CSE 542: Statistical Reinforcement Learning

Lecture 1: Course Logistics & MDP Fundamentals

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

What is this course about?

Reinforcement learning (RL): how should an agent act in an uncertain, sequential environment to maximize cumulative reward?

This course develops the mathematical and algorithmic foundations of RL:

- Provably efficient methods and their theoretical guarantees
- From classical tabular settings to modern function approximation
- Both offline (batch) and online (interactive) learning

By the end, you will be positioned to **read and contribute to RL research**.

Topics

Foundations

- MDP theory: value functions, Bellman equations
- Policy iteration, value iteration
- Planning complexity

Online RL

- Model-based/free exploration, UCB for tabular and linear MDPs with sample complexity guarantees
- Policy gradient

Offline RL

- Fitted value iteration
- Offline policy evaluation & optimization

Function Approximation

- Linear MDPs, LSVI-UCB
- Bellman rank, Eluder dimension
- Provably efficient algorithms

Prerequisites

Required:

- Linear algebra, probability, calculus (see HW0 self-test)
- Concentration inequalities: Hoeffding, Bernstein, Azuma–Hoeffding

Strongly recommended:

- Online learning and multi-armed bandits (CSE 541 or equivalent)
- Consult [Szepesvári–Lattimore] or CSE 541 (Winter 2026) materials

Self-test: Complete HW0 (not graded) in the first week to gauge your readiness.

Course Materials

Primary Textbook:

- *Reinforcement Learning: Theory and Algorithms*, Agarwal, Brantley, Jiang, Kakade, Sun
  `rltheorybook.github.io/rltheorybook_ABJKS.pdf`

Recommended Background:

- *Bandit Algorithms*, Szepesvári & Lattimore
  `https://tor-lattimore.com/downloads/book/book.pdf`

Additional References (Practical RL):

- *Reinforcement Learning: An Introduction*, Sutton & Barto
- *Reinforcement Learning: An Overview*, Murphy (2024)

Logistics at a Glance

Meeting Times

Mon/Wed 10:00–11:20 AM
ECE 045

Office Hours

Instructor: Tue 2:30–3:30, CSE 340
Mars Gao: TBD
Kevin Huang: TBD

Communication

Discussion board: **Ed** (preferred)
Email:
`cse542-staff@cs.washington.edu`

Grading

Homework 1	20%
Homework 2	20%
Homework 3	20%
Final Project	40%

Submission

Single PDF to **Gradescope**
Typeset (LaTeX recommended)
No photos/scans

Final Project

- Choose a topic in reinforcement learning
- Write a **summary and literature review**: a primer for someone about to enter research in that area
- Worth **40%** of your grade
- Details forthcoming

Late Policy (Homeworks)

24-hour grace period, no questions asked.
Multiple days: email instructor *before* the due date.
Late policy does *not* apply to the final project.

Collaboration & LLM Policy

Homeworks are individual

You may collaborate to discuss ideas, but **write your own solutions**.
List all collaborators on your submission.

LLM Use Policy

LLMs (ChatGPT, etc.) are allowed as a *learning aid*.
If you use an LLM, you must attach a link to the full transcript.

If you find yourself copying substantial derivations from the LLM $\Rightarrow$ you've crossed the line.

No transcript attached + *LLM suspected* $\Rightarrow$ report filed.

Academic Integrity

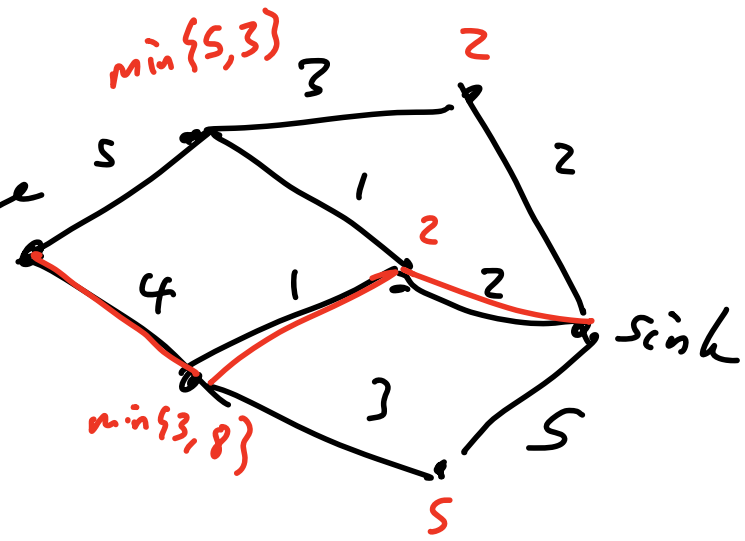
Do not use pre-existing solutions from the web, past courses, or other textbooks.
Violations are reported to Community Standards and Student Conduct.

