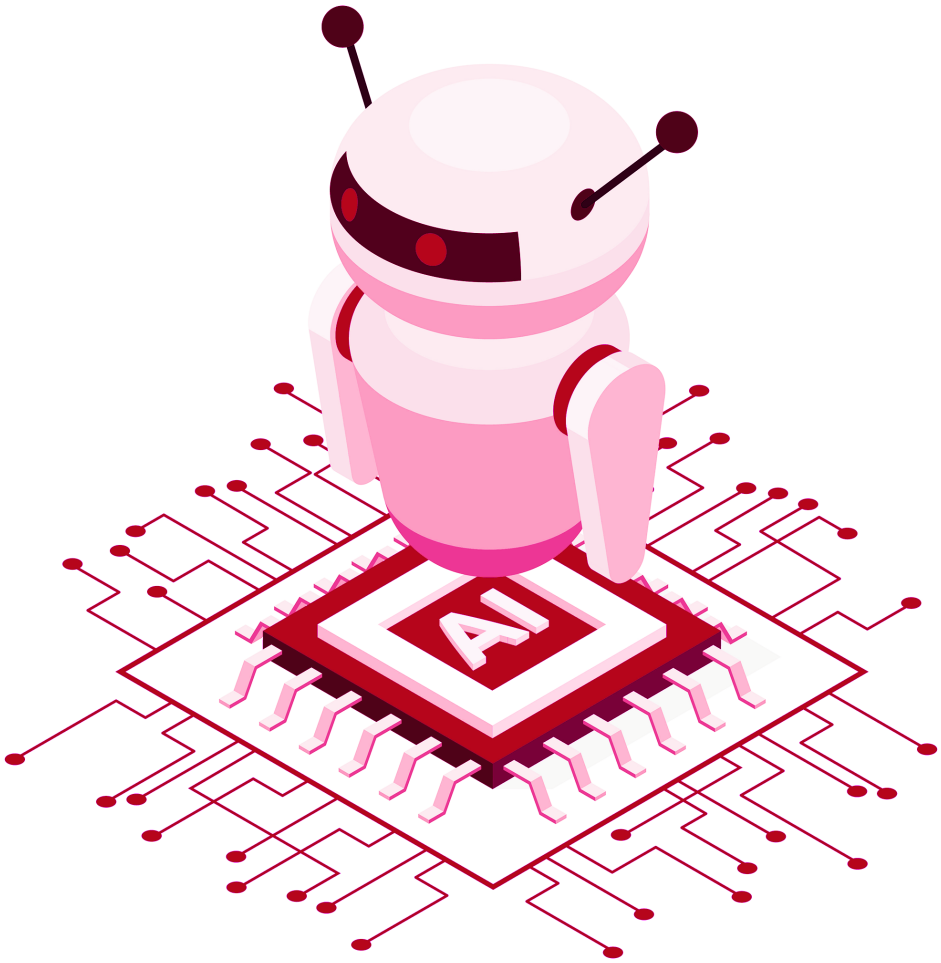
pgf@stop



Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 7:

Value-Based Theory

By:

Amir Kooshan Fattah Hesari

401102191

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^*.
   Answer : Because we have assumed that $R(s,a) \geq 0$ for all s, a, for deriving an upper bound we can presume that in every step we get a reward of value R_{max}. Hence the discounted reward sum (return) is going to be :

$$R(s,a) \leq R_{max} \xrightarrow{for\ upper\ bound} R(s,a) = R_{max} \quad \forall (s,a)$$
$$V_k^*(s) = max_\pi \mathbf{E}_\pi \left[\sum_{t=0}^{k-1} \gamma^t \cdot R(s_t, a_t) \middle| s_0 = s \right] \leq \sum_{t=0}^{k-1} \gamma^t \cdot R_{max}$$
$$\Rightarrow V_k^*(s) \leq R_{max} \cdot \sum_{t=0}^{k-1} \gamma^t = R_{max} \cdot \frac{1-\gamma^k}{1-\gamma} = \frac{R_{max}}{1-\gamma} \quad \forall s \in S$$

