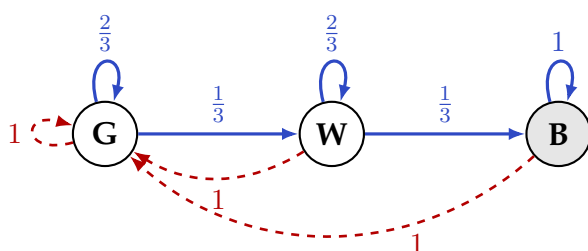


Lecture 6: Policy Iteration

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 a second run, Lecture 5's chain. Assumed known: Lecture 3's linear solve $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$; Lecture 4's V^* , Q^* , the operator $\mathcal{T}$, and the optimality test; Lecture 5's value iteration and contraction. Reading: [Sutton and Barto \(2018, Sections 4.1–4.3\)](#). 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 5 Left an Approximation

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

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

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

	On the board	Status
Lecture 3	$V^\pi = (27.5, 10, 0)$ for $\pi = \text{never repair}$; $Q^\pi(s, \text{repair}) = 14.75$	exact: one solve
Lecture 4	$V^* = (44.30, 32.41, 29.87)$, $\pi^* = (\text{run}, \text{run}, \text{repair})$ greedy w.r.t. V^π above: $(\text{run}, \text{repair}, \text{repair})$, worth (38.46, 24.62, 24.62)	exact: eight solves "better everywhere: proved later"
Lecture 5	$V_k = \mathcal{T}^k V_0 \rightarrow V^*$; π^* read off V_2 ; certified within 0.0099 at $k = 78$	never exact; the run could not know it held π^*

What Lectures 3–5 left. Today proves the third row's promise and replaces the fourth row's approximation.

Two things to iterate on:

tables (Lecture 5): $V_k \xrightarrow{\mathcal{T}} V_{k+1} \xrightarrow{\mathcal{T}} \dots \rightarrow V^*$, the policy a by-product;
policies (today): $\pi_k \xrightarrow{\text{solve}} V^{\pi_k} \xrightarrow{\text{greedy}} \pi_{k+1} \xrightarrow{\text{solve}} \dots$, the values a by-product.

Known from Lectures 3–5. For any table $V \in \mathbb{R}^{|S|}$ and any policy π ,

$$Q_V(s, a) := r(s, a) + \gamma \sum_{s'} P(s' | s, a) V(s'),$$

$$(\mathcal{T}^\pi V)(s) = \sum_a \pi(a | s) Q_V(s, a), \quad (\mathcal{T}V)(s) = \max_a Q_V(s, a) :$$

one lookahead table, two ways of collapsing a row.

- $V^\pi = \mathcal{T}^\pi V^\pi$, the only solution; $Q^\pi = Q_{V^\pi}$. [Lecture 3](#)
- $V^* = \mathcal{T}V^*$, the only solution; $Q^* = Q_{V^*}$. [Lecture 5](#)
- $\mathcal{T}^\pi V \leq \mathcal{T}V$ entry by entry; equality iff π is greedy w.r.t. V . [Lecture 4](#)
- Both are γ -contractions in $\|\cdot\|_\infty$: iterating either converges from any start. [Lecture 5](#)

Evaluate a policy exactly, act greedily on its values, repeat. Why does every round improve, why must the rounds stop, and why is the policy they stop at optimal?

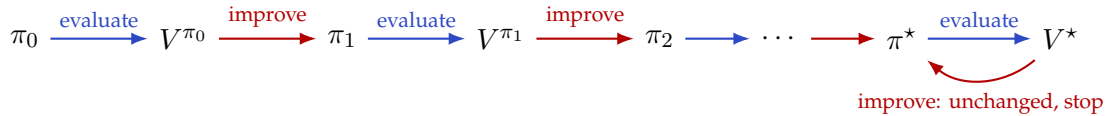
Today's targets.

