Lecture 1: RL Overview, MDPs

Supervised Learning

Before diving into reinforcement learning, it is helpful to recall supervised learning (SL) as a point of comparison. In SL, a fixed dataset is collected independently of the learner—someone else gathered the data, and every learner trains on the same examples. This "offline" and "passive" nature of SL stands in sharp contrast to RL, where the agent actively generates its own data through interaction with the environment.

Given $\{(x^{(i)}, y^{(i)})\}$, learn $f : x \mapsto y$

- Online version: for round $t = 1, 2, \ldots$, the learner
  - observes $x^{(t)}$
  - predicts $\hat{y}^{(t)}$
  - receives $y^{(t)}$ (the true label is revealed)
- Want to maximize # of correct predictions
- e.g., classifies if an image is about a dog, a cat, a plane, etc. (multi-class classification)
- Dataset is fixed for everyone: the learner has no control over which examples appear
- "Full information setting": after making a prediction, the learner observes the true label $y^{(t)}$, regardless of its prediction $\hat{y}^{(t)}$
- Core challenge: generalization—performing well on unseen data from the same distribution

Contextual bandits

Contextual bandits introduce two key challenges absent from supervised learning: *partial information* and *exploration*. Unlike SL where the true label is always revealed, here the learner only observes the reward for the action it actually chose—not the rewards it *would have* received for other actions. This creates a fundamental tension: should the learner *exploit* actions that have worked well so far, or *explore* new actions that might be even better?

For round $t = 1, 2, \ldots$, **the learner**

- Given context $x^{(t)}$, chooses from a set of actions $a^{(t)} \in A$

- Receives reward $r^{(t)} \sim \mathcal{R}(x^{(t)}, a^{(t)})$ (i.e., can be random)
- Want to maximize total reward $\sum_t r^{(t)}$
- You generate your own dataset $\{(x^{(t)}, a^{(t)}, r^{(t)})\}$! The data distribution depends on the learner's actions.
- e.g., for an image, the learner guesses a label, and is told whether correct or not (reward = 1 if correct and 0 otherwise). **Do not know what's the true label**—only whether the guess was right.
- e.g., for a user, the website recommends a movie, and observes whether the user likes it or not. **Do not know what movies the user really wants to see**—only feedback on the recommendation made.
- "Partial information setting": the learner never observes counterfactual rewards (what would have happened under different actions)
- Simplification: no context x, this reduces to Multi-Armed Bandits (MAB)

Reinforcement Learning

Reinforcement learning extends contextual bandits by adding *sequential structure*: the agent's actions affect not only immediate rewards but also future states and opportunities. This introduces *delayed consequences*—an action taken now may have effects that only become apparent much later. Examples of RL problems include:

- **Game playing:** In chess or Go, a single move may seem neutral but sets up a winning (or losing) position many moves later.
- **Robotics:** A robot learning to walk must coordinate many joint movements over time; falling down now prevents future progress.
- **Dialogue systems:** A conversational agent's early responses shape the user's subsequent questions and overall satisfaction.

For round $t = 1, 2, \ldots$,

- For time step $h = 1, 2, \ldots, H$, the learner
  - Observes state $x_h^{(t)}$
  - Chooses action $a_h^{(t)}$
  - Receives reward $r_h^{(t)} \sim \mathcal{R}(x_h^{(t)}, a_h^{(t)})$

  - Next state $x_{h+1}^{(t)}$ is generated as a function of $x_h^{(t)}$ and $a_h^{(t)}$ (or sometimes, all previous states and actions within round t)

  (The above steps repeat for each time step h, with the new state $x_{h+1}^{(t)}$ becoming the observed state in the next step.)

- RL = Bandits + "Delayed rewards/consequences": actions affect future states, not just immediate rewards
- The protocol here is for *episodic* RL (each round t is an *episode* of length H, we will discuss this setting more in later lectures).

	Generalization	**Interaction**	**Exploration**	**Credit Assignment**
Supervised Learning	✓			
Contextual Bandits	✓	✓	✓	
Reinforcement Learning	✓	✓	✓	✓

Table 1: Comparison of learning paradigms.

The table columns capture the following concepts:

- **Generalization:** Learn from observed data and perform well on unseen inputs.
- **Interaction:** The learner's actions influence what data it observes.
- **Exploration:** Must try different actions to discover which ones are best.
- **Credit Assignment:** Must determine which past actions were responsible for current rewards—the key challenge unique to RL.

Infinite-horizon discounted MDPs

A *Markov Decision Process* (MDP) is the standard mathematical framework for sequential decision-making under uncertainty. The "Markov" property means that the future depends only on the current state and action, not on the history of how we arrived at that state. This memoryless property makes MDPs tractable while still capturing a wide range of practical problems.

An MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, d_0)$

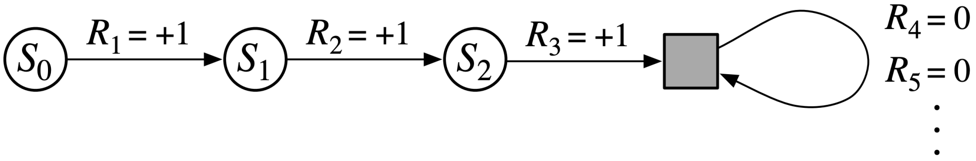
- State space $\mathcal{S}$.