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.
   Answer : Because we have assumed that $R(s,a) \geq 0 \quad \forall (s,a)$ and assuming that $V_k^\pi = V_k^*$ we have :

$$V_{k+1}^\pi = V_k^\pi + \gamma^k \cdot R(s,a) \geq V_k^\pi = V_k^*$$
$$\Rightarrow V_{k+1}^\pi \geq V_k^*$$

   And because $\{V_k^*\}_{k=1}^\infty$ is non-decreasing (because we have assumed that all $R(s,a) \geq 0$) and we derived an upper bound in the last part , hence the Value Iteration method converges to a solution satisfying yhe Bellman equation.

3. By taking the limit in the Bellman equation, prove that the V^* is optimal.
   Answer : By taking the limit $k \to \infty$ we have :

$$\lim_{k\to\infty} V_{k+1}^*(s) = \lim_{k\to\infty} max_a \sum_{s' \in S} p(s'|s,a) \cdot \left[R(s,a) + \gamma \cdot V_k^*(s') \right]$$
$$V^*(s) = max_a \sum_{s' \in S} p(s'|s,a) \cdot \left[R(s,a) + \gamma \cdot \lim_{k\to\infty} V_k^*(s') \right]$$
$$V^*(s) = max_a \sum_{s' \in S} p(s'|s,a) \cdot \left[R(s,a) + \gamma \cdot V_k^*(s') \right]$$

The reason that $\lim_{k\to\infty} V_k^*(s') = V_k^*(s')$ is that we can write the above equation in the form of Bellman optimality operator $\mathcal{T}_M : \mathbf{R}^{|s||a|} \to \mathbf{R}^{|s||a|}$ in the following format :

$$\mathcal{T}Q := r + \gamma P V_Q$$

and the preceding bellman operator is a **contraction mapping** and it has a unique fixed point and when applied iteratively (in every state we apply the $\mathcal{T}_M$ and try to find $\mathcal{T}V = V$) it converges to the unique fixed point.
Moreover, because the V^* satisfies the bellman optimality equation and it has a unique fixed point we conclude that it converges to the optimal solutions as $k \to \infty$

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^*.
   Answer : From last part we know that

$$\hat{V}_k^* = max_\pi \mathbf{E}_\pi \left[\sum_{t=0}^{k-1} \gamma^t \cdot \hat{R}(s,a) \middle| s_0 = s \right] = max_\pi \mathbf{E}_\pi \left[\sum_{t=0}^{k-1} \gamma^t \cdot r_o + \gamma^t \cdot R(s,a) \middle| s_0 = s \right]$$

$$\hat{V}_k^* = \sum_{t=0}^{k-1} \gamma^t \cdot r_0 + max_\pi \mathbf{E}_\pi \left[\sum_{t=0}^{k-1} \gamma^t \cdot R(s,a) \middle| s_0 = s \right] = r_0 \cdot \frac{1-\gamma^k}{1-\gamma} + V_k^*$$

   Now , for $K = 0$ (the base case) we have $\hat{V}_0^*(s) = V_0^*(s) = 0$. With this base :

$$\begin{aligned}
\hat{V}_k^* &= max_a \sum_{s' \in S} p(s'|s,a) \left[\hat{R}(s,a) + \gamma \cdot \hat{V}_{k-1}^* \right] \\
&= max_a \sum_{s' \in S} p(s'|s,a) \left[R(s,a) + r_0 + \gamma \cdot \left(V_{k-1}^* + r_0 \cdot \frac{1-\gamma^{(k-1)}}{1-\gamma} \right) \right] \\
\hat{V}_k^* &= r_0 + r_0 \cdot \frac{\gamma(1-\gamma^{k-1})}{1-\gamma} + max_a \sum_{s' \in S} p(s'|s,a) \left[R(s,a) + \gamma \cdot V_{k-1}^* \right] \\
&= r_0 \left(\frac{1-\gamma+\gamma-\gamma^k}{1-\gamma} \right) + V_k^* = r_0 \cdot \frac{1-\gamma^K}{1-\gamma} + V_k^* \\
\lim_{k\to\infty} \hat{V}_k^* &= \lim_{k\to\infty} V_k^* + \frac{r_0}{1-\gamma} \Rightarrow \hat{V}_k^* = V^* + \frac{r_0}{1-\gamma}
\end{aligned}$$