1. Policy iteration in full: exact evaluation, greedy improvement with a tie rule, stop when unchanged.
2. Two runs: the machine to the end, never repair $\rightarrow$ repair when worn $\rightarrow$ repair when broken $\rightarrow$ stop; Lecture 5's chain, one room per round.
3. The policy improvement theorem, proved; then read on Q tables and as advantages.
4. Why the run stops, and only at an optimal policy.

Policy Iteration

Policy iteration ([Howard, 1960](#)), ([Sutton and Barto, 2018](#), Section 4.3)

- Input: the model (P, r, γ) ; any deterministic π_0 .
- For $k = 0, 1, 2, \dots$:
 - *Evaluate, exactly*: $V^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$ (Lecture 3's linear solve), then its lookahead table $Q^{\pi_k}(s, a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) V^{\pi_k}(s')$.
 - *Improve, greedily*: for every s , $\pi_{k+1}(s) \leftarrow \operatorname{argmax}_a Q^{\pi_k}(s, a)$, keeping $\pi_k(s)$ whenever it is among the maximizers (the tie rule).
- Stop when $\pi_{k+1} = \pi_k$; return π_k and V^{π_k} .



Two alternating steps (Sutton and Barto, 2018, Section 4.3). Blue: a linear solve. Red: one lookahead per pair and a max per state.

- One round: one solve, $O(|S|^3)$, then one lookahead per pair, $O(|S|^2|A|)$. k counts rounds.
- The iterate is a policy; its values are exact at every round. Lecture 5's, on the machine, never were.
- The test compares action labels, $\pi_{k+1} = \pi_k$: no tolerance. What it certifies: two sections on.
- Sutton and Barto (2018, Section 4.3) evaluate by iteration to a tolerance; today exactly, approximately next time.

The Run on the Machine

Round k : evaluate π_k , read π_{k+1} off Q^{π_k} . Round 0 is Lecture 4's failure-mode row with the algorithm's names.

k	π_k	$V^{\pi_k}(G, W, B)$	$Q^{\pi_k}(\cdot, \text{run})$	$Q^{\pi_k}(\cdot, \text{repair})$	$\rightarrow \pi_{k+1}$	error
0	never repair	27.5, 10, 0	27.5, 10, 0	14.75	W, B $\rightarrow$ repair	29.87
1	repair when worn	38.46, 24.62, 24.62	38.46, 26.15, 22.15	24.62	W $\rightarrow$ run	7.79
2	repair when broken	44.30, 32.41, 29.87	44.30, 32.41, 26.89	29.87	none: stop	0

The run from “never repair”: three evaluations, two changes, W changing twice. Run column at the policy's own action = $V^{\pi_k}(s)$; repair column = $-10 + 0.9 V^{\pi_k}(G)$; error = $\|V^* - V^{\pi_k}\|_\infty$.

Example (row 0). π_0 = never repair; Q^{π_0} is on the board. Greedy?

G: run, $27.5 > 14.75$. W: repair, $14.75 > 10$. B: repair, $14.75 > 0$. $\pi_1 = (\text{run, repair, repair})$, repair when worn.

Row 1's evaluation by hand. W and B repair back to G, so

$$\begin{aligned}
 V(W) &= V(B) = -10 + 0.9 V(G), \\
 V(G) &= 8 + 0.9 \left(\frac{2}{3} V(G) + \frac{1}{3} (-10 + 0.9 V(G)) \right) = 5 + 0.87 V(G) \Rightarrow V(G) = 5/0.13 = 38.46, \\
 V(W) &= V(B) = -10 + 0.9 \cdot 38.46 = 24.62, \quad V^{\pi_1} - V^{\pi_0} = (10.96, 14.62, 24.62) > 0.
 \end{aligned}$$

Example (row 1). Under $V^{\pi_1} = (38.46, 24.62, 24.62)$: at W, run or repair?

Run $4 + 0.9 \cdot 24.62 = 26.15$; repair $-10 + 0.9 \cdot 38.46 = 24.62$. Run: W switches back.

G run $8 + 0.9(\frac{2}{3} \cdot 38.46 + \frac{1}{3} \cdot 24.62) = 38.46$; B run $0.9 \cdot 24.62 = 22.15$.