- Action space $\mathcal{A}$.

> We will only consider discrete and finite (but can be exponentially large) spaces in this course.

- Transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. $\Delta(\mathcal{S})$ is the probability simplex over $\mathcal{S}$, i.e., all non-negative vectors of length $|\mathcal{S}|$ that sum to 1
- Reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$, where $r \sim \mathcal{R}(s, a)$. We write $R(s, a) := \mathbb{E}[\mathcal{R}(s, a)]$ for the expected reward.

  Assumption (Bounded Rewards). We assume rewards are non-negative and bounded: $R(s, a) \in [0, R_{\max}]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
- Discount factor $\gamma \in [0, 1)$
- Initial state distribution d_0
- The agent starts in some state s_0, takes action a_0, receives reward $r_0 \sim \mathcal{R}(s_0, a_0)$, transitions to $s_1 \sim \mathcal{P}(s_0, a_0)$, takes action a_1, so on so forth — the process continues indefinitely
- Sometimes we define a terminal / absorbing state s_∞: a special state that transitions only to itself and generates only rewards of zero

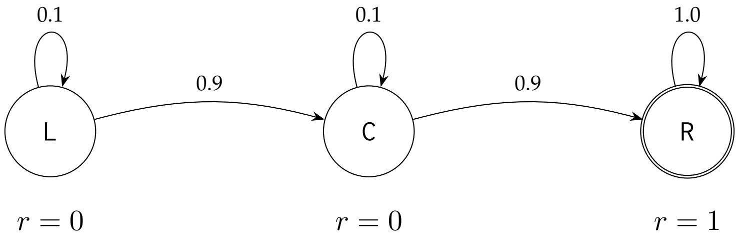


The rightmost shaded block represents the absorbing state s_∞. Figure adapted from Sutton & Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, 2018.

Example: Simple Navigation MDP. Consider a robot navigating a corridor with three locations: `Left`, `Center`, and `Right`. The goal is to reach `Right`.

- $\mathcal{S} = \{\texttt{Left}, \texttt{Center}, \texttt{Right}\}$
- $\mathcal{A} = \{\texttt{go-left}, \texttt{go-right}\}$
- Transitions: From `Center`, action `go-right` moves to `Right` with probability 0.9 and stays at `Center` with probability 0.1 (modeling possible slipping). Action `go-left` similarly moves to `Left` with probability 0.9. From `Left` or `Right`, actions toward the boundary have no effect.

- Rewards: $R(s,a) = 1$ if $s =$ `Right`, and $R(s,a) = 0$ otherwise.
- The state `Right` can be treated as absorbing (the robot stays there and collects reward 1 each step).



Transition diagram under the policy $\pi(s) =$ `go-right` for all s. Double circle indicates the absorbing state.

This simple example already illustrates the key MDP concepts: the agent must reason about how actions lead to state transitions and accumulate rewards over time.

Value and policy

The central objects in RL are *policies* and *value functions*. A policy specifies how the agent behaves: given the current state, which action should it take? A value function measures how good it is to be in a given state (or to take a given action in a given state) when following a particular policy. The goal of RL is to find a policy that maximizes value.

- Want to take actions in a way that maximizes value (or return):

$$\mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$$

  - This value depends on where you start and how you act

- Recall our bounded rewards assumption: $r_h \in [0, R_{\max}]$
  - What's the range of $\mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r_h\right]$? $\left[0, \frac{R_{\max}}{1-\gamma}\right]$

- A (deterministic) policy $\pi : \mathcal{S} \to \mathcal{A}$ describes how the agent acts: at state s_h, chooses action $a_h = \pi(s_h)$.

- More generally, the agent may choose actions randomly ($\pi : \mathcal{S} \to \Delta(\mathcal{A})$), or even in a way that varies across time steps ("non-stationary policies")

- **State-value function:** $V^\pi(s)$ answers "How good is it to be in state s if I follow policy

π?"

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r_h \,\middle|\, s_0 = s, \pi\right]$$

- **Action-value function (Q-function):** $Q^\pi(s, a)$ answers "How good is it to take action a in state s, then follow π?"

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s, a), \pi\right]$$

Above, the expectations are taken over trajectories generated by following policy π:

[for $V^\pi(s)$]: $s_0 = s, a_0 \sim \pi(\cdot \mid s_0), s_1 \sim \mathcal{P}(\cdot \mid s_0, a_0), a_1 \sim \pi(\cdot \mid s_1), s_2 \sim \mathcal{P}(\cdot \mid s_1, a_1), \ldots$

[for $Q^\pi(s, a)$]: $s_0 = s, a_0 = a, s_1 \sim \mathcal{P}(\cdot \mid s_0, a_0), a_1 \sim \pi(\cdot \mid s_1), s_2 \sim \mathcal{P}(\cdot \mid s_1, a_1), \ldots$

Lecture 2: Bellman Equation and Optimality

Bellman equation for policy evaluation

The *Bellman equation* is one of the most important equations in reinforcement learning. It expresses a recursive relationship: the value of a state equals the immediate reward plus the discounted value of the successor state. Intuitively, the value of being in state s can be decomposed into two parts:

1. The reward received in the current time step, and
2. The (discounted) value of wherever we end up next.

This recursive structure is the foundation of nearly all RL algorithms. Dynamic programming methods (policy iteration, value iteration) directly exploit it to compute value functions. Model-free algorithms like Q-learning and temporal-difference learning use it to define update rules. Understanding the Bellman equation is essential for the rest of this course.

