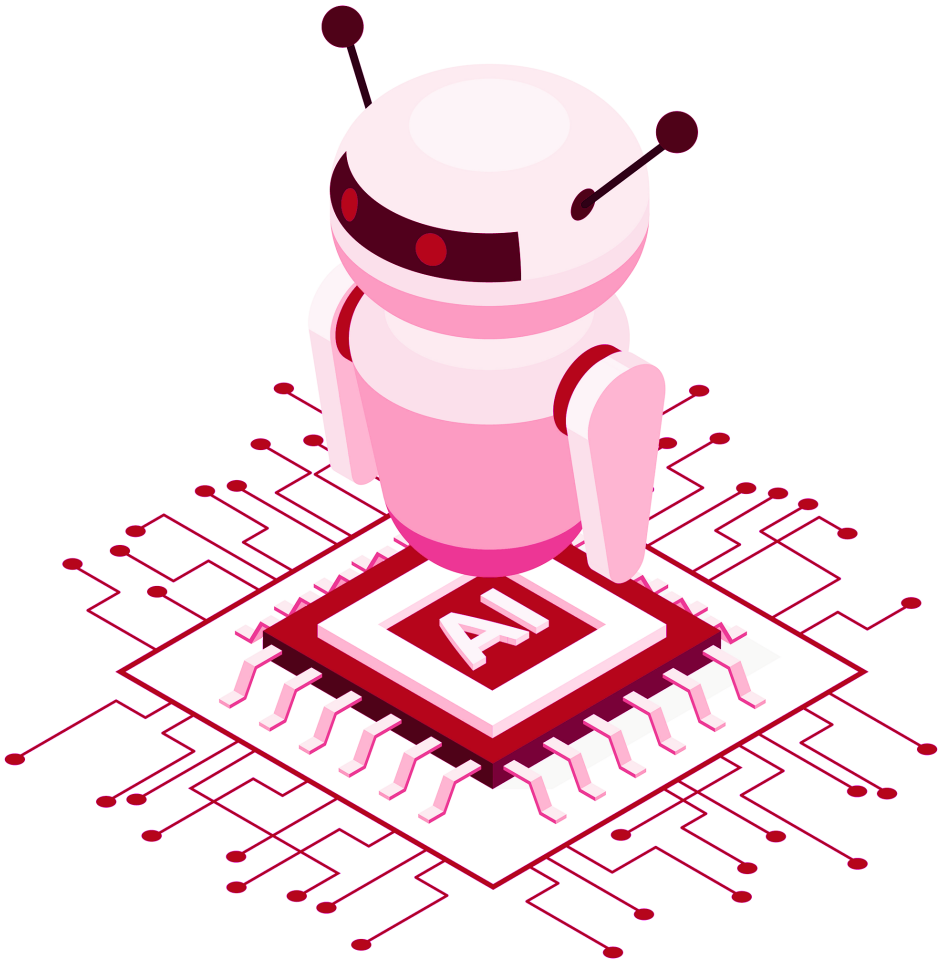

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 7:

Value-Based Theory

By:

Payam Taebi

400104867

RIML

Deep RL Course

Spring 2025

Contents

Grading

The grading will be based on the following criteria, with a total of 100 points:

Section	Points
Positive Rewards	15
General Rewards	10
Policy Turn	25
Bellman Operators	15
Bellman Residuals	35
Bonus 1: Writing your report in Latex	5
Bonus 2: Question 2.2.11	5

1 Iteration Family

Let $M = (S, A, R, P, \gamma)$ be a finite MDP with $|S| < \infty$, $|A| < \infty$, bounded rewards $|R(s,a)| \leq R_{\max}$ $\forall (s,a)$, and discount factor $\gamma \in [0,1)$. In this section, we will first explore an alternative proof approach for the value iteration algorithm, then we cover policy iteration which is discussed in the class more precisely.

1.1 Positive Rewards

Assume $R(s,a) \geq 0$ for all s, a.

1. Derive an upper bound for the optimal k-step value function V_k^*.

Upper Bound for the Optimal k-Step Value Function V_k^*

We will show by induction on k that for all $s \in S$,

$$V_k^*(s) \;\leq\; \frac{R_{\max}\big(1-\gamma^k\big)}{1-\gamma}\,.$$

Base Case ($k = 0$)

By definition of the 0-step value function,

$$V_0^*(s) = \max_{a \in A} R(s,a) \leq R_{\max}\,,$$

so the claimed bound holds for $k = 0$:

$$V_0^*(s) \leq \frac{R_{\max}(1-\gamma^0)}{1-\gamma} = \frac{R_{\max}(1-1)}{1-\gamma} = 0 \leq R_{\max}\,.$$

Inductive Step

Assume the bound holds for some $k \geq 0$, i.e.,

$$V_k^*(s) \;\leq\; \frac{R_{\max}\big(1-\gamma^k\big)}{1-\gamma} \quad \forall s \in S.$$

Consider the $(k+1)$-step value:

$$\begin{aligned}
V_{k+1}^*(s) &= \max_{a \in A}\Big\{R(s,a) \;+\; \gamma \sum_{s' \in S} P(s'|s,a)\, V_k^*(s')\Big\} \\
&\leq \max_{a \in A}\Big\{R_{\max} \;+\; \gamma \sum_{s'} P(s'|s,a)\, \frac{R_{\max}(1-\gamma^k)}{1-\gamma}\Big\} \quad \text{(by the inductive hypothesis)} \\
&= R_{\max} \;+\; \gamma\, \frac{R_{\max}(1-\gamma^k)}{1-\gamma} = \frac{R_{\max}\big((1-\gamma) + \gamma - \gamma^{k+1}\big)}{1-\gamma} \\
&= \frac{R_{\max}\big(1-\gamma^{k+1}\big)}{1-\gamma}\,.
\end{aligned}$$

Conclusion

By induction, for every $k \geq 0$ and $s \in S$,

$$\boxed{V_k^*(s) \;\leq\; \frac{R_{\max}\big(1-\gamma^k\big)}{1-\gamma}\,.}$$

This completes the derivation of the desired upper bound.

2. Prove V_k^* is non-decreasing in k. Giving a policy π such that:

$$V_{k+1}^{\pi} \geq V_k^*.$$

Use this to show convergence of Value Iteration to a solution satisfying the Bellman equation.

Monotonicity and Convergence of Value Iteration

We prove in full detail that V_k^* is non-decreasing in k, by exhibiting a policy π with

$$V_{k+1}^{\pi} \;\geq\; V_k^*,$$

and then show that the sequence $\{V_k^*\}$ converges to the unique solution of the Bellman optimality equation.

Setup and Notation

Define the Bellman optimality operator T on the space $\mathcal{B}(S)$ of bounded functions $V : S \to \mathbb{R}$ by

$$(TV)(s) \;=\; \max_{a\in A}\Big\{\, R(s,a) \;+\; \gamma \sum_{s'\in S} P(s' \mid s,a)\, V(s')\Big\}.$$

Likewise for a fixed policy $\pi : S \to A$, define the policy-evaluation operator T_π:

$$(T_\pi V)(s) \;=\; R\big(s,\pi(s)\big) \;+\; \gamma \sum_{s'} P\big(s' \mid s,\pi(s)\big)\, V(s').$$

Recall the k-step optimal value function satisfies the recursion

$$V_0^*(s) = \max_a R(s,a), \quad V_{k+1}^* \;=\; T\,V_k^*.$$

1. Monotonicity of T and T_π

- If $V(s) \leq W(s)$ for all s, then for each s, a,

$$R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V(s') \;\leq\; R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, W(s').$$

Taking $\max_a$ preserves this, so

$$V \leq W \quad\Longrightarrow\quad TV \leq TW.$$

Thus T is *monotone*. The same argument shows T_π is monotone:

$$V \leq W \;\Longrightarrow\; T_\pi V \leq T_\pi W \quad \forall \pi.$$

2. Greedy Policy Ensures $V_{k+1}^{\pi} \geq V_k^*$

Let V_k^* be given. Define the *greedy policy* π_k with respect to V_k^* by

$$\pi_k(s) \in \arg\max_{a \in A} \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V_k^*(s') \Big\}.$$