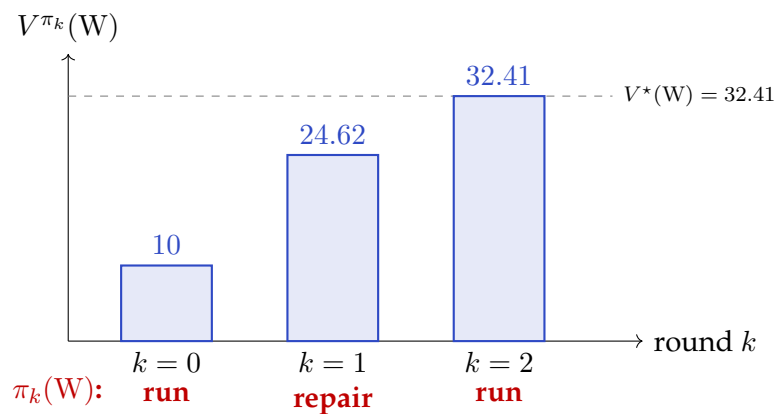
$\pi_2 = (\text{run}, \text{run}, \text{repair})$, repair when broken.

Row 2's evaluation: Lecture 4's hand solve,

$$V^{\pi_2} = (3500, 2560, 2360)/79 = (44.30, 32.41, 29.87).$$

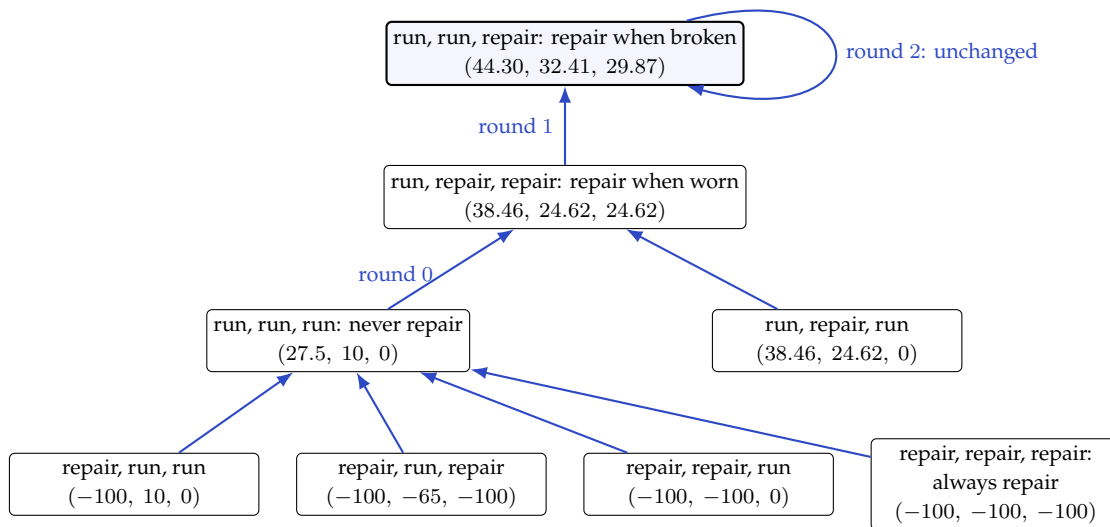
Example (row 2). Any row of Q^{π_2} whose maximum is not π_2 's own action?

G $44.30 > 29.87$; W $32.41 > 29.87$; B $29.87 > 0.9 \cdot 29.87 = 26.89$. None: $\pi_3 = \pi_2$, stop.



W's value rises every round while its action goes run, repair, run: monotone in values, not in actions.

- Each round is greedy w.r.t. a different table: V^{π_0} prices W at 10, V^{π_1} at 24.62.
- Better in every state at every round (the table's columns); value iteration's errors were 44.30, 36.30, 30.30, ... (Lecture 5).
- Stop $\Rightarrow V^{\pi_2} = \mathcal{T}V^{\pi_2} \Rightarrow V^{\pi_2} = V^*$ exactly: the certificate value iteration never had. Two sections on.