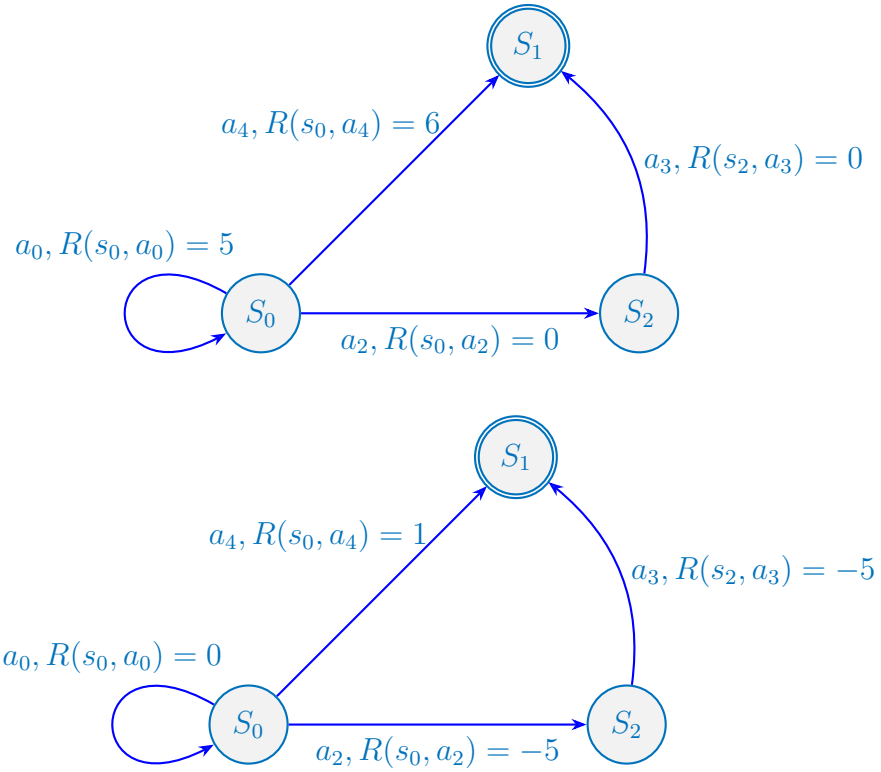
   We see that in the limit the new MDP still converges the optimal value function with a added constant of $\frac{r_0}{1-\gamma}$.

5. Why is it necessary to assume the absence of a terminating state? Try to explain with a counterexample.
   when a terminating state is reached, the process ends. Any reward added to the transition leading to the terminal state is a one-time benefit. The value of the terminal state itself is not affected by

the added reward because no further actions or rewards occur after reaching it.
The constant shift in rewards doesn't propagate consistently to the value of terminating states. The value of a terminating state remains fixed at its defined terminal value (usually 0), while the values of non-terminating states are shifted by a factor related to $\frac{r_0}{1-\gamma}$.
This difference in how the added reward affects terminating and non-terminating states breaks the uniform relationship

$$\hat{V}^*(s) = V^*(s) + \frac{r_0}{1-\gamma}$$

across all states.
Consider this scenario :



In the above exmaple we see that in the lower diagram the best course of action is to take the a_4 action and reach the terminal state S_1 and get a reward of $+1$.
But in the above diagaram the best course of action is to take action a_0 infinitely many times in order to get the maximum reward which is $5 \cdot \frac{1}{1-\gamma}$.

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.

Answer : Let π_k be the current policy at iteration k, with its value function denoted by V^{π_k}. Now, construct a new policy π_{k+1} by acting greedily with respect to V^{π_k}:

$$\pi_{k+1}(s) = \arg\max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V^{\pi_k}(s')\right]$$

In other words, π_{k+1} chooses the action that looks best if we trust the value estimates from π_k. We want to prove that this new policy gives at least as high a value as the old one in every state:

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

and that the inequality is strict in at least one state unless π_k is already optimal.