Today's Plan

Motivation: Sequential Decision Making

Many problems involve **sequential decisions under uncertainty**:

- Robot navigation: reach a goal efficiently
- Conversational agents: complete user tasks
- Games: chess, Go, poker
- Resource allocation, clinical trials, ...



Key challenges:

- Actions affect future states (not i.i.d.!)
- Reward may be delayed
- Need to balance exploration vs. exploitation

The (Infinite-Horizon Discounted) MDP

A **Markov Decision Process** is a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$:

$\Delta(S) = \{p \in \mathbb{R}_+^{|S|} : \sum_i p_i = 1\}$ $\lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r(s_t, a_t)\right]$

- $\mathcal{S}$: state space (assume finite)
- $\mathcal{A}$: action space (assume finite)
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$: transition kernel
  $P(s' \mid s, a)$: prob. of moving to s'
- $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$: reward function
- $\gamma \in [0, 1)$: discount factor
- $\mu \in \Delta(\mathcal{S})$: initial state distribution

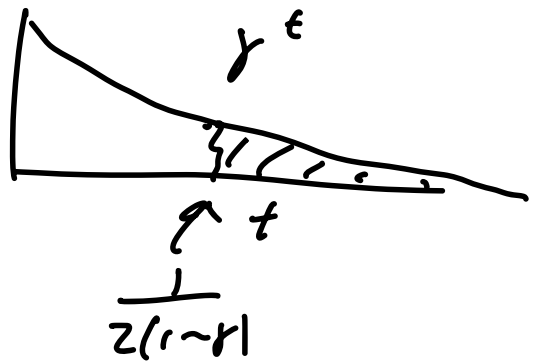
Goal: Maximize $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right]$

$\leq \sum_{t=0}^{\infty} \gamma^t = \frac{1}{1-\gamma}$

Interaction Protocol

$s_0 \sim \mu$. At each step $t = 0, 1, 2, \ldots$:

1. Observe s_t
2. Select $a_t \in \mathcal{A}$
3. Receive $r_t = r(s_t, a_t)$
4. Transition $s_{t+1} \sim P(\cdot \mid s_t, a_t)$



Examples

Navigation

State = current location. Actions = {N, S, E, W}. Deterministic transitions.
Reward = 1 at goal, else 0. Discount γ incentivizes shortest paths.
Optimal policy: greedy shortest path. Value $\approx \gamma^d$ for path length d.

Inventory Management

State = current stock level. Action = how much inventory to order each day.
Demand is stochastic: reward = sales revenue (only if item is in stock) − ordering cost.
Tension: understocking loses sales; overstocking ties up capital.

Strategic Games

State = board position. Actions = legal moves.
Reward = eventual win/loss.
RL has achieved superhuman performance in Go, Chess, Poker, Backgammon.

Planning vs. Statistical RL

Planning (dynamics known)

The model P, r is given. The challenge is purely *computational*: how do we efficiently find the optimal policy?

Examples:

- Shortest path / navigation
- Board games (Chess, Go): simulator provides exact next state
- Video games: emulator is the model

Statistical RL (dynamics unknown)

We don't know P or r. We must *learn* what actions lead to what outcomes by interacting with the environment.

Examples:

- Personalized medical treatment: how does this patient's condition evolve if we prescribe drug A vs. drug B?
- Inventory management: what is the true demand distribution?

Planning vs. Statistical RL

Planning (dynamics known)

The model P, r is given. The challenge is purely *computational*: how do we efficiently find the optimal policy?

Examples:

- Shortest path / navigation
- Board games (Chess, Go): simulator provides exact next state
- Video games: emulator is the model

Statistical RL (dynamics unknown)

We don't know P or r. We must *learn* what actions lead to what outcomes by interacting with the environment.

Examples:

- Personalized medical treatment: how does this patient's condition evolve if we prescribe drug A vs. drug B?
- Inventory management: what is the true demand distribution?