One round from each of the eight policies, by computer (Lecture 4's values in the boxes). Every path funnels through "repair when worn"; none visits a policy twice; the longest takes four evaluations.

Example. Can a run ever arrive at "always repair"? Which policies are visited only as π_0 ?

No: no arrow enters it. After round 0 every policy visited is greedy w.r.t. some V^π , and at G run beats repair in all eight lookahead tables (by computer), so the funnel's bottom row is never entered.

Example. Start from "always repair". Round 0?

$V(s) = -10 + 0.9 V(G)$ for every s , so $V^{\pi_0} = (-100, -100, -100)$. Run column $(8, 4, 0) - 90 = (-82, -86, -90) > -100$: $\pi_1 = \text{never repair}$, up by $(127.5, 110, 100)$; then the run above. Four evaluations in all.

Why Every Round Improves

The improvement step, as an inequality. With V^π exact and π' greedy with respect to it,

$$Q^\pi(s, \pi'(s)) = \max_a Q^\pi(s, a) \geq Q^\pi(s, \pi(s)) = V^\pi(s) \quad \text{at every } s,$$

$$\text{i.e.} \quad \mathcal{T}^{\pi'} V^\pi = \mathcal{T} V^\pi \geq \mathcal{T}^\pi V^\pi = V^\pi.$$

One week of π' , then π forever, is never worse than π (Lecture 3's Q^π). Round 0: $(27.5, 14.75, 14.75) \geq (27.5, 10, 0)$.

Theorem (policy improvement) (Sutton and Barto, 2018, Section 4.2). Let π and π' be deterministic policies with

$$Q^\pi(s, \pi'(s)) \geq V^\pi(s) \text{ for every } s, \quad \text{i.e.} \quad \mathcal{T}^{\pi'} V^\pi \geq V^\pi.$$

Then $V^{\pi'}(s) \geq V^\pi(s)$ for every s . If the hypothesis is strict at some s , so is the conclusion at that s . **For a greedy π' , improvement is strict in at least one state unless π is optimal.**

Proof. Plan: $(\mathcal{T}^{\pi'})^k V^\pi$ never falls as k grows (Step 2) and tends to $V^{\pi'}$ (Limit).

Step 1. $U \leq W \Rightarrow \mathcal{T}^{\pi'} U \leq \mathcal{T}^{\pi'} W$ $\mathcal{T}^{\pi'} W - \mathcal{T}^{\pi'} U = \gamma P^{\pi'}(W - U)$ and $P^{\pi'} \geq 0$ (Lecture 3)

Step 2. $V^\pi \leq \mathcal{T}^{\pi'} V^\pi$ *the hypothesis*

$$\mathcal{T}^{\pi'} V^\pi \leq \mathcal{T}^{\pi'} (\mathcal{T}^{\pi'} V^\pi) \quad \text{Step 1 with } U = V^\pi, W = \mathcal{T}^{\pi'} V^\pi$$

$$V^\pi \leq \mathcal{T}^{\pi'} V^\pi \leq (\mathcal{T}^{\pi'})^2 V^\pi \leq \dots \leq (\mathcal{T}^{\pi'})^k V^\pi \leq \dots \quad \text{repeat Step 1}$$

Limit. $V^{\pi'} = \mathcal{T}^{\pi'} V^{\pi'} = (\mathcal{T}^{\pi'})^2 V^{\pi'} = \dots = (\mathcal{T}^{\pi'})^k V^{\pi'}$ *fixed point of $\mathcal{T}^{\pi'}$ (Lecture 3)*