Define the action-value function under π_k:

$$Q^{\pi_k}(s,a) = \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V^{\pi_k}(s')\right]$$

By how we defined π_{k+1}, we have:

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

This means the new action is at least as good as the old one, according to the old value estimates.

Let T^π be the Bellman operator for policy π:

$$(T^\pi V)(s) = \sum_{s'} P(s'|s,\pi(s))\left[R(s,\pi(s),s') + \gamma V(s')\right]$$

This operator pulls any value function closer to the true value function of π. Since:

$$V^{\pi_k} = T^{\pi_k} V^{\pi_k}, \quad \text{and} \quad T^{\pi_{k+1}} V^{\pi_k} \geq T^{\pi_k} V^{\pi_k}$$

we get:

$$T^{\pi_{k+1}} V^{\pi_k} \geq V^{\pi_k}$$

Now, repeatedly applying $T^{\pi_{k+1}}$ (i.e., iterating it), we converge to the true value of π_{k+1}:

$$V^{\pi_{k+1}} = \lim_{n\to\infty} (T^{\pi_{k+1}})^n V^{\pi_k} \geq V^{\pi_k}$$

This shows that π_{k+1} performs at least as well as π_k in all states.

If π_k is not optimal, then there exists at least one state s where the greedy action improves on the current one:

$$Q^{\pi_k}(s, \pi_{k+1}(s)) > V^{\pi_k}(s) \Rightarrow V^{\pi_{k+1}}(s) > V^{\pi_k}(s)$$

So the new policy is strictly better in at least one state unless π_k was already the best possible.

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.
Answer : Since the MDP is finite, both the state space $|S|$ and the action space $|A|$ are finite. Consequently, the total number of deterministic policies is finite, at most $|A|^{|S|}$. From the previous result, we know that each policy improvement strictly increases the value function unless the policy is already optimal.

   This means a suboptimal policy cannot be selected more than once during the policy iteration process. If some suboptimal policy were chosen again, it would contradict the strict improvement property—implying it was optimal, which it is not. Hence, due to the finite number of possible policies and strict improvement at each step, policy iteration must eventually reach a policy π_k such that:

$$\pi_{k+1} = \pi_k$$

   Now, suppose for the sake of contradiction that this policy π_k is not optimal, i.e., its value function V^{π_k} does not satisfy the Bellman optimality equation. Then there exists at least one state $s \in S$ such that:

$$\max_a \sum_{s'} P(s'|s,a) \left[R(s,a) + \gamma V^{\pi_k}(s')\right] > \sum_{s'} P(s'|s,\pi_k(s)) \left[R(s,\pi_k(s)) + \gamma V^{\pi_k}(s')\right] = V^{\pi_k}(s)$$

   This inequality implies that choosing action $\pi_k(s)$ is not optimal in state s, and thus we could perform another policy improvement step. That would yield $\pi_{k+1} \neq \pi_k$, which contradicts our assumption that the algorithm has terminated.

   Conclusion: The contradiction implies that V^{π_k} satisfies the Bellman optimality equation. Therefore, the final policy returned by policy iteration is indeed optimal.

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?
Answer : Both value iteration and policy iteration are guaranteed to converge to the same optimal value function V^* in a finite MDP. Although the algorithms operate differently, they both rely on the Bellman optimality operator, which has a unique fixed point.

   In value iteration, we iteratively apply the Bellman optimality update:

$$V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a) \left[R(s,a) + \gamma V_k(s')\right]$$

   This update defines a contraction mapping, since the Bellman operator T satisfies:

$$\|TV - TW\|_\infty \leq \gamma \|V - W\|_\infty, \quad \text{for any } V, W$$

Therefore, by Banach's Fixed Point Theorem, the sequence $\{V_k\}$ converges to a unique fixed point V^* such that:

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

On the other hand, policy iteration alternates between evaluating a policy and improving it. After a finite number of steps, it reaches a policy π^* for which:

$$V^{\pi^*}(s) = \sum_{s'} P(s'|s,\pi^*(s))\,\left[R(s,\pi^*(s)) + \gamma V^{\pi^*}(s')\right]$$

and

$$\pi^*(s) = \arg\max_a \sum_{s'} P(s'|s,a)\,\left[R(s,a) + \gamma V^{\pi^*}(s')\right]$$

This means V^{π^*} also satisfies the Bellman optimality equation, and thus $V^{\pi^*} = V^*$.