We now derive the Bellman equations for both V^π and Q^π, then express them in matrix form.

Derivation for V^π.

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h r_h \,\middle|\, s_0 = s, \pi\right] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathbb{E}\left[r_0 + \sum_{h=1}^{\infty}\gamma^h r_h \,\middle|\, s_0 = s, a_0 = a, \pi\right]\right] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[R(s,a) + \sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)\mathbb{E}\left[\gamma\sum_{h=1}^{\infty}\gamma^{h-1} r_h \,\middle|\, s_1 = s', \pi\right]\right] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[R(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)\mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^{h} r_h \,\middle|\, s_0 = s', \pi\right]\right] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[R(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)V^\pi(s')\right]
\end{aligned}$$

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

Derivation for Q^π.

$$\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h r_h \,\middle|\, s_0 = s, a_0 = a, \pi\right] \\
&= \mathbb{E}\left[r_0 + \sum_{h=1}^{\infty}\gamma^h r_h \,\middle|\, s_0 = s, a_0 = a, \pi\right] \\
&= R(s,a) + \sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)\mathbb{E}\left[\gamma\sum_{h=1}^{\infty}\gamma^{h-1} r_h \,\middle|\, s_0 = s, a_0 = a, s_1 = s', \pi\right] \\
&= R(s,a) + \sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)\mathbb{E}\left[\gamma\sum_{h=1}^{\infty}\gamma^{h-1} r_h \,\middle|\, s_1 = s', \pi\right] \\
&= R(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)\mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h r_h \,\middle|\, s_0 = s', \pi\right] \\
&= R(s,a) + \gamma\sum_{s'\in\mathcal{S}}\mathcal{P}(s' \mid s,a)V^\pi(s') \\
&= R(s,a) + \gamma\langle\mathcal{P}(\cdot \mid s,a), V^\pi(\cdot)\rangle
\end{aligned}$$

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

Matrix form. Define

- V^π as the $|\mathcal{S}| \times 1$ vector $[V^\pi(s)]_{s\in\mathcal{S}}$
- R^π as the $|\mathcal{S}| \times 1$ vector whose s-th entry is $\mathbb{E}_{a\sim\pi(\cdot|s)}[R(s,a)]$
- $\mathcal{P}^\pi$ as the $|\mathcal{S}| \times |\mathcal{S}|$ matrix whose (s,s')-entry is $\mathbb{E}_{a\sim\pi(\cdot|s)}[\mathcal{P}(s' \mid s,a)]$

Then the Bellman equation can be written as:

$$\begin{aligned}
V^\pi &= R^\pi + \gamma\mathcal{P}^\pi V^\pi \\
(I - \gamma\mathcal{P}^\pi)V^\pi &= R^\pi \\
V^\pi &= (I - \gamma\mathcal{P}^\pi)^{-1}R^\pi
\end{aligned}$$

$(I - \gamma\mathcal{P}^\pi)$ is always invertible. Proof?

Proof of invertibility. There are two common ways to see this. First, define the vector ∞-norm and the induced matrix ∞-norm by

$$\|x\|_\infty := \max_j |x_j|, \qquad \|M\|_\infty := \max_i \sum_j |M_{ij}|,$$

where $|\cdot|$ denotes the modulus on $\mathbb{C}$, so that $\|Mx\|_\infty \le \|M\|_\infty \|x\|_\infty$ and in particular $\|M^t\|_\infty \le \|M\|_\infty^t$. Since $\mathcal{P}^\pi$ is row-stochastic, $\mathcal{P}^\pi_{ij} \ge 0$ and $\sum_j \mathcal{P}^\pi_{ij} = 1$ for each row i, hence $\|\mathcal{P}^\pi\|_\infty = 1$.

Spectral view (quick invertibility check). Let λ be an eigenvalue of $\mathcal{P}^\pi$, i.e., $\mathcal{P}^\pi v = \lambda v$ for some nonzero $v \in \mathbb{C}^{|\mathcal{S}|}$. Taking $\|\cdot\|_\infty$ on both sides,

$$|\lambda| \, \|v\|_\infty = \|\lambda v\|_\infty = \|\mathcal{P}^\pi v\|_\infty \le \|\mathcal{P}^\pi\|_\infty \|v\|_\infty = \|v\|_\infty,$$

so $|\lambda| \le 1$. Therefore every eigenvalue of $\gamma \mathcal{P}^\pi$ has magnitude at most $\gamma < 1$, and thus $1 - \gamma\lambda \ne 0$ for all eigenvalues. Hence $I - \gamma\mathcal{P}^\pi$ has no zero eigenvalues and is invertible.