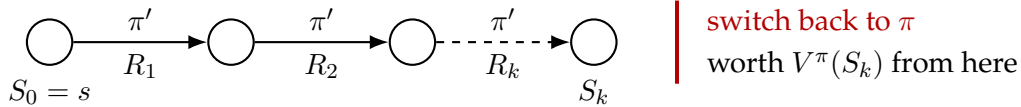
$$\left\| (\mathcal{T}^{\pi'})^k V^\pi - V^{\pi'} \right\|_\infty = \left\| (\mathcal{T}^{\pi'})^k V^\pi - (\mathcal{T}^{\pi'})^k V^{\pi'} \right\|_\infty \quad \text{the line above}$$

$$\leq \gamma^k \left\| V^\pi - V^{\pi'} \right\|_\infty \rightarrow 0 \quad \text{Lecture 5's contraction, } k \text{ times}$$

$$V^\pi \leq \mathcal{T}^{\pi'} V^\pi \leq V^{\pi'} \quad k \rightarrow \infty \text{ in the chain: } \leq \text{ survives the limit}$$

Strict. $V^{\pi'}(s) \geq (\mathcal{T}^{\pi'} V^\pi)(s) = Q^\pi(s, \pi'(s)) > V^\pi(s)$

the line above, at an s where the hypothesis is strict $\square$



$(\mathcal{T}^{\pi'})^k V^\pi(s)$: π' for k weeks, then π forever (Lecture 5's terminal-payout reading). The chain: a longer deviation is never worse; π' forever is best of all.

k	G	W	B	$(\mathcal{T}^{\pi'})^k V^\pi$: π' for k weeks, then π
0	27.5	10	0	π = never repair, forever: V^π
1	27.5	14.75	14.75	one week of π' : the hypothesis, $\mathcal{T}^{\pi'} V^\pi = \mathcal{T} V^\pi$
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	$\pi' =$ repair when worn, forever: $V^{\pi'}$

The chain on the machine, round 0: each row is at least the one above, entry by entry. G moves only from $k = 2$: a week for W's improvement to reach it.

- Step 2 in weeks: rows $k + 1$ and k both start with a week of π' , then pay out rows k and $k - 1$; the larger payout wins (Step 1).
- Nothing new: Step 1 is Lecture 3's matrix form, the limit Lecture 5's contraction, the hypothesis Lecture 4's $\mathcal{T}^\pi V \leq \mathcal{T}V$; $\gamma < 1$ only in the limit.
- Strict where the hypothesis is strict, and possibly elsewhere: at G, $27.5 = 27.5$, yet $V^{\pi'}(G) = 38.46$; G leads to W.
- Needs V^π exact: with an approximation the first $\leq$ can fail and the step can make things worse. Next time.

Example. From “never repair”, two changes satisfy the hypothesis: B alone, giving π^* (44.30 at G); W and B, the greedy step (38.46). Why does the theorem not rank them? It compares each π' with π only, $V^{\pi'} \geq V^\pi$; nothing between two π' . Greedy takes every change V^π can see; round 1 undid W.

Corollary: the same on the lookahead table.

$$Q^{\pi'}(s, a) - Q^\pi(s, a) = \gamma \sum_{s'} P(s' | s, a) [V^{\pi'}(s') - V^\pi(s')] \geq 0$$

$Q^\pi = Q_{V^\pi}$ (Lecture 3), then the theorem

So $Q^{\pi_{k+1}} \geq Q^{\pi_k}$ entry by entry. Machine: read the run table's Q columns downward, $14.75 \rightarrow 24.62 \rightarrow 29.87$ and $(27.5, 10, 0) \rightarrow (38.46, 26.15, 22.15) \rightarrow (44.30, 32.41, 26.89)$. On Q tables (Sutton and Barto, 2018, Exercise 4.5): $\pi_{k+1} = \text{greedy}(Q^{\pi_k})$, then $Q^{\pi_{k+1}}(s, a) = r(s, a) + \gamma \sum_{s'} P(s' | s, a) Q^{\pi_{k+1}}(s', \pi_{k+1}(s'))$: the form that survives without a model, later in the course.

The step, in the language of advantages. $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$: what one week of a , then π forever, gains over π .

