net/book/the-book-2nd.html>.
- Ronald J. Williams and Leemon C. Baird. Tight performance bounds on greedy policies based on imperfect value functions. Technical Report NU-CCS-93-14, College of Computer Science, Northeastern University, 1993.
- Yinyu Ye. The simplex and policy-iteration methods are strongly polynomial for the Markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36(4):593–603, 2011.