Neumann series (explicit inverse). Let $A := \gamma \mathcal{P}^\pi$. Then $\|A\|_\infty = \gamma < 1$, so the Neumann series $S := \sum_{t=0}^\infty A^t$ converges. For the partial sum $S_T := \sum_{t=0}^T A^t$, we have the telescoping identity

$$(I - A) S_T = S_T (I - A) = I - A^{T+1}.$$

Moreover, $\|A^{T+1}\|_\infty \le \|A\|_\infty^{T+1} \to 0$, so letting $T \to \infty$ gives $(I - A)S = S(I - A) = I$, hence $S = (I - A)^{-1}$. Substituting back $A = \gamma\mathcal{P}^\pi$ yields

$$(I - \gamma \mathcal{P}^\pi)^{-1} = \sum_{t=0}^\infty (\gamma \mathcal{P}^\pi)^t.$$

State occupancy. The matrix $(1-\gamma) \cdot (I - \gamma\mathcal{P}^\pi)^{-1}$ has a natural interpretation: each row (indexed by s) is the *normalized discounted state occupancy distribution* $d^{\pi,s}$, whose (s')-th entry is

$$d^{\pi,s}(s') = (1-\gamma)\,\mathbb{E}\left[\sum_{h=0}^\infty \gamma^h \mathbb{1}[s_h = s'] \;\middle|\; s_0 = s, \pi\right].$$

Proof. From the Neumann series,

$$(1-\gamma)(I - \gamma\mathcal{P}^\pi)^{-1} = (1-\gamma)\sum_{t=0}^\infty (\gamma\mathcal{P}^\pi)^t = (1-\gamma)\sum_{t=0}^\infty \gamma^t (\mathcal{P}^\pi)^t.$$

The (s, s')-entry of this matrix is therefore

$$\left[(1-\gamma)(I-\gamma\mathcal{P}^\pi)^{-1}\right]_{s,s'} = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\left[(\mathcal{P}^\pi)^t\right]_{s,s'}.$$

Now, $(\mathcal{P}^\pi)^t$ is the t-step transition matrix under policy π, so $[(\mathcal{P}^\pi)^t]_{s,s'} = \Pr(s_t = s' \mid s_0 = s, \pi)$. Therefore:

$$\begin{aligned}\left[(1-\gamma)(I-\gamma\mathcal{P}^\pi)^{-1}\right]_{s,s'} &= (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\Pr(s_t = s' \mid s_0 = s, \pi)\\ &= (1-\gamma)\sum_{h=0}^{\infty}\gamma^h\mathbb{E}\left[\mathbb{1}[s_h = s'] \mid s_0 = s, \pi\right]\\ &= (1-\gamma)\,\mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h\mathbb{1}[s_h = s'] \,\middle|\, s_0 = s, \pi\right] = d^{\pi,s}(s').\end{aligned}$$

Thus $[(1-\gamma)(I-\gamma\mathcal{P}^\pi)^{-1}]_{s,s'} = d^{\pi,s}(s')$.

To verify row-stochasticity, note that $d^{\pi,s}(s') \geq 0$ and summing over s' gives

$$\sum_{s'\in\mathcal{S}} d^{\pi,s}(s') = (1-\gamma)\,\mathbb{E}\left[\sum_{h=0}^{\infty}\gamma^h\underbrace{\sum_{s'\in\mathcal{S}}\mathbb{1}[s_h = s']}_{=1} \,\middle|\, s_0 = s, \pi\right] = (1-\gamma)\sum_{h=0}^{\infty}\gamma^h = 1.$$

Therefore each row of $(1-\gamma)(I-\gamma\mathcal{P}^\pi)^{-1}$ is a probability distribution, so the matrix is row-stochastic. □

- The factor $(1-\gamma)$ is the normalization constant that makes each row of the matrix sum to 1, i.e., the matrix $(1-\gamma)(I-\gamma\mathcal{P}^\pi)^{-1}$ is row-stochastic.
- Using this, we can express the value function as:

  $$V^\pi(s) = \frac{1}{1-\gamma}\langle d^{\pi,s}, R^\pi\rangle,$$

  i.e., it is the dot product between the distribution $d^{\pi,s}$ and the reward vector R^π, scaled by $1/(1-\gamma)$.
- Alternatively, for a fixed $s' \in \mathcal{S}$, if we define an indicator reward function $R_{s'}(s, a) :=$

$\mathbb{1}[s = s']$, then

$$\frac{1}{1-\gamma} d^{\pi,s}(s') = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h R_{s'}(s_h, \pi(s_h)) \,\middle|\, s_0 = s, \pi\right]$$

is the corresponding value function.

Bellman Optimality

For infinite-horizon discounted MDPs (with finite $\mathcal{S}$ and $\mathcal{A}$), we establish three fundamental results that characterize optimal policies and their relationship to the Bellman optimality equations.

Main Results

Theorem 1 (Bellman Optimality for $V^\star$)**.** *There exists a* ***unique*** *function $V^\star$ that satisfies the Bellman optimality equation:*

$$V(s) = \max_{a \in \mathcal{A}} \left(R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[V(s')]\right)$$

Conversely, any function V satisfying this equation must equal $V^\star$.