Then for each s,

$$\big(T_{\pi_k} V_k^*\big)(s) = R\big(s, \pi_k(s)\big) + \gamma \sum_{s'} P(s' \mid s, \pi_k(s))\, V_k^*(s') = \big(T V_k^*\big)(s) = V_{k+1}^*(s).$$

By monotonicity of T_{π_k} and since $V_k^* = T\, V_{k-1}^*$,

$$V_{k+1}^{\pi} \;=\; T_{\pi_k} V_k^* \;\geq\; T_{\pi_k} V_{k-1}^* \;=\; T\, V_{k-1}^* \;=\; V_k^*.$$

Moreover, because $V_{k+1}^* = \max_{\tilde{\pi}} V_{k+1}^{\tilde{\pi}}$, we conclude

$$\boxed{V_{k+1}^*(s) \;\geq\; V_k^*(s) \quad \forall s,\; k \geq 0.}$$

3. Boundedness and Pointwise Convergence

From the bound derived earlier,

$$V_k^*(s) \;\leq\; \frac{R_{\max}(1-\gamma^k)}{1-\gamma} \;\leq\; \frac{R_{\max}}{1-\gamma} \quad \forall s, k.$$

Hence for each state s, the sequence $\{V_k^*(s)\}_{k=0}^{\infty}$ is non-decreasing and bounded above, so it converges:

$$V_\infty(s) \;=\; \lim_{k \to \infty} V_k^*(s).$$

4. V_∞ Satisfies the Bellman Equation

Since $V_{k+1}^* = T\, V_k^*$, taking the limit as $k \to \infty$ and invoking continuity of T in the sup-norm (or by the contraction property below) gives

$$V_\infty = \lim_{k \to \infty} V_{k+1}^* = \lim_{k \to \infty} T\, V_k^* = T\left(\lim_{k \to \infty} V_k^*\right) = T\, V_\infty.$$

Thus V_∞ is a fixed point of T, i.e., it satisfies

$$V_\infty(s) = \max_{a \in A} \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V_\infty(s') \Big\}.$$

5. Uniqueness of the Bellman Fixed Point

To see why the Bellman operator T admits exactly one fixed point in $\mathcal{B}(S)$, we invoke its contraction property under the sup–norm. We give a self-contained proof:

Define the sup–norm for any bounded functions $V, W : S \to \mathbb{R}$ by

$$\|V - W\|_\infty \;=\; \max_{s \in S} \big|V(s) - W(s)\big|.$$

We recall that T is a γ-contraction:

$$\begin{aligned}\|TV - TW\|_\infty = \max_s \Big| & \max_a\{R(s,a) + \gamma \sum_{s'} P(s'|s,a)V(s')\} \\ & - \max_a\{R(s,a) + \gamma \sum_{s'} P(s'|s,a)W(s')\} \Big| \\ \leq \gamma \|V - W\|_\infty.\end{aligned}$$

Suppose, for the sake of contradiction, that there are two distinct fixed points V^α and V^β satisfying

$$V^\alpha = TV^\alpha, \quad V^\beta = TV^\beta, \quad V^\alpha \neq V^\beta.$$

Then applying the contraction inequality,

$$\|V^\alpha - V^\beta\|_\infty = \|TV^\alpha - TV^\beta\|_\infty \ \leq \ \gamma \|V^\alpha - V^\beta\|_\infty.$$

Since $0 \leq \gamma < 1$, we can rearrange:

$$(1-\gamma)\|V^\alpha - V^\beta\|_\infty \ \leq \ 0 \quad \Longrightarrow \quad \|V^\alpha - V^\beta\|_\infty = 0.$$

Hence $V^\alpha(s) = V^\beta(s)$ for all s, contradicting the assumption that they differ. Therefore there can be no two distinct fixed points.

By the Contraction Mapping (Banach) Theorem, T has exactly one fixed point in $\mathcal{B}(S)$. Since we have shown Value Iteration converges to $\lim_{k\to\infty} V_k^* = V_\infty$ and V_∞ is a fixed point of T, it must coincide with this unique fixed point, namely the optimal value function V^*.

Conclusion

We have shown:

- $V_{k+1}^* \geq V_k^*$ (monotonicity),
- $\{V_k^*\}$ is bounded above,
- the pointwise limit V_∞ satisfies $V_\infty = T\,V_\infty$,
- and by contraction, this fixed point is unique and equals V^*.

Hence Value Iteration converges:

$$\lim_{k\to\infty} V_k^* \ = \ V^*,$$

the unique solution of the Bellman optimality equation.

3. By taking the limit in the Bellman equation, prove that the V^* is optimal.

Optimality of V^* via the Bellman Equation

Let $V^* \in \mathcal{B}(S)$ denote the pointwise limit

$$V^*(s) \ = \ \lim_{k\to\infty} V_k^*(s),$$

which we have already shown satisfies the Bellman optimality equation

$$V^*(s) = \max_{a\in A}\Big\{R(s,a) + \gamma \sum_{s'\in S} P(s' \mid s,a)\, V^*(s')\Big\} \ = \ (TV^*)(s).$$

We now prove that V^* indeed equals the value function of some policy and dominates every other policy's value function, thus establishing its optimality.

1. V^* Dominates Any Fixed Policy

Fix any deterministic policy π. Its value function V^π satisfies the Bellman equation for policy π:

$$V^\pi = T_\pi V^\pi,$$

where

$$(T_\pi V)(s) = R\big(s, \pi(s)\big) + \gamma \sum_{s'} P\big(s' \mid s, \pi(s)\big)\, V(s').$$

Since $V^* = TV^*$ and T is monotone and satisfies $TV^* \geq T_\pi V^*$, we have

$$V^* \;=\; TV^* \;\geq\; T_\pi V^*.$$

Applying T_π repeatedly and using its monotonicity gives

$$V^* \;\geq\; T_\pi V^* \;\geq\; T_\pi^2 V^* \;\geq\; \cdots \;\geq\; \lim_{k\to\infty} T_\pi^k V^* = \; V^\pi.$$

Thus

$$\boxed{V^*(s) \;\geq\; V^\pi(s) \quad \forall s \in S,\, \forall \pi.}$$

2. Existence of a Policy Achieving V^*

Let π^* be any *greedy policy* with respect to V^*:

$$\pi^*(s) \;\in\; \arg\max_{a\in A}\Big\{R(s,a) + \gamma \sum_{s'} P(s' \mid s, a)\, V^*(s')\Big\}.$$

By construction,

$$\big(T_{\pi^*} V^*\big)(s) = \max_{a\in A}\{R(s,a) + \gamma P\, V^*(s')\} = (TV^*)(s) = V^*(s).$$

Thus V^* is a fixed point of T_{π^*} as well, and by standard policy-evaluation convergence,

$$V^{\pi^*} = \lim_{k\to\infty} T_{\pi^*}^k V^* = V^*.$$

Hence π^* attains value V^*.

3. Conclusion: V^* Is Optimal

Combining the two parts:

$$V^*(s) \;\geq\; V^\pi(s) \quad \forall \pi \quad \text{and} \quad V^{\pi^*}(s) = V^*(s),$$

we conclude that V^* is indeed the maximum achievable value function and thus is optimal. Equivalently,

$$V^* \;=\; \max_\pi V^\pi,$$

and π^* is an optimal policy.

1.2 General Rewards

Remove the non-negativity constraint on $R(s,a)$. Assume no terminating states exist. Consider a new MDP defined by adding a constant reward r_0 to all rewards of the current MDP. That is, for all (s,a), the new reward is:

$$\hat{R}(s,a) = R(s,a) + r_0$$

4. By deriving the optimal action and V_k^* in terms of the original MDP's values and r_0, show that Value Iteration still converges to the optimal value function V^* (and optimal policy) of the original MDP even if rewards are negative. Also compute the new value V^*.