Focus of this course

We study the **statistical** problem: how many interactions does it take to learn a near-optimal policy? Planning serves as a subroutine, but the core challenge is exploration and generalization.

Policies and Value Functions

History: $\tau_t = (s_0, a_0, r_0, s_1, \ldots, s_t)$.
A **policy** π maps histories to distributions over actions.

Stationary Policy

$\pi : \mathcal{S} \to \Delta(\mathcal{A})$ — actions depend only on current state s_t.
A **deterministic** stationary policy is $\pi : \mathcal{S} \to \mathcal{A}$.

$V^\pi \in \mathbb{R}^{|S|}$ $\quad r(s,a) \in [0,1]$

Value function $V^\pi : \mathcal{S} \to \mathbb{R}$:

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \;\middle|\; \pi,\ s_0 = s\right], \qquad 0 \le V^\pi(s) \le \frac{1}{1-\gamma}.$$

Action-value (Q-value) function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \;\middle|\; \pi,\ s_0 = s,\ a_0 = a\right].$$

Goal: find π maximizing $V^\pi(s)$ for all s (or for $s \sim \mu$).

Bellman Consistency Equations

Lemma 1.4 (Bellman consistency for π)

For any stationary policy π, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$: If deterministic $V^\pi(s) = Q^\pi(s, \pi(s))$

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)]$$
$$Q^\pi(s,a) = r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$$

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 = s, a_0 = a\right]$$

$$= r(s,a) + \mathbb{E}_{s_1 \sim P(s,a)}\left[\mathbb{E}_\pi\left[\sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_1\right]\right]$$

$$= r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(s,a)}\, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) \mid \pi, s_1\right]$$

$$= r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(s,a)}\left[Q^\pi(s', \pi(s'))\right]$$

$$= r(s,a) + \gamma \, \mathbb{E}_{s'}[V^\pi(s')]$$

Bellman Consistency Equations

Lemma 1.4 (Bellman consistency for π)

For any stationary policy π, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s,a)]$$
$$Q^\pi(s,a) = r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$$

Proof. Unroll one step using the Markov property and tower property:

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s, a_0 = a\right] \\ &= r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_{t+1}, a_{t+1}) \,\middle|\, \pi, s_1 = s'\right]\right] \\ &= r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]\,. \end{aligned}$$

The first equation in the lemma then follows by averaging over $a \sim \pi(\cdot|s)$. □

Matrix Form & Corollary 1.5

In matrix/vector form (treating Q^π, r as vectors over $\mathcal{S} \times \mathcal{A}$):

$$Q^\pi = r + \gamma P V^\pi = r + \gamma P^\pi Q^\pi,$$

where $P^\pi_{(s,a),(s',a')} = P(s'|s,a)\,\pi(a'|s')$ is the induced state-action transition matrix.

Corollary 1.5. (Closed-form solution)

$$Q^\pi = (I - \gamma P^\pi)^{-1} r.$$

A is invertible if $Ax \neq 0$ if $x \neq 0$

$$\|(I - \gamma P^\pi)x\|_\infty \geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty$$

$$\leq \|x\|_\infty$$

$$\geq \|x\|_\infty - \gamma \|x\|_\infty$$

$$= (1-\gamma)\|x\|_\infty > 0$$

Matrix Form & Corollary 1.5

In matrix/vector form (treating Q^π, r as vectors over $\mathcal{S} \times \mathcal{A}$):

$$Q^\pi = r + \gamma P V^\pi = r + \gamma P^\pi Q^\pi,$$

$\|x\|_\infty = \max_{i=1,\dots,d} |x_i|$

where $P^\pi_{(s,a),(s',a')} = P(s'|s,a)\,\pi(a'|s')$ is the induced state-action transition matrix.

Corollary 1.5. (Closed-form solution)

$Q^\pi = (I - \gamma P^\pi)^{-1} r.$

Proof. It suffices to show $I - \gamma P^\pi$ is invertible. For any nonzero $x \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$\begin{aligned}
\|(I - \gamma P^\pi)x\|_\infty &\geq \|x\|_\infty - \gamma \|P^\pi x\|_\infty && \text{(triangle inequality)} \\
&\geq \|x\|_\infty - \gamma \|x\|_\infty && (P^\pi \text{ row-stochastic}) \\
&= (1-\gamma)\|x\|_\infty > 0. && (\gamma < 1,\ x \neq 0)
\end{aligned}$$

So $I - \gamma P^\pi$ has trivial kernel, hence is invertible. Then $Q^\pi = r + \gamma P^\pi Q^\pi$ gives $(I - \gamma P^\pi)Q^\pi = r$. □

Optimal Policy: The Key Theorem

Theorem 1.7 (Existence of optimal stationary policy)

Define $V^*(s) \triangleq \sup_\pi V^\pi(s)$ and $Q^*(s,a) \triangleq \sup_\pi Q^\pi(s,a)$.