Theorem 2 (Existence of Optimal Markov Policy)**.** *There exists a* ***stationary deterministic*** *policy $\pi^\star$ such that*

$$V^{\pi^\star}(s) \geq V^\pi(s) \quad \textit{for all } s \in \mathcal{S} \textit{ and } \textbf{\textit{all}} \textit{ policies } \pi,$$

where π ranges over all policies including non-stationary, stochastic, and history-dependent ones. In particular, for $V^\star$ defined in Theorem 1 and all $s \in \mathcal{S}$,

$$V^{\pi^\star}(s) = V^\star(s).$$

Theorem 3 (Bellman Optimality for $Q^\star$)**.** *There exists a* ***unique*** *function $Q^\star$ that satisfies the Q-Bellman optimality equation:*

$$Q(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q(s',a')\right]$$

Moreover, $Q^\star$ coincides with the optimal action-value function: for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$Q^\star(s,a) = Q^{\pi^\star}(s,a).$$

In particular, the optimal policy satisfies $\pi^\star(s) \in \operatorname{argmax}_{a\in\mathcal{A}} Q^\star(s,a)$.

Corollary 4 ($V^\star$ and $Q^\star$ Relationship)**.** *The optimal value functions are related by:*

$$V^\star(s) = \max_{a\in\mathcal{A}} Q^\star(s,a), \qquad Q^\star(s,a) = R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}V^\star(s').$$

The remainder of this section proves these results using the Banach Fixed Point Theorem.

Preliminaries: Bellman Operators and Contraction

We define the space of bounded value functions $\mathcal{B}(\mathcal{S}) = \mathbb{R}^{|\mathcal{S}|}$ equipped with the ℓ_∞ norm $\|V\|_\infty = \max_{s\in\mathcal{S}} |V(s)|$. This is a complete metric space (a Banach space).

1. Bellman Operators. For any *stationary* policy π (possibly stochastic), we define the *Bellman operator* $\mathcal{T}^\pi : \mathcal{B}(\mathcal{S}) \to \mathcal{B}(\mathcal{S})$ as:

$$(\mathcal{T}^\pi V)(s) := \mathbb{E}_{a\sim\pi(\cdot|s)}\left[R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[V(s')]\right].$$

We define the *Bellman optimality operator* $\mathcal{T}^\star : \mathcal{B}(\mathcal{S}) \to \mathcal{B}(\mathcal{S})$ as:

$$(\mathcal{T}^\star V)(s) := \max_{a\in\mathcal{A}}\left(R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[V(s')]\right).$$

Lemma 5 (Contraction)**.** *Both* $\mathcal{T}^\pi$ *and* $\mathcal{T}^\star$ *are* γ*-contractions under the* ℓ_∞ *norm. That is, for any* $U, V \in \mathcal{B}(\mathcal{S})$*:*

$$\|\mathcal{T}^\pi U - \mathcal{T}^\pi V\|_\infty \le \gamma\|U-V\|_\infty, \quad \textit{and} \quad \|\mathcal{T}^\star U - \mathcal{T}^\star V\|_\infty \le \gamma\|U-V\|_\infty.$$

Proof. For $\mathcal{T}^\pi$, the proof is straightforward linearity:

$$\begin{aligned}&|(\mathcal{T}^\pi U)(s) - (\mathcal{T}^\pi V)(s)| = \left|\gamma\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[U(s') - V(s')]\right| \\ \le\ &\gamma\mathbb{E}_{a\sim\pi(\cdot|s)}\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}|U(s') - V(s')| \le \gamma\|U-V\|_\infty.\end{aligned}$$

For $\mathcal{T}^\star$, fix s and let

$$a_s^\star \in \operatorname*{argmax}_{a\in\mathcal{A}}\left\{R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[U(s')]\right\}.$$

Then

$$\begin{aligned}(\mathcal{T}^\star U)(s) - (\mathcal{T}^\star V)(s) &= \left(R(s,a_s^\star) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a_s^\star)}[U(s')]\right) - \max_{a\in\mathcal{A}}\left(R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[V(s')]\right) \\ &\le \left(R(s,a_s^\star) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a_s^\star)}[U(s')]\right) - \left(R(s,a_s^\star) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a_s^\star)}[V(s')]\right) \\ &= \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a_s^\star)}[U(s') - V(s')] \le \gamma\|U-V\|_\infty.\end{aligned}$$

By symmetry, $(\mathcal{T}^\star V)(s) - (\mathcal{T}^\star U)(s) \le \gamma \|V - U\|_\infty$, which implies the result. □

Theorem 6 (Banach Fixed Point Theorem). *Let (X, d) be a complete metric space, and let $T : X \to X$ be a γ-contraction mapping (i.e., $d(T(x), T(y)) \le \gamma d(x, y)$ for some $\gamma \in [0, 1)$ and all $x, y \in X$). Then T has a unique fixed point $x^\star \in X$ (i.e., $T(x^\star) = x^\star$). Furthermore, for any $x_0 \in X$, the sequence defined by $x_{k+1} = T(x_k)$ converges to $x^\star$ as $k \to \infty$.*

Proof of Theorem 1 (Bellman Optimality for $V^\star$)

We now apply the Banach Fixed Point Theorem to show that the Bellman optimality operator has a unique fixed point, and then connect this fixed point to the optimal value function.