Value Iteration with an Added Constant Reward

We transform the original MDP $M = (S, A, R, P, \gamma)$ by defining

$$\hat{R}(s, a) = R(s, a) + r_0\,, \quad \forall s \in S,\ a \in A,$$

obtaining $\hat{M} = (S, A, \hat{R}, P, \gamma)$. We prove:

1. Greedy actions under $\hat{R}$ coincide with those under R. 2. For each k, the k-step value functions satisfy

$$\hat{V}_k(s) = V_k(s) + \frac{r_0\big(1 - \gamma^k\big)}{1 - \gamma} \quad \forall s \in S.$$

3. As $k \to \infty$, $\hat{V}_k \to \hat{V}^*$ and $\hat{V}^*(s) = V^*(s) + \frac{r_0}{1-\gamma}$, so the same policy is optimal.

1. Greedy Actions Unchanged

(a) At iteration $k + 1$, Value Iteration on $\hat{M}$ computes

$$\hat{V}_{k+1}(s) = \max_{a \in A}\Big\{\hat{R}(s, a) + \gamma \sum_{s'} P(s' \mid s, a)\, \hat{V}_k(s')\Big\}.$$

(b) Substitute $\hat{R}(s, a) = R(s, a) + r_0$:

$$\hat{V}_{k+1}(s) = \max_{a}\Big\{R(s, a) + r_0 + \gamma \sum_{s'} P(s' \mid s, a)\, \hat{V}_k(s')\Big\}.$$

(c) Since r_0 is constant w.r.t. a, it does not affect the maximizing action:

$$\arg\max_a\{R(s, a) + \gamma\, \mathbb{E}[\hat{V}_k]\} = \arg\max_a\{R(s, a) + r_0 + \gamma\, \mathbb{E}[\hat{V}_k]\}.$$

(d) Thus the policy sequence produced under $\hat{R}$ matches that of the original MDP.

2. Relationship Between $\hat{V}_k$ and V_k

We prove by induction that for all $k \geq 0$ and $s \in S$,

$$\hat{V}_k(s) = V_k(s) + \frac{r_0\big(1 - \gamma^k\big)}{1 - \gamma}.$$

Base Case ($k = 0$).

$$V_0(s) = \max_a R(s, a),$$
$$\hat{V}_0(s) = \max_a \hat{R}(s, a) = \max_a\big(R(s, a) + r_0\big) = \max_a R(s, a) + r_0 = V_0(s) + r_0.$$

Note that $1 - \gamma^0 = 0$, so $\frac{r_0(1-\gamma^0)}{1-\gamma} = 0$ and indeed $\hat{V}_0 = V_0 + r_0$.

Inductive Step. Assume for some $k \geq 0$ that

$$\hat{V}_k(s) = V_k(s) + C_k, \quad C_k = \frac{r_0(1-\gamma^k)}{1-\gamma}.$$

Then

$$\begin{aligned}
\hat{V}_{k+1}(s) &= \max_a \Big\{ R(s,a) + r_0 + \gamma \sum_{s'} P(s'|s,a)\, \hat{V}_k(s') \Big\} \\
&= \max_a \Big\{ R(s,a) + r_0 + \gamma \sum_{s'} P(s'|s,a) \big[V_k(s') + C_k \big] \Big\} \\
&= \max_a \Big\{ R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V_k(s') \Big\} \; + \; r_0 + \gamma\, C_k \\
&= V_{k+1}(s) + r_0 + \gamma \frac{r_0(1-\gamma^k)}{1-\gamma} = V_{k+1}(s) + \frac{r_0\big(1-\gamma^{k+1}\big)}{1-\gamma}.
\end{aligned}$$

This completes the inductive argument.

3. Limit as $k \to \infty$

(a) Since $\gamma \in [0,1)$, $\gamma^k \to 0$. Taking limits,

$$\hat{V}^*(s) = \lim_{k\to\infty} \hat{V}_k(s) = \lim_{k\to\infty} \Big[V_k(s) + \frac{r_0(1-\gamma^k)}{1-\gamma} \Big] = V^*(s) + \frac{r_0}{1-\gamma}.$$

(b) The Bellman optimality equation for $\hat{V}^*$ is

$$\hat{V}^*(s) = \max_a \Big\{ R(s,a) + r_0 + \gamma \sum_{s'} P(s' \mid s,a)\, \hat{V}^*(s') \Big\}.$$

(c) Subtract $\frac{r_0}{1-\gamma}$ from both sides:

$$V^*(s) = \max_a \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^*(s') \Big\},$$

showing V^* satisfies the original Bellman equation.

4. Preservation of the Optimal Policy

Because adding r_0 simply shifts all value estimates by the same constant, the ordering of action values at each state is unchanged. Hence any greedy policy for $\hat{V}^*$ is also greedy for V^*, and thus optimal for the original MDP.

$$\boxed{\hat{V}^*(s) = V^*(s) + \frac{r_0}{1-\gamma}, \quad \text{and Optimal Policy is Unchanged.}}$$

5. Why is it necessary to assume the absence of a terminating state? Try to explain with a counterexample.

Necessity of "No Terminating State": A Counterexample

In the proof above, we crucially used the fact that every trajectory is infinite (no terminating state), so adding a constant reward r_0 simply shifts every value by the same constant $\frac{r_0}{1-\gamma}$ without changing the ordering of action-values. If we allowed a terminating (absorbing) state, this property can fail. We demonstrate this with a simple two-action MDP.

MDP Definition

- States: $S = \{s_0, s_T\}$, where s_0 is the start state and s_T is an absorbing terminal state.
- Actions in s_0: a_{end} ("terminate") and a_{loop} ("loop").
- Transitions:

$$P(s_T \mid s_0, a_{\text{end}}) = 1, \quad P(s_0 \mid s_0, a_{\text{loop}}) = 1, \quad P(s_T \mid s_T, \cdot) = 1.$$

- Rewards (original MDP):

$$R(s_0, a_{\text{end}}) = 0, \quad R(s_0, a_{\text{loop}}) = 0, \quad R(s_T, \cdot) = 0.$$

Under these rewards, both actions from s_0 yield the same infinite-horizon discounted return:

$$V^{\text{orig}}(s_0 \mid \pi_{\text{end}}) = 0, \quad V^{\text{orig}}(s_0 \mid \pi_{\text{loop}}) = 0.$$

Thus any policy is optimal for the original MDP.

Adding a Constant $r_0 > 0$

Define $\hat{R}(s, a) = R(s, a) + r_0$. Then:

$$\hat{V}(s_0 \mid \pi_{\text{end}}) = \hat{R}(s_0, a_{\text{end}}) + \gamma \underbrace{\hat{V}(s_T)}_{=0} = r_0,$$

$$\hat{V}(s_0 \mid \pi_{\text{loop}}) = \hat{R}(s_0, a_{\text{loop}}) + \gamma \hat{V}(s_0 \mid \pi_{\text{loop}}) = r_0 + \gamma \hat{V}(s_0 \mid \pi_{\text{loop}}),$$

so

$$\hat{V}(s_0 \mid \pi_{\text{loop}}) = \frac{r_0}{1-\gamma} > r_0 = \hat{V}(s_0 \mid \pi_{\text{end}}).$$

Consequently, under $\hat{R}$ the "loop" action strictly dominates "terminate," whereas under the original rewards they were equally good. The optimal policy changes merely because we added a constant to the rewards.

Conclusion

This counterexample shows that if trajectories can terminate, adding a constant reward r_0

$$\hat{R}(s, a) = R(s, a) + r_0$$

does not simply shift every value by the same constant; it can alter the optimal action. Hence the "no terminating state" assumption is essential to ensure Value Iteration—and the policy derived from it—remains invariant under adding a constant to all rewards.