There exists a stationary deterministic policy π^* such that for all s, a:

$$V^{\pi^*}(s) = V^*(s), \qquad Q^{\pi^*}(s,a) = Q^*(s,a).$$

$V^{\pi_*}(s) = \max_a Q^{\pi_*}(s,a)$

Why is this surprising? The supremum is over *all* policies (non-stationary, randomized, history-dependent). Yet a simple "look only at current state" policy suffices.

$$V^*(s) = \sup_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s \right]$$

$$= \sup_\pi \mathbb{E}_\pi \left[r(s_0, a_0) + \sum_{t=1}^{\infty} \gamma^t r(s_t, a_t) \mid s_1 = s \right]$$

Theorem 1.7: Proof sketch

Step 1. $\sup_\pi \mathbb{E}[\sum_{t=1}^{\infty} \gamma^t r(S_t, A_t) \mid \pi,\, S_1 = s'] = \gamma V^*(s')$. *(shift time index)*

Step 2. $\tilde{\pi}(s) \in \arg\max_a \mathbb{E}_{\tilde{\pi}}[r(s,a) + \gamma V^*(S_1) \mid s]$.

Step 3. Trivially we have $V^{\tilde{\pi}}(s_0) \leq V^*(s_0)$. Now show $V^*(s_0) \leq V^{\tilde{\pi}}(s_0)$:

$$
\begin{aligned}
V^*(s_0) &= \sup_\pi \mathbb{E}_\pi \left[r(S_0, A_0) + \sum_{t=1}^{\infty} \gamma^t r(S_t, A_t) \mid s_0 \right] \\
&= \sup_\pi \mathbb{E}_\pi \left[r(S_0, A_0) + \mathbb{E}_\pi \left[\sum_{t=1}^{\infty} \gamma^t r(S_t, A_t) \mid \pi, s_0, A_0, R_0, S_1 \right] \,\middle|\, s_0 \right] \\
&\leq \sup_\pi \mathbb{E}_\pi \left[r(S_0, A_0) + \sup_{\pi'} \mathbb{E}_{\pi'} \left[\sum_{t=1}^{\infty} \gamma^t r(S_t, A_t) \mid \pi', s_0, A_0, R_0, S_1 \right] \,\middle|\, s_0 \right] \\
&= \sup_\pi \mathbb{E}_\pi \left[r(S_0, A_0) + \gamma V^*(S_1) \,\middle|\, s_0 \right] && \text{(Step 1)} \\
&= \mathbb{E}_{\tilde{\pi}} \left[r(S_0, A_0) + \gamma V^*(S_1) \,\middle|\, s_0 \right] && \text{(def. } \tilde{\pi}\text{)} \\
&\leq \mathbb{E}_{\tilde{\pi}} \left[r(S_0, A_0) + \gamma r(S_1, A_1) + \gamma^2 V^*(S_2) \mid s_0 \right] && \text{(recurse)} \\
&\leq V^{\tilde{\pi}}(s_0)
\end{aligned}
$$

Bellman Optimality Equations

Define the **Bellman optimality operator** $T_M : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$(TQ)(s,a) \triangleq r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s',a') \right].$$

Write $V_Q(s) \triangleq \max_a Q(s,a)$ and $\pi_Q(s) \in \arg\max_a Q(s,a)$ (greedy policy).

Theorem 1.8 (Bellman optimality equations)

$Q = Q^*$ **if and only if** $Q = TQ$.

Moreover, π_{Q^*} is an optimal policy.

Bellman Optimality Equations

Define the **Bellman optimality operator** $T_M : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$(TQ)(s,a) \triangleq r(s,a) + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,a)} \left[\max_{a'} Q(s',a') \right].$$

Write $V_Q(s) \triangleq \max_a Q(s,a)$ and $\pi_Q(s) \in \arg\max_a Q(s,a)$ (greedy policy).