Theorem 7 (Policy Evaluation as a Fixed Point). *For any* stationary *policy π, the Bellman operator $\mathcal{T}^\pi$ has a unique fixed point in $\mathcal{B}(\mathcal{S})$. This fixed point is exactly the value function V^π, i.e., $\mathcal{T}^\pi V^\pi = V^\pi$. Moreover, for any $V_0 \in \mathcal{B}(\mathcal{S})$, the iterates $V_{k+1} \coloneqq \mathcal{T}^\pi V_k$ converge to V^π.*

Proof. Since $\mathcal{T}^\pi$ is a γ-contraction on the Banach space $\mathcal{B}(\mathcal{S})$, the **Banach Fixed Point Theorem** implies that it has a unique fixed point and that the iterates $V_{k+1} \coloneqq \mathcal{T}^\pi V_k$ converge to it. Moreover, the Bellman equation for policy evaluation gives $V^\pi = \mathcal{T}^\pi V^\pi$, so V^π is a fixed point. By uniqueness, it must be the unique fixed point of $\mathcal{T}^\pi$. □

Theorem 8 (Existence and Uniqueness of Bellman Fixed Point). *There exists a unique $V^\star \in \mathcal{B}(\mathcal{S})$ such that $\mathcal{T}^\star V^\star = V^\star$. That is, this $V^\star$ satisfies the* Bellman Optimality Equation*:*

$$V^\star(s) = \max_{a \in \mathcal{A}} \left(R(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[V^\star(s')] \right)$$

Moreover, for any $V_0 \in \mathcal{B}(\mathcal{S})$, the iterates $V_{k+1} \coloneqq \mathcal{T}^\star V_k$ converge to $V^\star$.

Proof. Since $\mathcal{T}^\star$ is a γ-contraction on the Banach space $\mathcal{B}(\mathcal{S})$, by the **Banach Fixed Point Theorem**, there exists a unique fixed point $V^\star$, and the iterates converge to it. (To see uniqueness, suppose there are two fixed points V and V'; then $\|V - V'\|_\infty = \|\mathcal{T}^\star V - \mathcal{T}^\star V'\|_\infty \le \gamma \|V - V'\|_\infty$. Since $\gamma < 1$, this implies $\|V - V'\|_\infty = 0$.) The Bellman Optimality Equation is simply the fixed-point condition $V^\star = \mathcal{T}^\star V^\star$ written out explicitly. □

Proof of Theorem 2 (Existence of Optimal Markov Policy)

Definition 1 (Greedy Policy). *Define the* greedy policy *$\pi^\star$ with respect to $V^\star$ as:*

$$\pi^\star(s) \in \operatorname*{argmax}_{a \in \mathcal{A}} \left(R(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[V^\star(s')] \right).$$

Remark 1. *The policy $\pi^\star$ is **stationary** (it depends only on the current state s, not on time) and **deterministic** (it selects a single action at each state).*

Theorem 9 (Greedy Policy Achieves $V^\star$)**.** *The greedy policy $\pi^\star$ achieves the value function $V^\star$, i.e., $V^{\pi^\star} = V^\star$.*

Proof. By definition of the greedy policy, for each state s:

$$(\mathcal{T}^{\pi^\star} V^\star)(s) = R(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,\pi^\star(s))}[V^\star(s')] = (\mathcal{T}^\star V^\star)(s).$$

Since $V^\star$ is the fixed point of $\mathcal{T}^\star$, we have $\mathcal{T}^{\pi^\star} V^\star = \mathcal{T}^\star V^\star = V^\star$. Thus $V^\star$ is a fixed point of $\mathcal{T}^{\pi^\star}$. Since $\mathcal{T}^{\pi^\star}$ is a γ-contraction, it has a *unique* fixed point, which is $V^{\pi^\star}$. Therefore, $V^{\pi^\star} = V^\star$. □

We now prove that the greedy policy $\pi^\star$ dominates all policies, including non-stationary and history-dependent ones.

Theorem 10 ($V^\star$ Dominates All Policies)**.** *For any policy π (including non-stationary, stochastic, and history-dependent policies), we have $V^\star(s) \geq V^\pi(s)$ for all $s \in \mathcal{S}$.*

Proof. Let $\pi = (\pi_0, \pi_1, \pi_2, \ldots)$ be any (possibly non-stationary, possibly stochastic, possibly history-dependent) policy, where at time h, $\pi_h(\cdot \mid s_0, a_0, \ldots, s_h)$ is a distribution over actions given the entire history. Define the T-step truncated value function:

$$V_T^\pi(s) := \mathbb{E}\left[\sum_{h=0}^{T-1} \gamma^h R(s_h, a_h) \,\middle|\, s_0 = s,\, \pi\right].$$

Also define $\widetilde{V}_T := (\mathcal{T}^\star)^T \mathbf{0}$, where $\mathbf{0}$ is the zero function. By the Banach Fixed Point Theorem, $\widetilde{V}_T \to V^\star$ as $T \to \infty$.

Claim: $\widetilde{V}_T(s) \geq V_T^\pi(s)$ for all $T \geq 0$, all policies π, and all states s.