1.3 Policy Turn

In this part we want to dive into the mathematical proof of policy iteration.

6. Let π_k be the policy at iteration k. Prove the following:

$$V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s) \quad \forall s \in S,$$

with strict inequality for at least one state unless π_k is already optimal. Use the definition of the greedy policy and explain why policy improvement leads to a better or equal value function.

Monotonic Improvement in Policy Iteration

Let π_k be the policy at iteration k, and let

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \Big\{ Q^{\pi_k}(s,a) \Big\},$$

be the *greedy policy* with respect to the action–value function of π_k:

$$Q^{\pi_k}(s,a) = R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^{\pi_k}(s').$$

We will show that

$$V^{\pi_{k+1}}(s) \geq V^{\pi_k}(s) \quad \forall s \in S,$$

and that the inequality is strict at some state unless π_k is already optimal.

1. Policy Improvement Inequality

By definition of π_{k+1},

$$Q^{\pi_k}\big(s, \pi_{k+1}(s)\big) = \max_a Q^{\pi_k}(s,a) \geq Q^{\pi_k}\big(s, \pi_k(s)\big) = V^{\pi_k}(s).$$

Thus for every state s,

$$Q^{\pi_k}\big(s, \pi_{k+1}(s)\big) \geq V^{\pi_k}(s).$$

2. From One-Step Improvement to Full Evaluation

Recall the Bellman operator for a fixed policy π:

$$(T_\pi V)(s) = R\big(s, \pi(s)\big) + \gamma \sum_{s'} P(s' \mid s, \pi(s))\, V(s').$$

Since $V^{\pi_{k+1}}$ is the unique fixed point of $T_{\pi_{k+1}}$, we have

$$V^{\pi_{k+1}} = T_{\pi_{k+1}} V^{\pi_{k+1}} \geq T_{\pi_{k+1}} V^{\pi_k},$$

where the inequality follows from monotonicity of $T_{\pi_{k+1}}$ and $V^{\pi_{k+1}} \geq V^{\pi_k}$ will be established next. But first observe

$$\big(T_{\pi_{k+1}} V^{\pi_k}\big)(s) = Q^{\pi_k}\big(s, \pi_{k+1}(s)\big) \geq V^{\pi_k}(s) = \big(T_{\pi_k} V^{\pi_k}\big)(s).$$

Therefore

$$V^{\pi_{k+1}}(s) = \big(T_{\pi_{k+1}} V^{\pi_{k+1}}\big)(s) \geq \big(T_{\pi_{k+1}} V^{\pi_k}\big)(s) \geq \big(T_{\pi_k} V^{\pi_k}\big)(s) = V^{\pi_k}(s).$$

3. Strict Improvement Unless Already Optimal

- If π_k is not optimal, then there exists some state s_0 where $\pi_{k+1}(s_0) \neq \pi_k(s_0)$. For that s_0, the maximization $\max_a Q^{\pi_k}(s_0, a)$ is strictly greater than $Q^{\pi_k}(s_0, \pi_k(s_0))$, so

$$Q^{\pi_k}\big(s_0, \pi_{k+1}(s_0)\big) > V^{\pi_k}(s_0).$$

This strict inequality propagates through the evaluation operator, yielding $V^{\pi_{k+1}}(s_0) > V^{\pi_k}(s_0)$.
- If no such state exists, then $\pi_{k+1}(s) = \pi_k(s)$ for all s, which implies π_k is already greedy with respect to V^{π_k}, and hence optimal.

4. Conclusion

Combining the above, we obtain

$$\boxed{V^{\pi_{k+1}}(s) \;\geq\; V^{\pi_k}(s) \quad \forall s \in S,}$$

with strict inequality for at least one state precisely when π_k is not yet optimal. Thus each policy iteration step never decreases the value function and strictly improves it until optimality is reached.

7. Prove that Policy Iteration always converges to the optimal policy in a finite MDP. Specifically, show that after a finite number of policy evaluations and improvements, the algorithm reaches a policy π^* that satisfies the Bellman optimality equation. You may use theorems discussed in class, but if a result was not proven, please provide a full justification.

Finite-Step Convergence of Policy Iteration

We consider Policy Iteration on a finite MDP $M = (S, A, R, P, \gamma)$ with $|S| < \infty$ and $|A| < \infty$. Recall that each iteration k consists of

1. **Policy Evaluation:** Compute V^{π_k}, the unique solution of

$$V(s) = R\big(s, \pi_k(s)\big) + \gamma \sum_{s'} P(s' \mid s, \pi_k(s))\, V(s') \quad \forall s \in S.$$

2. **Policy Improvement:** Define

$$\pi_{k+1}(s) \in \arg\max_{a \in A} \Big\{ R(s, a) + \gamma \sum_{s'} P(s' \mid s, a)\, V^{\pi_k}(s') \Big\}.$$

We prove that this process terminates in a finite number of steps at an optimal policy π^*.

1. Strict Improvement or Termination

From the Policy Improvement Theorem, we have for every s:

$$V^{\pi_{k+1}}(s) \;\geq\; V^{\pi_k}(s),$$

with strict inequality for at least one state unless $\pi_{k+1} = \pi_k$. Thus either

- $\pi_{k+1} \neq \pi_k$ and $V^{\pi_{k+1}} > V^{\pi_k}$ (strictly), or - $\pi_{k+1} = \pi_k$.

2. Finiteness of the Policy Space

There are only finitely many deterministic stationary policies, at most $|A|^{|S|}$. Since Policy Iteration never revisits a policy with strictly lower (or equal) value, the sequence $\{\pi_k\}$ cannot cycle through infinitely many distinct policies: each strict improvement moves to a new policy with strictly higher value at some state, and there are only finitely many policies.

3. Termination Implies Optimality

When the algorithm reaches an iteration K such that $\pi_{K+1} = \pi_K$, Policy Improvement has no further effect. By definition of the greedy update,

$$\pi_K(s) \in \arg\max_a \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^{\pi_K}(s') \Big\},$$

so π_K is greedy with respect to its own value function V^{π_K}. Therefore V^{π_K} satisfies the Bellman optimality equation:

$$V^{\pi_K}(s) = \max_a \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^{\pi_K}(s') \Big\},$$

and by uniqueness of the Bellman optimal solution, $V^{\pi_K} = V^*$ and π_K is an optimal policy π^*.

4. Conclusion

- **Strict improvements** ensure that as long as $\pi_{k+1} \neq \pi_k$, we move to a strictly better policy. - **Finiteness** of the policy set guarantees that we cannot improve strictly more than $|A|^{|S|} - 1$ times. - **Termination condition** $\pi_{K+1} = \pi_K$ implies the Bellman optimality equation holds, so the final policy is optimal.

Hence Policy Iteration converges in at most $|A|^{|S|}$ iterations to a policy π^* satisfying

$$V^{\pi^*}(s) = \max_a \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^{\pi^*}(s') \Big\},$$

i.e. the unique optimal policy.

8. Prove that Value Iteration and Policy Iteration both converge to the same optimal value function V^*, even if the policies may differ. How the policies are still optimal despite possible differences?

Equivalence of Value Iteration and Policy Iteration

We show that both algorithms converge to the same optimal value function V^* and that any policy they produce (even if different) is optimal.

1. Both Algorithms Converge to the Unique Fixed Point of T

Value Iteration. Recall that Value Iteration produces the sequence

$$V_{k+1} = T\, V_k$$

and we have shown

$$\|V_{k+1} - V^*\|_\infty = \|T\, V_k - T\, V^*\|_\infty \leq \gamma \|V_k - V^*\|_\infty.$$

By the contraction mapping theorem, $V_k \to V^*$, the unique fixed point of T.

Policy Iteration. Policy Iteration produces a sequence of policies π_k and values V^{π_k}. We showed

$$V^{\pi_0} \leq V^{\pi_1} \leq \cdots \leq V^*,$$

and that the process terminates in finitely many steps with π_K satisfying

$$V^{\pi_K} = T\,V^{\pi_K},$$

hence $V^{\pi_K} = V^*$. Thus Policy Iteration converges to the same V^*.