Theorem 1.8 (Bellman optimality equations)

$Q = Q^*$ **if and only if** $Q = TQ$.

Moreover, π_{Q^*} is an optimal policy.

This almost gives us an algorithm for identifying π^*: just find some Q that satisfies the fixed point equation $Q = TQ$!

Q-Value Iteration

Algorithm: Initialize $Q^{(0)} = 0$. Repeat:

$$Q^{(k+1)} \leftarrow TQ^{(k)}, \quad \text{i.e.,} \quad Q^{(k+1)}(s,a) = r(s,a) + \gamma \, \mathbb{E}_{s'} \left[\max_{a'} Q^{(k)}(s',a') \right].$$

Extract policy: $\pi^{(k)} = \pi_{Q^{(k)}}$ (greedy w.r.t. $Q^{(k)}$).

Q-Value Iteration

Algorithm: Initialize $Q^{(0)} = 0$. Repeat:

$$\underbrace{Q^{(k+1)} \leftarrow TQ^{(k)}}_{Q^* = TQ^*}, \quad \text{i.e.,} \quad Q^{(k+1)}(s,a) = r(s,a) + \gamma \, \mathbb{E}_{s'}\left[\max_{a'} Q^{(k)}(s',a')\right].$$

Extract policy: $\pi^{(k)} = \pi_{Q^{(k)}}$ (greedy w.r.t. $Q^{(k)}$).

Why does this converge? The operator T is a γ-**contraction** in $\|\cdot\|_\infty$, and Q^* is its unique fixed point.

Lemma 1.10 (Contraction)

For any $Q, Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$\|TQ - TQ'\|_\infty \leq \gamma \, \|Q - Q'\|_\infty .$$

Q-Value Iteration

Algorithm: Initialize $Q^{(0)} = 0$. Repeat:

$$Q^{(k+1)} \leftarrow TQ^{(k)}, \quad \text{i.e.,} \quad Q^{(k+1)}(s,a) = r(s,a) + \gamma \, \mathbb{E}_{s'} \left[\max_{a'} Q^{(k)}(s',a') \right].$$

Extract policy: $\pi^{(k)} = \pi_{Q^{(k)}}$ (greedy w.r.t. $Q^{(k)}$).

Why does this converge? The operator T is a γ-**contraction** in $\|\cdot\|_\infty$, and Q^* is its unique fixed point.

Lemma 1.10 (Contraction)

For any $Q, Q' \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$\|TQ - TQ'\|_\infty \leq \gamma \, \|Q - Q'\|_\infty .$$

Recall $TQ^* = Q^*$ and $Q_0 = 0$. Thus,

$$\|Q_k - Q^*\|_\infty = \|TQ_{k-1} - TQ^*\|_\infty \leq \gamma \|Q_{k-1} - Q^*\|_\infty \leq \gamma^k \|Q^*\|_\infty \leq \frac{\gamma^k}{1-\gamma}.$$

Lemma 1.10: Contraction — Proof

Step 1. Fix any s, WLOG $V_Q(s) \geq V_{Q'}(s)$. Let $a^* \in \arg\max_a Q(s,a)$.

$V_Q(s) - V_{Q'}(s) = Q(s,a^*) - \max_a Q'(s,a) \leq Q(s,a^*) - Q'(s,a^*) \leq \|Q - Q'\|_\infty$.

Thus, $\|V_Q - V_{Q'}\|_\infty \leq \|Q - Q'\|_\infty$.

Step 2.

$$\|TQ - TQ'\|_\infty = \gamma \|P(V_Q - V_{Q'})\|_\infty \leq \gamma \|V_Q - V_{Q'}\|_\infty \leq \gamma \|Q - Q'\|_\infty ,$$

where $\|Px\|_\infty \leq \|x\|_\infty$ because each row of P is a probability distribution. □

Q-Error Amplification (Lemma 1.11)

Lemma 1.11 (Suboptimality of greedy policy)

For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$V^{\pi_Q} \geq V^* - \frac{2\left\|Q - Q^*\right\|_\infty}{1-\gamma} \cdot \mathbf{1}.$$

We showed above that $\left\|Q_k - Q^*\right\|_\infty \leq \frac{\gamma^k}{1-\gamma}$. Thus, $V^{\pi_{Q_k}} \geq V^* - \frac{2\gamma^k}{(1-\gamma)^2} \cdot \mathbf{1}$.