Now, even though both methods converge to the same value function V^*, the optimal policies they produce may differ. This happens when there are multiple actions that achieve the same maximum value in a state. For example, if two actions a_1 and a_2 both satisfy:

$$Q^*(s,a_1) = Q^*(s,a_2) = V^*(s),$$

then choosing either action is valid. As a result, different greedy choices can lead to different policies, but all of them are optimal because they yield the same value function V^*.

Therefore, value iteration and policy iteration both converge to V^*, and although the resulting policies may differ, they are all optimal since they are greedy with respect to the same V^*.

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.
Answer : In policy iteration, the evaluation step involves solving a linear system derived from the Bellman expectation equation:

$$V^{\pi_k}(s) = \sum_{s'} P(s'|s,\pi_k(s))\,[R(s,\pi_k(s)) + \gamma V^{\pi_k}(s')] \quad \forall s \in S$$

This results in $|S|$ linear equations with $|S|$ unknowns, and solving such a system typically requires $\mathcal{O}(|S|^3)$ time using standard matrix inversion or linear solvers.

The improvement step updates the policy by selecting the action that maximizes expected return:

$$\pi_{k+1}(s) = \arg\max_{a\in A} \sum_{s'} P(s'|s,a)\,[R(s,a) + \gamma V^{\pi_k}(s')]$$

This maximization needs to be performed for each state. Evaluating the expected return for each action involves a sum over $|S|$ states, and since there are $|A|$ actions, the cost per state is $\mathcal{O}(|S||A|)$. Across all states, the total cost for this step becomes $\mathcal{O}(|S|^2|A|)$.

Combining both steps, the total time complexity for a single iteration of policy iteration is:

$$\mathcal{O}(|S|^3 + |S|^2|A|)$$

Now, consider value iteration. The update for each state is given by:

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

This operation, similar to the improvement step above, also takes $\mathcal{O}(|S||A|)$ time per state, leading to a full iteration complexity of $\mathcal{O}(|S|^2|A|)$.

Although value iteration requires fewer computations per iteration than policy iteration, it generally needs more iterations to converge. On the other hand, policy iteration often finishes in fewer iterations due to more substantial updates at each step, despite its higher per-iteration cost. In practice, this trade-off depends on the specifics of the environment and desired accuracy.

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.
Answer : When the discount factor $\gamma = 1$, the infinite-horizon MDP becomes undiscounted, meaning future rewards are valued equally with immediate rewards. In this case, the Bellman optimality equation no longer ensures contraction:

$$TV(s) = \max_a \sum_{s'} P(s'|s,a)\,[R(s,a) + V(s')]$$

Since the Bellman operator T is not a contraction for $\gamma = 1$, value iteration is not guaranteed to converge. Instead, it may oscillate or diverge unless additional assumptions like bounded expected return or the presence of absorbing states are imposed. In practice, value iteration is often unstable for $\gamma = 1$ without careful control.

Policy iteration, on the other hand, can still converge in the undiscounted setting if the MDP satisfies special conditions such as a proper policy or finite expected return. The evaluation step becomes solving a linear system of the form:

$$V^\pi(s) = \sum_{s'} P(s'|s,\pi(s))\,[R(s,\pi(s)) + V^\pi(s')]$$

Under suitable assumptions, this system has a well-defined solution, and the improvement step remains valid. However, care must be taken since policy iteration may cycle or fail to terminate if the assumptions do not hold. Overall, $\gamma = 1$ makes both methods more fragile and problem-dependent.