2. Uniqueness of the Optimal Value Function

Since the Bellman optimality operator T is a γ-contraction on $\mathcal{B}(S)$, it has exactly one fixed point V^*. Both algorithms converge to that unique fixed point, so they produce the same value function in the limit.

3. Possible Differences in Policies

Multiple Greedy Policies. Given V^*, any "greedy" policy defined by

$$\pi(s) \in \arg\max_{a \in A} \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^*(s') \Big\}$$

is optimal. If the maximizer is not unique at some state, different choices of $\pi(s)$ still attain the same value $V^*(s)$.

Algorithms' Tie-Breaking. - Value Iteration often extracts a policy π_k by choosing at each state one maximizer of the one-step update on V_k. - Policy Iteration refines policies via greedy improvement, possibly choosing different maximizers in tie cases.

When ties occur, the two algorithms may select different actions, but each respects

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^*(s'),$$

so both policies satisfy

$$V^{\pi}(s) = V^*(s) \quad \forall s.$$

4. Conclusion

- Both Value Iteration and Policy Iteration converge to the unique optimal value function V^*.
- Any policy that is greedy w.r.t. V^* is optimal, even if different algorithms break ties differently.
- Hence, despite potentially different action selections in tie-situations, all policies produced at convergence achieve the same optimal performance.

9. Compare and contrast the computational cost of one step of Policy Iteration (i.e., full Policy Evaluation + Policy Improvement) versus one iteration of Value Iteration.

Computational Cost: Policy Iteration vs. Value Iteration

We compare the cost of one full Policy Iteration step (Policy Evaluation + Policy Improvement) to one sweep of Value Iteration, in terms of $|S|$, $|A|$, and the number of nonzero transition probabilities (denote by $N \approx |S|\,|A|$ in the dense case).

1. One Policy Iteration Step

(a) Policy Evaluation. Given a fixed policy π, Policy Evaluation requires computing V^π by solving the system of linear equations

$$V(s) \;=\; R\big(s,\pi(s)\big) \;+\; \gamma \sum_{s'} P(s' \mid s,\pi(s))\, V(s'), \quad \forall s \in S.$$

- *Direct solution* (e.g. Gaussian elimination) costs $O(|S|^3)$.
- *Iterative evaluation* (e.g. Jacobi or Gauss–Seidel) requires N_{eval} sweeps, each sweep costing $O(N) \approx O(|S|\,|A|)$.

(b) Policy Improvement. Once V^π is known, updating the policy at each state entails

$$\pi'(s) \;\in\; \arg\max_{a \in A} \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^\pi(s') \Big\},$$

which costs $O(N)$ to evaluate all $Q^\pi(s,a)$ values.

Total cost per Policy Iteration step:

$$\underbrace{O(\,|S|^3\,) \text{ or } O\big(N_{\text{eval}} \cdot N\big)}_{\text{Policy Evaluation}} + \underbrace{O\big(N\big)}_{\text{Policy Improvement}}.$$

2. One Value Iteration Sweep

Value Iteration performs a single Bellman backup at each state:

$$V_{k+1}(s) = \max_{a \in A} \Big\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V_k(s') \Big\},$$

which costs $O(N)$ per sweep over all (s,a) pairs.

$$\text{Cost per iteration: } O\big(N\big) \approx O(|S|\,|A|).$$

3. Trade-Offs and Practical Considerations

- **Per-step cost:** Value Iteration is $O(N)$ per iteration, whereas Policy Iteration is dominated by evaluation $O(|S|^3)$ (or $O(N_{\text{eval}}\, N)$).
- **Number of iterations:** Value Iteration typically requires $O\big(\frac{\ln(1/\epsilon)}{1-\gamma}\big)$ sweeps to achieve $\|V_k - V^*\|_\infty \le \epsilon$. Policy Iteration often converges in far fewer than $|A|^{|S|}$ iterations (often tens of policy improvements), each expensive but few in number.
- **Sparse transitions:** When P is sparse, $N \ll |S|\,|A|$, both methods benefit, but Policy Evaluation savings accrue only if iterative methods converge quickly.
- **Hybrid methods:** Modified Policy Iteration (truncated evaluation) interpolates between the two, performing a small number of evaluation sweeps per improvement, balancing per-step cost and iteration count.

Conclusion: • Value Iteration has low cost per iteration but may need many iterations. • Policy Iteration has high cost per step (due to full evaluation) but typically converges in very few steps. • Practical choice depends on $|S|$, $|A|$, transition sparsity, and required precision ϵ.

10. In the context of a (MDP) with an infinite horizon, when the discount factor $\gamma = 1$, analyze how both Value Iteration and Policy Iteration behave.

Behavior of Value and Policy Iteration When $\gamma = 1$

We now consider the infinite-horizon case with discount factor $\gamma = 1$. In this setting the usual contraction-based arguments break down, and both algorithms require additional assumptions (e.g. proper policies or termination) to behave well. We analyze their key issues and possible remedies.

1. Loss of Contraction and Divergence of Value Iteration

When $\gamma < 1$, the Bellman optimality operator T is a γ–contraction in the sup–norm. But at $\gamma = 1$,

$$(TV)(s) = \max_a \Big\{ R(s,a) + \sum_{s'} P(s' \mid s,a)\, V(s') \Big\}$$

is only a *monotone* operator, not a contraction. Consequences include:

- **No guaranteed unique fixed point:** There may be multiple bounded solutions of $V = T\,V$, or none if total rewards diverge.
- **Value Iteration may oscillate or diverge:** Without a discount "shrink," the sequence

  $$V_{k+1} = T\,V_k$$

  need not converge. Even if it converges pointwise to some V_∞, this limit need not satisfy $V_\infty = T\,V_\infty$ or be optimal.
- **Unbounded returns:** If there exist cycles with positive total reward, the undiscounted return $\sum_{t=0}^{\infty} R(s_t, a_t)$ may diverge to $+\infty$, so $V^*(s) = +\infty$.

Remedy: One typically imposes a *stochastic shortest-path* structure or "proper policies" assumption: there is a designated terminal state, and all admissible policies reach it with probability one in finite expected time. Under such conditions, a modified Bellman operator becomes a contraction in a weighted norm, and Value Iteration converges.

2. Policy Iteration Without Discounting

Policy Iteration still alternates evaluation and greedy improvement:

$$\begin{aligned} V^{\pi_k} &\leftarrow \text{ solve } V = R_\pi + P_\pi V, \\ \pi_{k+1}(s) &\in \arg\max_a \{ R(s,a) + P(\cdot \mid s,a)\, V^{\pi_k} \}. \end{aligned}$$

However:

- **Evaluation may be ill-posed:** If there are recurrent cycles with non-zero net reward, the linear system $(I - P_\pi)V = R_\pi$ may have no finite solution.

- **Monotonic improvement still holds (when values are finite):** Whenever the evaluation succeeds, the policy improvement theorem remains valid, so $V^{\pi_{k+1}} \geq V^{\pi_k}$.
- **Finite termination under proper-policy assumptions:** If every policy is "proper" (i.e. reaches a terminal in finite expected time), then Policy Iteration still converges in finitely many steps to an optimal proper policy.

3. Summary of Key Points

(a) Without discounting ($\gamma = 1$), T is not a contraction. Value Iteration can fail to converge or produce unbounded values.

(b) Policy Iteration's guarantee of strict policy improvement remains, but evaluation can break if returns are infinite or no termination is guaranteed.

(c) To recover convergence, one must impose additional structure—commonly a proper-policy or stochastic-shortest-path framework—so that all returns remain finite and a contraction-like property holds.

Conclusion: In the undiscounted infinite-horizon case, both algorithms require strong model assumptions (e.g. guaranteed termination) to ensure meaningful convergence. Otherwise, Value Iteration may diverge, and Policy Iteration's evaluations may be ill-posed.