Q-Error Amplification (Lemma 1.11)

Lemma 1.11 (Suboptimality of greedy policy)

For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

$$V^{\pi_Q} \geq V^* - \frac{2\left\|Q - Q^*\right\|_\infty}{1-\gamma} \cdot \mathbf{1}.$$

We showed above that $\left\|Q_k - Q^*\right\|_\infty \leq \frac{\gamma^k}{1-\gamma}$. Thus, $V^{\pi_{Q_k}} \geq V^* - \frac{2\gamma^k}{(1-\gamma)^2} \cdot \mathbf{1}$.

Theorem 1.12 (Q-value iteration convergence)

Set $Q^{(0)} = 0$. After $k \geq \frac{1}{1-\gamma} \log \frac{2}{(1-\gamma)^2 \varepsilon}$ iterations,

$$V^{\pi^{(k)}} \geq V^* - \varepsilon \cdot \mathbf{1}.$$

Lemma 1.11: Proof

Fix s; let $a = \pi_Q(s)$ and π^* be optimal. Then:

$$V^*(s) - V^{\pi_Q}(s) = \underbrace{Q^*(s, \pi^*(s)) - Q^*(s, a)}_{\text{(I)}} + \underbrace{Q^*(s, a) - Q^{\pi_Q}(s, a)}_{\text{(II)}}.$$

Bounding (I): Since π_Q is greedy w.r.t. Q, we have $Q(s, \pi^*(s)) \leq Q(s, a)$, so:

$$\text{(I)} = Q^*(s, \pi^*(s)) - Q^*(s, a) \leq [Q^*(s, \pi^*) - Q(s, \pi^*)] + [Q(s, a) - Q^*(s, a)] \leq 2 \left\| Q - Q^* \right\|_\infty.$$

Bounding (II): By Bellman consistency for π_Q:

$$\text{(II)} = \gamma \, \mathbb{E}_{s'} \left[V^*(s') - V^{\pi_Q}(s') \right] \leq \gamma \left\| V^* - V^{\pi_Q} \right\|_\infty.$$

Combine: $\left\| V^* - V^{\pi_Q} \right\|_\infty \leq 2 \left\| Q - Q^* \right\|_\infty + \gamma \left\| V^* - V^{\pi_Q} \right\|_\infty$.
Rearranging:

$$\left\| V^* - V^{\pi_Q} \right\|_\infty \leq \frac{2}{1 - \gamma} \left\| Q - Q^* \right\|_\infty. \qquad \square$$

Policy Iteration

Start from any deterministic stationary policy π_0. Repeat for $k = 0, 1, 2, \ldots$:

Step 1: Policy Evaluation

Solve for Q^{π_k}: $\quad Q^{\pi_k} = (I - \gamma P^{\pi_k})^{-1} r$.

Step 2: Policy Improvement

$\pi_{k+1} = \pi_{Q^{\pi_k}}$ $\quad$ (greedy w.r.t. Q^{π_k}).

Key properties:

- Values improve monotonically: $Q^{\pi_{k+1}} \geq Q^{\pi_k}$ (elementwise).
- Converges geometrically to Q^*.
- Can terminate with *exact* π^* in finitely many steps.

Policy Iteration: Monotonicity & Contraction

Lemma 1.13

1. **(Sandwich)** $Q^{\pi_{k+1}} \geq TQ^{\pi_k} \geq Q^{\pi_k}$.
2. **(Geometric contraction)** $\left\|Q^* - Q^{\pi_{k+1}}\right\|_\infty \leq \gamma \left\|Q^* - Q^{\pi_k}\right\|_\infty$.

Policy Iteration: Monotonicity & Contraction

Lemma 1.13

1. **(Sandwich)** $Q^{\pi_{k+1}} \geq TQ^{\pi_k} \geq Q^{\pi_k}$.
2. **(Geometric contraction)** $\left\|Q^* - Q^{\pi_{k+1}}\right\|_\infty \leq \gamma \left\|Q^* - Q^{\pi_k}\right\|_\infty$.