Proof by induction on T:

Base case ($T = 0$): $\widetilde{V}_0(s) = 0 = V_0^\pi(s)$ for all s. ✓

Inductive step: Assume $\widetilde{V}_{T-1}(s) \geq V_{T-1}^{\pi'}(s)$ for *all* policies π' and all states s.[1]

Consider the value of policy π at state s. In the first step, π chooses an action a according to $\pi_0(\cdot \mid s)$ and transitions to $s' \sim \mathcal{P}(\cdot \mid s, a)$. After observing the one-step history (s, a, s'), the

[1] Formally, for any history $h_t = (s_0, a_0, \ldots, s_{t-1}, a_{t-1})$ observed up to time t, we can construct a new policy μ starting at time 0 by defining $\mu_k(a \mid s_0', a_0', \ldots, s_k') := \pi_{t+k}(a \mid h_t, s_t = s_0', a_t = a_0' \ldots, s_{t+k} = s_k')$. By the time-homogeneity of MDPs, the continuation value of π given history h_t equals the value of μ starting from s_t: $V_K^\pi(s_t \mid h_t) = V_K^\mu(s_t)$. The inductive hypothesis applies to *all* policies (including μ), so it implies $\widetilde{V}_{T-1}(s_t) \geq V_{T-1}^\pi(s_t \mid h_t)$ for any history h_t.

remainder of π induces a *continuation policy* $\pi^{(s,a,s')}$ for the remaining $T-1$ steps, which may be history-dependent. By the dynamic programming principle:

$$V_T^\pi(s) = \mathbb{E}_{a\sim\pi_0(\cdot|s),\, s'\sim\mathcal{P}(\cdot|s,a)}\left[R(s,a) + \gamma V_{T-1}^{\pi^{(s,a,s')}}(s')\right].$$

By the induction hypothesis (using the construction mentioned in the footnote), no matter what the continuation policy $\pi^{(s,a,s')}$ is (even if it depends on the history), its value is bounded by $\widetilde{V}_{T-1}$:

$$V_{T-1}^{\pi^{(s,a,s')}}(s') \le \widetilde{V}_{T-1}(s').$$

By definition of $\mathcal{T}^\star$:

$$\widetilde{V}_T(s) = (\mathcal{T}^\star \widetilde{V}_{T-1})(s) = \max_{a\in\mathcal{A}}\left(R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[\widetilde{V}_{T-1}(s')]\right).$$

Combining these:

$$\begin{aligned} V_T^\pi(s) &= \mathbb{E}_{a\sim\pi_0(\cdot|s),\, s'\sim\mathcal{P}(\cdot|s,a)}\left[R(s,a) + \gamma V_{T-1}^{\pi^{(s,a,s')}}(s')\right] \\ &\le \mathbb{E}_{a\sim\pi_0(\cdot|s)}\left[R(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[\widetilde{V}_{T-1}(s')]\right] \\ &\le \max_{a'\in\mathcal{A}}\left(R(s,a') + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a')}[\widetilde{V}_{T-1}(s')]\right) \\ &= \widetilde{V}_T(s). \quad \checkmark \end{aligned}$$

Taking $T\to\infty$: by the Banach Fixed Point Theorem, $\widetilde{V}_T \to V^\star$. For $V_T^\pi \to V^\pi$, note that π may be non-stationary, so we cannot directly apply Banach (there is no single operator $\mathcal{T}^\pi$). Instead, by the bounded rewards assumption:

$$|V^\pi(s) - V_T^\pi(s)| = \left|\mathbb{E}\left[\sum_{h=T}^\infty \gamma^h R(s_h,a_h) \,\middle|\, s_0 = s, \pi\right]\right| \le \mathbb{E}\left[\sum_{h=T}^\infty \gamma^h |R(s_h,a_h)| \,\middle|\, s_0 = s, \pi\right]$$
$$\le \sum_{h=T}^\infty \gamma^h R_{\max} = \frac{\gamma^T R_{\max}}{1-\gamma} \to 0.$$

Therefore, since $\widetilde{V}_T(s) \ge V_T^\pi(s)$ for all T and both sides converge as $T\to\infty$, we conclude $V^\star(s) \ge V^\pi(s)$ for all s. □

Corollary 11 (Equivalence of Bellman Fixed Point and Optimal Value)**.** *Let $V^\star_{\text{Bellman}}$ denote the unique fixed point of $\mathcal{T}^\star$, and define the* optimal value function *as $V^\star_{\text{sup}}(s) := \sup_\pi V^\pi(s)$, where the supremum is over all policies. Then:*

1. *$V^\star_{\text{Bellman}} = V^\star_{\text{sup}}$, i.e., the Bellman fixed point equals the optimal value function.*

2. $\pi^\star$ *(the greedy policy w.r.t.* $V^\star_{\text{Bellman}}$*) is optimal:* $V^{\pi^\star} = V^\star_{\text{sup}}$.

Henceforth, we write $V^\star$ *to denote this common value.*

Proof. From Theorem 10, $V^\star_{\text{Bellman}}(s) \geq V^\pi(s)$ for all policies π, so $V^\star_{\text{Bellman}}(s) \geq \sup_\pi V^\pi(s) = V^\star_{\text{sup}}(s)$.

Conversely, $\pi^\star$ is a policy, so $V^\star_{\text{sup}}(s) \geq V^{\pi^\star}(s) = V^\star_{\text{Bellman}}(s)$ (the last equality by Theorem "Greedy Policy Achieves $V^\star$").

Therefore $V^\star_{\text{Bellman}} = V^\star_{\text{sup}} = V^{\pi^\star}$. This proves that $\pi^\star$ is a **stationary, deterministic** policy that is optimal among **all** policies (Theorem 2), and that this optimal value is the unique solution to the Bellman optimality equation (Theorem 1). □

Remark 2. *A related result is that for any general policy* π *(possibly history-dependent), there always exists a Markov policy* μ *such that* $V^\pi(s) = V^\mu(s)$ *for all states. We will prove this claim in Homework 1.*

Proof of Theorem 3 (Bellman Optimality for $Q^\star$)

We now prove that $Q^\star$ is the unique fixed point of a Bellman optimality operator on Q-functions, paralleling the treatment for $V^\star$.