2 Bellman or Bellwoman

[1] Recall that a value function is a $|S|$-dimensional vector where $|S|$ is the number of states of the MDP. When we use the term V in these expressions as an "arbitrary value function", we mean that V is an arbitrary $|S|$-dimensional vector which need not be aligned with the definition of the MDP at all. On the other hand, V^π is a value function that is achieved by some policy π in the MDP. For example, say the MDP has 2 states and only negative immediate rewards. $V = [1, 1]$ would be a valid choice for V even though this value function can never be achieved by any policy π, but we can never have a $V^\pi = [1, 1]$. This distinction between V and V^π is important for this question and more broadly in reinforcement learning.

2.1 Bellman Operators

In the first part of this problem, we will explore some general and useful properties of the Bellman backup operator. We know that the Bellman backup operator B, defined below, is a contraction with the fixed point as V^*, the optimal value function of the MDP. The symbols have their usual meanings. γ is the discount factor and $0 \leq \gamma < 1$. In all parts, $\|v\| = \max_s |v(s)|$ is the infinity norm of the vector.

$$(BV)(s) = \max_a \left[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s') \right]$$

We also saw the contraction operator B^π with the fixed point V^π, which is the Bellman backup operator for a particular policy given below:

$$(B^\pi V)(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V(s')$$

In this case, we'll assume π is deterministic, but it doesn't have to be in general. You have seen that $\|BV - BV'\| \leq \gamma \|V - V'\|$ for two arbitrary value functions V and V'.

1. Show that the analogous inequality, $\|B^\pi V - B^\pi V'\| \leq \gamma \|V - V'\|$, holds.

Contraction of the Policy-Evaluation Operator B^π

We wish to show that for any two arbitrary value vectors $V, V' \in \mathbb{R}^{|S|}$,

$$\|B^\pi V \; - \; B^\pi V'\|_\infty \; \leq \; \gamma \, \|V - V'\|_\infty .$$

Proof

By definition, for each state $s \in S$,

$$(B^\pi V)(s) = r\big(s, \pi(s)\big) \; + \; \gamma \sum_{s'} P(s' \mid s, \pi(s)) \, V(s'),$$

and similarly for V'. Consider the difference at an arbitrary s:

$$\begin{aligned}\big|(B^\pi V)(s) \;-\; (B^\pi V')(s)\big| &= \Big|\gamma \sum_{s'} P(s' \mid s, \pi(s))\big[V(s') - V'(s')\big]\Big| \\ &\le\; \gamma \sum_{s'} P(s' \mid s, \pi(s))\big|V(s') - V'(s')\big| \quad \text{(triangle inequality)} \\ &\le\; \gamma \sum_{s'} P(s' \mid s, \pi(s))\, \|V - V'\|_\infty \;=\; \gamma\, \|V - V'\|_\infty,\end{aligned}$$

since $\sum_{s'} P(s' \mid s, \pi(s)) = 1$. Taking the maximum over s gives

$$\|B^\pi V - B^\pi V'\|_\infty = \max_s \big|(B^\pi V)(s) - (B^\pi V')(s)\big| \le \gamma\, \|V - V'\|_\infty,$$

as required.

2. Prove that the fixed point for B^π is unique. Recall that the fixed point is defined as V satisfying $V = B^\pi V$. You may assume that a fixed point exists.

Uniqueness of the Fixed Point of B^π

Assume there exists at least one fixed point V satisfying

$$V = B^\pi V.$$

Suppose, for the sake of contradiction, that there are two (possibly distinct) fixed points V and V' with

$$V = B^\pi V, \quad V' = B^\pi V'.$$

We will show $V = V'$ by exploiting the contraction property of B^π.

1. Apply the Contraction Inequality

From the result

$$\|B^\pi V - B^\pi V'\|_\infty \;\le\; \gamma\, \|V - V'\|_\infty,$$

and using $B^\pi V = V$ and $B^\pi V' = V'$, we get

$$\|V - V'\|_\infty = \big\|B^\pi V - B^\pi V'\big\|_\infty \;\le\; \gamma\, \|V - V'\|_\infty.$$

2. Conclude Uniqueness

Since $0 \le \gamma < 1$, rearrange the inequality:

$$(1 - \gamma)\, \|V - V'\|_\infty \;\le\; 0 \quad \Longrightarrow \quad \|V - V'\|_\infty = 0.$$

Hence $V(s) = V'(s)$ for all $s \in S$, proving that the fixed point of B^π is unique.

3. Suppose that V and V' are vectors satisfying $V(s) \le V'(s)$ for all s. Show that $B^\pi V(s) \le B^\pi V'(s)$ for all s. *Note: all of these inequalities are elementwise.*

Monotonicity of the Policy-Evaluation Operator B^π

Let $V, V' \in \mathbb{R}^{|S|}$ satisfy

$$V(s) \;\le\; V'(s) \quad \forall s \in S.$$

We show that

$$(B^\pi V)(s) \;\le\; (B^\pi V')(s) \quad \forall s \in S.$$

Proof

By definition of B^π, for each $s \in S$,

$$(B^\pi V)(s) = r\big(s, \pi(s)\big) \;+\; \gamma \sum_{s'} P(s' \mid s, \pi(s))\, V(s'),$$

and

$$(B^\pi V')(s) = r\big(s, \pi(s)\big) \;+\; \gamma \sum_{s'} P(s' \mid s, \pi(s))\, V'(s').$$

Since $V(s') \le V'(s')$ for every s' and all transition probabilities $P(s'|s, \pi(s))$ are nonnegative,

$$\sum_{s'} P(s' \mid s, \pi(s))\, V(s') \;\le\; \sum_{s'} P(s' \mid s, \pi(s))\, V'(s').$$

Multiplying by $\gamma \ge 0$ and adding the common reward term $r(s, \pi(s))$ preserves the inequality, yielding

$$(B^\pi V)(s) \;\le\; (B^\pi V')(s).$$

Because this holds for every s, we conclude elementwise

$$B^\pi V \;\le\; B^\pi V'.$$

2.2 Bellman Residuals

We can extract a greedy policy π from an arbitrary value function V using the equation below:

$$\pi(s) = \arg\max_a \left[r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V(s') \right]$$

It is often helpful to know what the performance will be if we extract a greedy policy from an arbitrary value function. To see this, we introduce the notion of a Bellman residual.

Define the Bellman residual to be $(BV - V)$ and the Bellman error magnitude to be $\|BV - V\|$.

4. For what value function V does the Bellman error magnitude $\|BV - V\|$ equal 0? Why?

Zero Bellman Error and the Optimal Value Function

We recall the Bellman residual for an arbitrary value vector V is

$$BV - V,$$

and its magnitude is measured by the sup-norm $\|BV - V\|_\infty$. We ask: for what V does $\|BV - V\|_\infty = 0$?

Characterization of Zero Residual

(a) *Zero residual means $BV = V$.* By definition of the norm,

$$\|BV - V\|_\infty = 0 \quad \Longleftrightarrow \quad (BV)(s) - V(s) = 0 \quad \forall s \in S \quad \Longleftrightarrow \quad BV = V.$$

(b) *Fixed-point of the Bellman optimality operator.* The equation

$$V = BV$$

is exactly the Bellman optimality equation. It is well-known (via the contraction argument) that this equation has a unique solution, namely the optimal value function V^*.

Conclusion

Therefore,

$$\boxed{\|BV - V\|_\infty = 0 \quad \Longleftrightarrow \quad V = V^*,}$$

because only the optimal value function V^* satisfies the fixed-point equation $BV^* = V^*$.