(1a) $TQ^{\pi_k} \geq Q^{\pi_k}$**:** Since $V_{Q^{\pi_k}}(s) = \max_a Q^{\pi_k}(s,a) \geq Q^{\pi_k}(s, \pi_k(s)) = V^{\pi_k}(s)$:

$$TQ^{\pi_k}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V_{Q^{\pi_k}}(s')] \geq r(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi_k}(s')] = Q^{\pi_k}(s,a).$$

(1b) $Q^{\pi_{k+1}} \geq TQ^{\pi_k}$**:** Let $\pi' = \pi_{k+1}$. The policy evaluation operator $T^{\pi'}V(s) = r(s, \pi'(s)) + \gamma \mathbb{E}_{s'}[V(s')]$ satisfies:

$$T^{\pi'}V^{\pi_k}(s) = Q^{\pi_k}(s, \pi'(s)) = V_{Q^{\pi_k}}(s) \geq V^{\pi_k}(s).$$

Iterating monotonicity: $V^{\pi_{k+1}} = \lim_{n\to\infty}(T^{\pi'})^n V^{\pi_k} \geq V_{Q^{\pi_k}}$. Therefore:

$$Q^{\pi_{k+1}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi_{k+1}}(s')] \geq r(s,a) + \gamma \mathbb{E}_{s'}[V_{Q^{\pi_k}}(s')] = TQ^{\pi_k}(s,a).$$

Policy Iteration: Monotonicity & Contraction

Lemma 1.13

1. **(Sandwich)** $Q^{\pi_{k+1}} \geq TQ^{\pi_k} \geq Q^{\pi_k}$.
2. **(Geometric contraction)** $\left\|Q^* - Q^{\pi_{k+1}}\right\|_\infty \leq \gamma \left\|Q^* - Q^{\pi_k}\right\|_\infty$.

(1a) $TQ^{\pi_k} \geq Q^{\pi_k}$**:** Since $V_{Q^{\pi_k}}(s) = \max_a Q^{\pi_k}(s,a) \geq Q^{\pi_k}(s, \pi_k(s)) = V^{\pi_k}(s)$:

$$TQ^{\pi_k}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V_{Q^{\pi_k}}(s')] \geq r(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi_k}(s')] = Q^{\pi_k}(s,a).$$

(1b) $Q^{\pi_{k+1}} \geq TQ^{\pi_k}$**:** Let $\pi' = \pi_{k+1}$. The policy evaluation operator $T^{\pi'}V(s) = r(s,\pi'(s)) + \gamma \mathbb{E}_{s'}[V(s')]$ satisfies:

$$T^{\pi'}V^{\pi_k}(s) = Q^{\pi_k}(s,\pi'(s)) = V_{Q^{\pi_k}}(s) \geq V^{\pi_k}(s).$$

Iterating monotonicity: $V^{\pi_{k+1}} = \lim_{n\to\infty}(T^{\pi'})^n V^{\pi_k} \geq V_{Q^{\pi_k}}$. Therefore:

$$Q^{\pi_{k+1}}(s,a) = r(s,a) + \gamma \mathbb{E}_{s'}[V^{\pi_{k+1}}(s')] \geq r(s,a) + \gamma \mathbb{E}_{s'}[V_{Q^{\pi_k}}(s')] = TQ^{\pi_k}(s,a).$$

(2) Contraction: Since $Q^* \geq Q^{\pi_{k+1}} \geq TQ^{\pi_k}$:

$$0 \leq Q^* - Q^{\pi_{k+1}} \leq Q^* - TQ^{\pi_k} = TQ^* - TQ^{\pi_k}.$$

Taking $\|\cdot\|_\infty$ and applying Lemma 1.10: $\left\|Q^* - Q^{\pi_{k+1}}\right\|_\infty \leq \gamma \left\|Q^* - Q^{\pi_k}\right\|_\infty$. □

Policy Iteration: Convergence

Theorem 1.14 (Policy iteration convergence)

Starting from any π_0:

$$\|Q^* - Q^{\pi_k}\|_\infty \leq \gamma^k \|Q^* - Q^{\pi_0}\|_\infty \leq \frac{\gamma^k}{1-\gamma}.$$

For $k \geq \frac{\log(1/((1-\gamma)\varepsilon))}{1-\gamma}$: $\quad Q^{\pi_k} \geq Q^* - \varepsilon \cdot \mathbf{1}$.

Policy iteration vs. value iteration:

- Same geometric rate γ^k, but PI terminates with *exact* π^* (finitely many policies).
- Naive bound: $|\mathcal{A}|^{|\mathcal{S}|}$ iterations.
- Each PI iteration costs $O(|\mathcal{S}|^3 + |\mathcal{S}|^2|\mathcal{A}|)$ (policy evaluation + greedy step).