Bellman Operators for Q. Define the space $\mathcal{B}(\mathcal{S} \times \mathcal{A}) = \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ equipped with $\|Q\|_\infty = \max_{s,a} |Q(s,a)|$.

For any stationary policy π, define the Bellman operator $\mathcal{T}^\pi_Q : \mathcal{B}(\mathcal{S} \times \mathcal{A}) \to \mathcal{B}(\mathcal{S} \times \mathcal{A})$:

$$(\mathcal{T}^\pi_Q Q)(s,a) := R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \mathbb{E}_{a' \sim \pi(\cdot|s')}[Q(s',a')].$$

Define the Bellman optimality operator $\mathcal{T}^\star_Q : \mathcal{B}(\mathcal{S} \times \mathcal{A}) \to \mathcal{B}(\mathcal{S} \times \mathcal{A})$:

$$(\mathcal{T}^\star_Q Q)(s,a) := R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right].$$

Lemma 12 (Contraction for Q)**.** *Both* $\mathcal{T}^\pi_Q$ *and* $\mathcal{T}^\star_Q$ *are* γ*-contractions under the* ℓ_∞ *norm.*

Proof. For $\mathcal{T}^\star_Q$, fix (s,a) and let $a^\star_{s'} \in \operatorname{argmax}_{a'} Q_1(s',a')$. Then:

$$\begin{aligned} |(\mathcal{T}^\star_Q Q_1)(s,a) - (\mathcal{T}^\star_Q Q_2)(s,a)| &= \gamma \left| \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[\max_{a'} Q_1(s',a') - \max_{a'} Q_2(s',a') \right] \right| \\ &\leq \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} |Q_1(s',a^\star_{s'}) - Q_2(s',a^\star_{s'})| \\ &\leq \gamma \|Q_1 - Q_2\|_\infty. \end{aligned}$$

The proof for $\mathcal{T}_Q^\pi$ is similar (and simpler due to linearity). □

By the Banach Fixed Point Theorem, $\mathcal{T}_Q^\star$ has a unique fixed point. We now characterize this fixed point and relate it to $V^\star$.

Step 1: Express $Q^\star$ in terms of $V^\star$. Define $\bar{Q} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ by

$$\bar{Q}(s,a) \coloneqq R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^\star(s').$$

Then, using $V^\star = \mathcal{T}^\star V^\star$,

$$\begin{aligned}
(\mathcal{T}_Q^\star \bar{Q})(s,a) &= R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} \bar{Q}(s',a') \right] \\
&= R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} \left(R(s',a') + \gamma \mathbb{E}_{s'' \sim \mathcal{P}(\cdot|s',a')} V^\star(s'') \right) \right] \\
&= R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} (\mathcal{T}^\star V^\star)(s') \\
&= R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^\star(s') = \bar{Q}(s,a).
\end{aligned}$$

Thus $\bar{Q}$ is a fixed point of $\mathcal{T}_Q^\star$. By uniqueness of the fixed point, $\bar{Q} = Q^\star$. In particular,

$$Q^\star(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^\star(s'), \qquad V^\star(s) = \max_{a \in \mathcal{A}} Q^\star(s,a).$$

Step 2: Equivalence of Bellman Fixed Point and Optimal Action-Value. Let $Q^\star_{\text{Bellman}}$ be the unique fixed point of $\mathcal{T}_Q^\star$ (which we identified as $\bar{Q}$ in Step 1), and let $Q^\star_{\text{sup}}(s,a) \coloneqq \sup_\pi Q^\pi(s,a)$. For any policy π, by Theorem 10 we have $V^\pi(s) \le V^\star(s)$ for all s, hence for all (s,a),

$$Q^\pi(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^\pi(s') \le R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^\star(s') = Q^\star_{\text{Bellman}}(s,a).$$

Therefore $\sup_\pi Q^\pi(s,a) \le Q^\star_{\text{Bellman}}(s,a)$, i.e., $Q^\star_{\text{sup}} \le Q^\star_{\text{Bellman}}$.

Conversely, we first show that $Q^\star_{\text{Bellman}} = Q^{\pi^\star}$. Recall from Step 1 that $Q^\star_{\text{Bellman}}(s,a) = R(s,a) + \gamma \mathbb{E}_{s'} V^\star(s')$. Since we proved that $\pi^\star$ achieves $V^\star$ (i.e., $V^{\pi^\star} = V^\star$), we have

$$Q^{\pi^\star}(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^{\pi^\star}(s') = R(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V^\star(s') = Q^\star_{\text{Bellman}}(s,a).$$

Since $\pi^\star$ is a valid policy,

$$Q^\star_{\text{Bellman}}(s,a) = Q^{\pi^\star}(s,a) \le \sup_\pi Q^\pi(s,a) = Q^\star_{\text{sup}}(s,a).$$

Combining these gives $Q^\star_{\text{sup}} = Q^\star_{\text{Bellman}} (\eqqcolon Q^\star)$. □