5. Prove the following statements for an arbitrary value function V and any policy π.

$$\|V - V^\pi\| \leq \frac{\|V - B^\pi V\|}{1 - \gamma}$$
$$\|V - V^*\| \leq \frac{\|V - BV\|}{1 - \gamma}$$

Performance Bounds via Bellman Residuals

We prove the two inequalities

$$\|V - V^\pi\|_\infty \leq \frac{\|V - B^\pi V\|_\infty}{1 - \gamma}, \qquad \|V - V^*\|_\infty \leq \frac{\|V - BV\|_\infty}{1 - \gamma},$$

for any arbitrary value vector V, any policy π, and the optimal operator B.

1. Bound for $\|V - V^\pi\|$

Recall that V^π is the unique fixed point of B^π: $V^\pi = B^\pi V^\pi$, and B^π is a γ–contraction. We write

$$V - V^\pi = \big(V - B^\pi V\big) + \big(B^\pi V - B^\pi V^\pi\big).$$

Taking the sup-norm and using the contraction and triangle inequalities:

$$\begin{aligned}\|V - V^\pi\|_\infty &\leq \|V - B^\pi V\|_\infty + \|B^\pi V - B^\pi V^\pi\|_\infty \\ &\leq \|V - B^\pi V\|_\infty + \gamma \|V - V^\pi\|_\infty.\end{aligned}$$

Rearranging gives

$$(1 - \gamma) \|V - V^\pi\|_\infty \leq \|V - B^\pi V\|_\infty \implies \boxed{\|V - V^\pi\|_\infty \leq \frac{\|V - B^\pi V\|_\infty}{1 - \gamma}.}$$

2. Bound for $\|V - V^*\|$

An identical argument applies with B and its fixed point $V^* = BV^*$. Write

$$V - V^* = \big(V - BV\big) + \big(BV - BV^*\big),$$

so

$$\|V - V^*\|_\infty \le \|V - BV\|_\infty + \|BV - BV^*\|_\infty \le \|V - BV\|_\infty + \gamma \|V - V^*\|_\infty.$$

Rearranging gives

$$\boxed{\|V - V^*\|_\infty \le \frac{\|V - BV\|_\infty}{1 - \gamma}.}$$

These two bounds quantify how the Bellman residual controls the suboptimality of an arbitrary value estimate V.

6. Let V be an arbitrary value function and π be the greedy policy extracted from V. Let $\varepsilon = \|BV - V\|$ be the Bellman error magnitude for V. Prove the following for any state s.

$$V^\pi(s) \ge V^*(s) - \frac{2\varepsilon}{1 - \gamma}$$

Performance of the Greedy Policy from an Approximate Value

Let V be any value vector, and let π be the greedy policy extracted from V:

$$\pi(s) \in \arg\max_a \Big\{ r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V(s') \Big\}.$$

Define the Bellman error magnitude $\varepsilon = \|BV - V\|_\infty$. We show that for every state s,

$$V^\pi(s) \;\ge\; V^*(s) \;-\; \frac{2\,\varepsilon}{1 - \gamma}.$$

1. Relating ε to the Policy-Evaluation Residual

Since π is greedy w.r.t. V, we have for each s:

$$(B^\pi V)(s) = r\big(s, \pi(s)\big) + \gamma \sum_{s'} P(s'|s,\pi(s))\, V(s') = \max_a \Big\{ r(s,a) + \gamma \sum_{s'} P\, V \Big\} = (BV)(s).$$

Hence elementwise

$$B^\pi V \;=\; BV,$$

and therefore

$$\| B^\pi V - V\|_\infty = \| BV - V\|_\infty = \varepsilon.$$

2. Bounding $\|V^\pi - V\|$ and $\|V - V^*\|$

From part (e) we have the two Bellman-residual bounds:

$$\|V - V^\pi\|_\infty \;\le\; \frac{\|V - B^\pi V\|_\infty}{1 - \gamma} \;=\; \frac{\varepsilon}{1 - \gamma}, \quad \|V - V^*\|_\infty \;\le\; \frac{\|V - BV\|_\infty}{1 - \gamma} \;=\; \frac{\varepsilon}{1 - \gamma}.$$

3. Deriving the Lower Bound on V^π

For any state s,

$$V^\pi(s) \;=\; V(s) \;+\; \big[V^\pi(s) - V(s)\big] \;\geq\; V(s) \;-\; \|V^\pi - V\|_\infty \;\geq\; V(s) \;-\; \frac{\varepsilon}{1-\gamma}.$$

Also,

$$V(s) \;\geq\; V^*(s) \;-\; \|V - V^*\|_\infty \;\geq\; V^*(s) \;-\; \frac{\varepsilon}{1-\gamma}.$$

Combining these two inequalities gives

$$V^\pi(s) \;\geq\; \Big[V^*(s) - \tfrac{\varepsilon}{1-\gamma}\Big] \;-\; \tfrac{\varepsilon}{1-\gamma} \;=\; V^*(s) \;-\; \frac{2\,\varepsilon}{1-\gamma}.$$

Conclusion

Thus for every $s \in S$,

$$\boxed{V^\pi(s) \;\geq\; V^*(s) \;-\; \frac{2\,\varepsilon}{1-\gamma},}$$

showing that the greedy policy from an approximate value function is near-optimal, with suboptimality bounded by $\frac{2\varepsilon}{1-\gamma}$.

7. Give an example real-world application or domain where having a lower bound on $V^\pi(s)$ would be useful.

Example Application: Risk-Averse Portfolio Management

In automated portfolio management, the "state" s encodes current market conditions and portfolio holdings, and the "reward" reflects portfolio return (possibly adjusted for risk). A policy π specifies how to rebalance assets over time. Here a lower bound on the value

$$V^\pi(s) \;=\; \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t) \;\Big|\; s_0 = s, \pi\Big]$$

means we can guarantee that, starting from any market state s, our strategy will achieve at least a minimum expected cumulative return (or utility) even in the worst-case scenarios.

Such a bound is critical for:

- **Regulatory compliance:** Ensuring the investment strategy meets mandated minimum performance or capital-preservation requirements.
- **Risk management:** Providing formal assurances to risk-averse investors that losses will not exceed a prespecified threshold.
- **Automated safeguards:** Triggering conservative fallback policies if the estimated lower bound drops below acceptable levels.

By using the inequality $V^\pi(s) \geq V^*(s) - \frac{2\varepsilon}{1-\gamma}$, we can certify that—even if our value-approximation V has error ε—the deployed policy still achieves a guaranteed performance margin, which is invaluable in high-stakes financial applications.

8. Suppose we have another value function V' and extract its greedy policy π'. $\|BV' - V'\| = \varepsilon = \|BV - V\|$. Does the above lower bound imply that $V^\pi(s) = V^{\pi'}(s)$ at any s?

Does Equal Bellman Residual Imply Equal Greedy-Policy Values?

No. Even if two arbitrary value vectors V and V' have the same Bellman error magnitude

$$\|BV - V\|_\infty \;=\; \|BV' - V'\|_\infty \;=\; \varepsilon,$$

and we extract their greedy policies π and π', the bound

$$V^\pi(s) \;\geq\; V^*(s) \;-\; \frac{2\varepsilon}{1-\gamma} \quad \text{and} \quad V^{\pi'}(s) \;\geq\; V^*(s) \;-\; \frac{2\varepsilon}{1-\gamma}$$

only guarantees that both policies are *near-optimal*, not that they coincide. In particular:

1. The lower bound is one-sided: it constrains how far below V^* each can be, but places no upper bound on differences between V^π and $V^{\pi'}$. 2. Greedy-policy extraction can break ties arbitrarily when the action-value gaps are small. Two different approximate value functions with equal residuals can induce different tie-breaking, leading to distinct π and π'. 3. Consequently, at a given state s, it need not hold that

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

All we know is both $V^\pi(s)$ and $V^{\pi'}(s)$ lie within $\frac{2\varepsilon}{1-\gamma}$ of the true optimum $V^*(s)$, but they may differ from each other.