Finite-Horizon MDPs

A **finite-horizon MDP** $M = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1}, H, \mu)$:

- Horizon $H \in \mathbb{N}$; rewards $r_h(s,a) \in [0,1]$; transitions P_h may depend on h.

Value functions: For $h \in \{0, \ldots, H\}$, $s \in \mathcal{S}$:

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} r_t(S_t, A_t) \,\middle|\, \pi, S_h = s\right], \qquad V_H^\pi(s) \equiv 0.$$

Finite-Horizon MDPs

A **finite-horizon MDP** $M = (\mathcal{S}, \mathcal{A}, \{P_h\}_{h=0}^{H-1}, \{r_h\}_{h=0}^{H-1}, H, \mu)$:

- Horizon $H \in \mathbb{N}$; rewards $r_h(s,a) \in [0,1]$; transitions P_h may depend on h.

Value functions: For $h \in \{0, \dots, H\}$, $s \in \mathcal{S}$:

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{t=h}^{H-1} r_t(S_t, A_t) \,\middle|\, \pi, S_h = s\right], \qquad V_H^\pi(s) \equiv 0.$$

Stage-h Bellman optimality operator:

$$(T_h f)(s,a) = r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\max_{a'} f(s', a')\right].$$

Theorem 1.9 (Bellman optimality: finite horizon)

The unique solution to $Q_H^* \equiv 0$, $Q_h^* = T_h Q_{h+1}^*$ equals $\sup_\pi Q_h^\pi$.
The policy $\pi^*(s,h) \in \arg\max_a Q_h^*(s,a)$ is optimal.

Proof by backward induction. (Exercise.)

Discounted vs. Finite Horizon

Infinite-Horizon Discounted

- Stationary dynamics/rewards
- Tabular size: $O(|\mathcal{S}||\mathcal{A}|)$
- Convergence via contraction ($\gamma < 1$)
- Effective horizon: $\frac{1}{1-\gamma}$

Finite Horizon (Time-Dependent)

- P_h, r_h may vary with h
- Tabular size: $O(H|\mathcal{S}||\mathcal{A}|)$
- Backward induction (exact, no contraction needed)
- Sharper theoretical rates

Convention in this course

Both settings studied. The finite-horizon time-dependent formulation is often **more convenient theoretically** (backward induction, sharper rates).

Can fold time h into the state to recover a stationary MDP on $\mathcal{S} \times [H]$.

Backward Induction (Section 1.3.3)

In the finite-horizon setting, backward induction computes Q_h^* **exactly** in $O(H|\mathcal{S}|^2|\mathcal{A}|)$ time.

Algorithm

Initialize $Q_H(s,a) \equiv 0$. For $h = H-1, H-2, \ldots, 0$:

$$Q_h(s,a) \leftarrow r_h(s,a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}\left[\max_{a'} Q_{h+1}(s',a')\right].$$

By Theorem 1.9, the resulting $Q_h = Q_h^*$. The optimal policy is $\pi^*(s,h) = \arg\max_a Q_h^*(s,a)$.

No contraction needed: backward induction is exact. This is a major advantage of the finite-horizon formulation.

Complexity: $O(H|\mathcal{S}|^2|\mathcal{A}|)$ with naive matrix-vector products, $O(H|\mathcal{S}||\mathcal{A}|)$ with sparse transitions.

Summary

MDP Basics:

- MDP tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$
- Policies, value functions V^π, Q^π
- Bellman consistency (Lemma 1.4)
- Optimal policy exists and is stationary (Thm. 1.7)
- Bellman optimality equations (Thm. 1.8)
- Finite-horizon formulation (Thm. 1.9)

Planning:

- Value iteration: T is γ-contraction (Lem. 1.10), convergence (Thm. 1.12)
- Suboptimality of greedy policy (Lem. 1.11): $\frac{2}{1-\gamma} \|Q - Q^*\|_\infty$
- Policy iteration: monotone + geometric convergence (Lem. 1.13, Thm. 1.14)
- Backward induction: exact, $O(H|\mathcal{S}|^2|\mathcal{A}|)$

Next

Offline RL: given a fixed dataset of transitions, how do we evaluate or optimize a policy?