$A^\pi(s, \pi(s)) = 0$, $\max_a A^\pi(s, a) \geq 0$, $\pi'(s) = \arg\max_a A^\pi(s, a)$ is the greedy step;
the hypothesis: $A^\pi(s, \pi'(s)) \geq 0 \ \forall s$; the stop: $\max_a A^{\pi_k}(s, a) = 0 \ \forall s$.

k	π_k	$A^{\pi_k}(\cdot, \text{run}): \text{G, W, B}$			$A^{\pi_k}(\cdot, \text{repair}): \text{G, W, B}$			positive entries
0	never repair	0	0	0	-12.75	4.75	14.75	W, B: repair
1	repair when worn	0	1.54	-2.46	-13.85	0	0	W: run
2	repair when broken	0	0	-2.99	-14.43	-2.53	0	none: stop

The run as advantages. The policy's own action has 0; the step changes the action wherever an entry is positive; the stop is a table with no positive entry.

Where the gain comes from: the performance difference lemma (Jiang, 2024, Proposition 2); stated here, proved with policy gradient methods. With $x_s^{\pi'}(s') := \sum_{t \geq 0} \gamma^t \Pr_{\pi'}(S_t = s' \mid S_0 = s)$, the discounted weeks π' spends at s' from s ,

$$V^{\pi'}(s) - V^{\pi}(s) = \sum_{s'} x_s^{\pi'}(s') A^{\pi}(s', \pi'(s')) :$$

the gain of π' over π is the advantage of π' 's action under π 's prices, summed over the states π' visits. Under the hypothesis every term is ≥ 0 : the theorem a third time. Machine, round 0 from G: π_1 spends $x = 2.31$ discounted weeks at W and none at B, so $V^{\pi_1}(G) - V^{\pi_0}(G) = 2.31 \cdot 4.75 = 10.96 = 38.46 - 27.5$.

The same proof, unrolled (Sutton and Barto, 2018, Section 4.2); reading.

$$\begin{aligned} V^{\pi}(s) &\leq Q^{\pi}(s, \pi'(s)) = \mathbb{E}_{\pi'}[R_1 + \gamma V^{\pi}(S_1) \mid S_0 = s] \leq \mathbb{E}_{\pi'}[R_1 + \gamma Q^{\pi}(S_1, \pi'(S_1)) \mid S_0 = s] \\ &= \mathbb{E}_{\pi'}[R_1 + \gamma R_2 + \gamma^2 V^{\pi}(S_2) \mid S_0 = s] \leq \dots \\ &\leq \mathbb{E}_{\pi'}\left[\sum_{t=1}^k \gamma^{t-1} R_t + \gamma^k V^{\pi}(S_k) \mid S_0 = s\right] \xrightarrow[k \rightarrow \infty]{} V^{\pi'}(s). \end{aligned}$$

Each $\leq$ is the hypothesis one week later, each = Lecture 3's equation for Q^{π} ; line k is row k of the table. Agarwal et al. (2022, Lemma 1.13) write the chain on Q ; Jiang (2024, Section 2.1) proves it again via the *performance difference lemma*, back with policy gradient methods.

Why It Stops, and Only at an Optimal Policy

Unchanged means optimal.

$$\pi_{k+1} = \pi_k \Rightarrow \pi_k(s) \in \operatorname{argmax}_a Q^{\pi_k}(s, a) \quad \forall s \quad \text{the stop, read off } Q^{\pi_k}$$

$$\Rightarrow V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} = \mathcal{T} V^{\pi_k} \quad \mathcal{T}^{\pi} V = \mathcal{T} V \text{ iff } \pi \text{ greedy w.r.t. } V \text{ (Lecture 4)}$$

$$\Rightarrow V^{\pi_k} = V^*, \quad \pi_k \text{ optimal} \quad V = \mathcal{T} V \text{ has one solution (Lecture 5)}$$

$$\text{Conversely, } \pi \text{ optimal} \Rightarrow \pi(s) \in \operatorname{argmax}_a Q^{\pi}(s, a) \quad \forall s \quad \text{Lecture 4's test}$$

$$\Rightarrow \text{the tie rule keeps every } \pi(s) : \text{ unchanged}$$

so the run stops at the first optimal policy

A crude termination bound: at most $|\mathcal{A}|^{|\mathcal{S}|}$ evaluations.