Say $V \leq V'$ if $\forall s$, $V(s) \leq V'(s)$.
What if our algorithm returns a V that satisfies $V^* \leq V$? I.e., it returns a value function that is better than the optimal value function of the MDP. Once again, remember that V can be any vector, not necessarily achievable in the MDP, but we would still like to bound the performance of V^π where π is extracted from said V. We will show that if this condition is met, then we can achieve an even tighter bound on policy performance.

9. Using the same notation and setup as part 5, if $V^* \leq V$, show the following holds for any state s. *Recall that for all π, $V^\pi \leq V^*$ (why?)*

$$V^\pi(s) \geq V^*(s) - \frac{\varepsilon}{1-\gamma}$$

Tighter Performance Bound When $V^* \leq V$

Let V be any value vector satisfying

$$V^*(s) \;\leq\; V(s) \quad \forall s \in S,$$

and let π be the greedy policy extracted from V. Denote the Bellman error magnitude by $\varepsilon = \|BV - V\|_\infty$. We show that for every state s,

$$\boxed{V^\pi(s) \;\geq\; V^*(s) \;-\; \frac{\varepsilon}{1-\gamma}.}$$

Proof

1. From the greedy definition, as in part (8), we have
$$B^\pi V \;=\; BV,$$
so the policy-evaluation residual satisfies $\|V - B^\pi V\|_\infty = \varepsilon$.

2. By the result of part (e), for any policy π,
$$\|V - V^\pi\|_\infty \;\le\; \frac{\|V - B^\pi V\|_\infty}{1-\gamma} \;=\; \frac{\varepsilon}{1-\gamma}.$$
Hence for each s,
$$V^\pi(s) \;\ge\; V(s) \;-\; \|V - V^\pi\|_\infty \;\ge\; V(s) \;-\; \frac{\varepsilon}{1-\gamma}.$$

3. Since $V^*(s) \le V(s)$, we combine to obtain
$$V^\pi(s) \;\ge\; V(s) \;-\; \frac{\varepsilon}{1-\gamma} \;\ge\; V^*(s) \;-\; \frac{\varepsilon}{1-\gamma}.$$

This completes the proof of the tighter bound when V overestimates the optimal value.

Intuition: A useful way to interpret the results from parts (8) and (9) is based on the observation that a constant immediate reward of r at every time-step leads to an overall discounted reward of
$$r + \gamma r + \gamma^2 r + \cdots = \frac{r}{1-\gamma}$$
Thus, the above results say that a state value function V with Bellman error magnitude ε yields a greedy policy whose reward per step (on average), differs from optimal by at most 2ε. So, if we develop an algorithm that reduces the Bellman residual, we're also able to bound the performance of the policy extracted from the value function outputted by that algorithm, which is very useful!

10. It's not easy to show that the condition $V^* \le V$ holds because we often don't know V^* of the MDP. Show that if $BV \le V$ then $V^* \le V$. Note that this sufficient condition is much easier to check and does not require knowledge of V^*.

    Hint: Try to apply induction. What is $\lim_{n\to\infty} B^n V$?

Sufficient Condition $BV \le V \implies V^* \le V$

We show that if an arbitrary value vector V satisfies the *Bellman dominance* condition
$$(BV)(s) \;\le\; V(s) \quad \forall s \in S,$$
then necessarily
$$V^*(s) \;\le\; V(s) \quad \forall s \in S.$$

1. Monotonicity of B

Recall that the Bellman optimality operator B is monotone:
$$V_1 \le V_2 \implies BV_1 \le BV_2,$$
where all inequalities are elementwise.

2. Iterating B Produces a Decreasing Sequence

Assume $BV \leq V$. Then by monotonicity,

$$B^2V = B(BV) \;\leq\; BV \;\leq\; V,$$

and inductively,

$$B^nV \;\leq\; B^{n-1}V \;\leq\; \cdots \;\leq\; BV \;\leq\; V \quad \forall n \geq 1.$$

Hence the sequence $\{B^nV\}_{n=0}^{\infty}$ (with $B^0V = V$) is nonincreasing and bounded below.

3. Limit of B^nV Is V^*

Since B is a γ–contraction in the sup-norm, it has a unique fixed point V^*, and for any starting vector V,

$$\lim_{n\to\infty} B^nV = V^*.$$

(This follows from the Banach fixed-point theorem or from the convergence proof of Value Iteration.)

4. Conclusion

Because each $B^nV \leq V$ and the limit of B^nV is V^*, taking $n \to \infty$ yields

$$V^* \;=\; \lim_{n\to\infty} B^nV \;\leq\; V.$$

Thus $BV \leq V$ is a sufficient (and easily checkable) condition guaranteeing $V^* \leq V$.

11. (Bonus) It is possible to make the bounds from parts (9) and (10) tighter. Let V be an arbitrary value function and π be the greedy policy extracted from V. Let $\varepsilon = \|BV - V\|$ be the Bellman error magnitude for V. Prove the following for any state s:

$$V^\pi(s) \geq V^*(s) - \frac{2\gamma\varepsilon}{1-\gamma}$$

Further, if $V^* \leq V$, prove for any state s

$$V^\pi(s) \geq V^*(s) - \frac{\gamma\varepsilon}{1-\gamma}$$

Tighter Performance Bounds (Bonus)

Let V be any value vector, π the greedy policy w.r.t. V, and $\varepsilon = \|BV - V\|_\infty$. We prove the refined bounds:

$$V^\pi(s) \;\geq\; V^*(s) \;-\; \frac{2\gamma\,\varepsilon}{1-\gamma}, \qquad \text{and if } V^* \leq V, \quad V^\pi(s) \;\geq\; V^*(s) \;-\; \frac{\gamma\,\varepsilon}{1-\gamma}.$$

Refinement for the General Case

1. **One-step evaluation bound.** Since π is greedy on V, we have

$$(B^\pi V)(s) = (BV)(s), \quad \|V - B^\pi V\|_\infty = \|BV - V\|_\infty = \varepsilon.$$

From the Bellman-evaluation bound (part (e)),

$$\|V^\pi - V\|_\infty \le \frac{\|V - B^\pi V\|_\infty}{1-\gamma} = \frac{\varepsilon}{1-\gamma}.$$

Hence

$$V^\pi(s) \;\ge\; (B^\pi V)(s) \;-\; \gamma\,\|V^\pi - V\|_\infty = (BV)(s) \;-\; \frac{\gamma\,\varepsilon}{1-\gamma}.$$

2. **Relate BV to V^*.** By definition of the Bellman operator, $BV \ge BV^* - \|BV - BV^*\|$, but $BV^* = V^*$ and $\|BV - BV^*\| \le \varepsilon$. Thus

$$(BV)(s) \;\ge\; V^*(s) \;-\; \varepsilon.$$

Combining,

$$V^\pi(s) \;\ge\; V^*(s) \;-\; \varepsilon \;-\; \frac{\gamma\,\varepsilon}{1-\gamma} \;=\; V^*(s) \;-\; \frac{\varepsilon(1-\gamma) + \gamma\,\varepsilon}{1-\gamma} \;=\; V^*(s) \;-\; \frac{2\gamma\,\varepsilon}{1-\gamma}.$$

Tighter Bound When $V^* \le V$

If in addition $V^*(s) \le V(s)$ for all s, then instead of $BV(s) \ge V^*(s) - \varepsilon$ we have

$$(BV)(s) \;\ge\; V(s) \;-\; \varepsilon \;\ge\; V^*(s) \;-\; \varepsilon.$$

Repeating the same steps as above gives

$$V^\pi(s) \;\ge\; (BV)(s) \;-\; \frac{\gamma\,\varepsilon}{1-\gamma} \;\ge\; V^*(s) \;-\; \varepsilon \;-\; \frac{\gamma\,\varepsilon}{1-\gamma} \;=\; V^*(s) \;-\; \frac{\gamma\,\varepsilon}{1-\gamma}.$$

This completes the proof of the tighter performance guarantees.

References

[1] Baesed on CS 234: Reinforcement Learning, Stanford University. Spring 2024.

[2] Cover image designed by freepik