- (1) $V^{\pi_{k+1}} \geq V^{\pi_k}, \quad Q^{\pi_{k+1}} \geq Q^{\pi_k}$ *the theorem and its Q corollary: no entry falls*
 if $\pi_{k+1}(s) \neq \pi_k(s)$, then
 $V^{\pi_{k+1}}(s) = Q^{\pi_{k+1}}(s, \pi_{k+1}(s))$ $V^\pi(s) = Q^\pi(s, \pi(s))$
 $\geq Q^{\pi_k}(s, \pi_{k+1}(s))$ $Q^{\pi_{k+1}} \geq Q^{\pi_k}$
 $> Q^{\pi_k}(s, \pi_k(s)) = V^{\pi_k}(s)$ $\pi_{k+1}(s)$ a maximizer; $\pi_k(s)$ not (tie rule)
- (2) $V^{\pi_0} \leq V^{\pi_1} \leq \dots$, each step strict somewhere $\Rightarrow$ all distinct *from (1), before the stop*
 $\Rightarrow \pi_0, \pi_1, \dots$ distinct $\Rightarrow$ at most $|\mathcal{A}|^{|\mathcal{S}|}$ rounds *finitely many deterministic policies*
- (3) stop $\Rightarrow$ optimal *above*

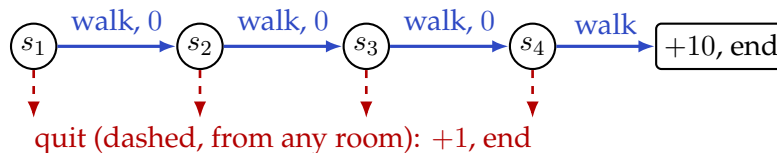
Machine: 8 policies; the longest run takes 4 evaluations, the run from “never repair” 3.

Example. Does the step leave “repair when worn” unchanged? “Repair when broken”?
 No: at W, run 26.15 > 24.62; strict there, so $V^{\pi_2}(W) > 24.62$ (it is 32.41). Yes: every maximum of Q^* is π^* ’s own action (Lecture 4).

- $|\mathcal{A}|^{|\mathcal{S}|}$ is Lecture 4’s brute-force count, 262,144 on the Gridworld. 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); improved by Scherrer (2016).
- In practice a handful: S&B’s 4×4 gridworld from the random policy, one round (Sutton and Barto, 2018, Figure 4.1); Jack’s car rental, 441 states, four (Sutton and Barto, 2018, Example 4.2).
- Value iteration certifies within $\gamma\theta/(1 - \gamma)$; policy iteration ends exactly. A round against an iteration: next time.

A Second Run: The Chain

Lecture 5's chain: four rooms; walk moves one room on for 0 and from s_4 collects 10 and ends; quit collects 1 and ends; $\gamma = 0.9$; $V^* = (7.29, 8.1, 9, 10)$, walk everywhere. Policy iteration from $\pi_0 = \text{quit everywhere}$.



The chain of Lecture 5. $A^\pi(s_i, \text{walk}) = 0.9V^\pi(s_{i+1}) - V^\pi(s_i)$ for $i < 4$, and $10 - V^\pi(s_4)$ at s_4 .

k	π_k at $s_1, \dots, s_4$	V^{π_k}	$A^{\pi_k}(\cdot, \text{walk})$	$\rightarrow \pi_{k+1}$	$\ V^* - V^{\pi_k}\ _\infty$
0	q q q q	1, 1, 1, 1	-0.1, -0.1, -0.1, 9	$s_4 \rightarrow \text{walk}$	9
1	q q q w	1, 1, 1, 10	-0.1, -0.1, 8, 0	$s_3 \rightarrow \text{walk}$	8
2	q q w w	1, 1, 9, 10	-0.1, 7.1, 0, 0	$s_2 \rightarrow \text{walk}$	7.1
3	q w w w	1, 8.1, 9, 10	6.29, 0, 0, 0	$s_1 \rightarrow \text{walk}$	6.29
4	w w w w	7.29, 8.1, 9, 10	0, 0, 0, 0	none: stop	0

Policy iteration on the chain from “quit everywhere”: five evaluations, four changes, one room per round, back to front. q: quit, w: walk.

Example. In round 0 the prize is already worth 8.1 from s_2 . Why does s_2 not switch until round 2?

The step sees the prize only through V^{π_k} : in round 0, $V^{\pi_0}(s_3) = 1$ (quit), so walking from s_2 is worth $0.9 < 1$. One room per round, for the reason W flipped on the machine: each round is greedy w.r.t. the current policy's prices.

- $V^{\pi_k} = V_k$ of Lecture 5 for $k = 1, \dots, 4$: here a round gains exactly one iteration; the errors fall by 0.89, 0.89, 0.89, then to 0.
- Never less than one iteration per round, on any model: next time. Machine: 0.26, then 0.
- $|\mathcal{A}|^{|S|} = 16$ policies; the run visits 5. The count that matters is the distance to the prize, not the number of policies.

	Value iteration (Lecture 5)	Policy iteration (today)
Iterates on	tables V_k	policies π_k
One step	one lookahead per pair, then a max	a solve, then one lookahead per pair and a max
Guarantee per step	$\ V_{k+1} - V^*\ _\infty \leq \gamma \ V_k - V^*\ _\infty$	$V^{\pi_{k+1}} \geq V^{\pi_k}$, strict where the policy changed
Stops	at θ : within $\gamma\theta/(1-\gamma)$ of V^*	when unchanged: at π^* and V^* exactly
On the machine	78 iterations to 0.01; π^* from $k = 2$, unknowingly	3 evaluations; π^* at round 2, certified

Summary

Lecture 6 on one board

- Policy iteration: any π_0 ; $V^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r^{\pi_k}$; $\pi_{k+1}(s) = \operatorname{argmax}_a Q^{\pi_k}(s, a)$, keep $\pi_k(s)$ at ties; stop when $\pi_{k+1} = \pi_k$.
- The run (table): never repair $\rightarrow$ repair when worn $\rightarrow$ repair when broken $\rightarrow$ stop. Three evaluations, two changes, W twice; errors 29.87, 7.79, 0.
- Theorem: $\mathcal{T}^{\pi'} V^\pi \geq V^\pi \Rightarrow V^{\pi'} \geq V^\pi$, strict where strict. Proof: $P^{\pi'} \geq 0$, then $V^\pi \leq \mathcal{T}^{\pi'} V^\pi \leq (\mathcal{T}^{\pi'})^2 V^\pi \leq \dots \rightarrow V^{\pi'}$. The greedy step has $\mathcal{T}^{\pi'} V^\pi = \mathcal{T} V^\pi \geq V^\pi$: every round improves.
- Unchanged $\Rightarrow V^{\pi_k} = \mathcal{T} V^{\pi_k} \Rightarrow V^{\pi_k} = V^*$. Tie rule $\Rightarrow$ every change raises a value $\Rightarrow$ no policy returns $\Rightarrow$ at most $|\mathcal{A}|^{|\mathcal{S}|}$ rounds.
- Advantages $A^\pi(s, a) = Q^\pi(s, a) - V^\pi(s)$: the step changes the action wherever an entry is positive; the stop is a table with none. Also $Q^{\pi_{k+1}} \geq Q^{\pi_k}$.
- The chain from quit everywhere: one room per round, $V^{\pi_k} = V_k$, five evaluations of 16 policies.
- Next: a round against an iteration; m evaluation iterations per improvement, $m = 1$ value iteration, $m = \infty$ policy iteration; what a policy off approximate values is worth.

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.
- Ronald A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.
- 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.
- Bruno Scherrer. Improved and generalized upper bounds on the complexity of policy iteration. *Mathematics of Operations Research*, 41(3):758–774, 2016.
- 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>.
- 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.