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 will assume that π is deterministic, but it does not 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.
   Answer : Let $V,\ V' : S \to \mathbb{R}$ be two arbitrary bounded-value functions and fix a deterministic policy π. For every state $s \in S$ we have

$$\begin{aligned} \left|B^\pi V(s) - B^\pi V'(s)\right| &= \Big| r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V(s') \\ &\quad - \Big[r(s,\pi(s)) + \gamma \sum_{s'} p(s'|s,\pi(s)) V'(s') \Big] \Big| \\ &= \gamma \Big| \sum_{s'} p(s'|s,\pi(s)) \big[V(s') - V'(s') \big] \Big| \end{aligned}$$

   Because the transition probabilities form a distribution, Jensen's (or simply the triangle) inequality gives

$$\begin{aligned} \Big| \sum_{s'} p(s'|s,\pi(s)) \big[V(s') - V'(s') \big] \Big| &\leq \sum_{s'} p(s'|s,\pi(s)) \big| V(s') - V'(s') \big| \\ &\leq \max_{s''} \big| V(s'') - V'(s'') \big| = \|V - V'\|_\infty \end{aligned}$$

Hence, for each state s,

$$\left|B^\pi V(s) - B^\pi V'(s)\right| \le \gamma \|V - V'\|_\infty.$$

Taking the supremum over all $s \in S$ yields the desired contraction bound

$$\boxed{\|B^\pi V - B^\pi V'\|_\infty \le \gamma \, \|V - V'\|_\infty}.$$

Since $0 \le \gamma < 1$, the operator B^π is a γ-contraction under the infinity norm, guaranteeing a unique fixed point V^π via Banach's fixed-point theorem.

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.
Answer : Assume V and W are two fixed points of the Bellman operator B^π; that is,

$$V = B^\pi V, \qquad W = B^\pi W.$$

Now we have

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

From the previous result, B^π is a γ-contraction under the infinity norm with $0 \le \gamma < 1$, so

$$\|B^\pi V - B^\pi W\|_\infty \le \gamma \, \|V - W\|_\infty.$$

Combining the two displays gives

$$\|V - W\|_\infty \le \gamma \, \|V - W\|_\infty.$$

After subtracting $\gamma\|V - W\|_\infty$ from both sides we obtain

$$(1 - \gamma) \, \|V - W\|_\infty \le 0.$$

Because $1 - \gamma > 0$, the only possibility is $\|V - W\|_\infty = 0$, which implies $V = W$ for every state. Therefore 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.*
Answer :

$$\begin{aligned} B^\pi V(s) &= r\big(s, \pi(s)\big) + \gamma \sum_{s'} p\big(s' \mid s, \pi(s)\big) \, V(s') \\ &\le r\big(s, \pi(s)\big) + \gamma \sum_{s'} p\big(s' \mid s, \pi(s)\big) \, V'(s') \\ &= B^\pi V'(s) \\ &\Rightarrow B^\pi V(s) \le B^\pi V'(s) \end{aligned}$$

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?
   Answer : The error vanishes precisely when $BV = V$. Because the Bellman optimality equation has a unique solution, the error is zero whenever V equals the optimal value function V^*.

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

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

Answer : Add and subtract $T^\pi V$:

$$V^\pi - V = \big(T^\pi V^\pi - T^\pi V\big) + \big(T^\pi V - V\big) = \gamma P^\pi (V^\pi - V) + \big(T^\pi V - V\big),$$

where P^π is the state–transition matrix under π. Taking the infinity norm and using $\|P^\pi\|_\infty = 1$,

$$\|V^\pi - V\|_\infty \leq \gamma \|V^\pi - V\|_\infty + \|T^\pi V - V\|_\infty.$$

Rearranging gives

$$(1 - \gamma)\,\|V^\pi - V\|_\infty \leq \|V - T^\pi V\|_\infty, \qquad \Longrightarrow \qquad \boxed{\|V - V^\pi\|_\infty \leq \frac{\|V - B^\pi V\|_\infty}{1 - \gamma}}$$

Repeat the same argument with the optimal operator T:

$$V^* - V = (TV^* - TV) + (TV - V) = \gamma P^{\pi_V} (V^* - V) + (TV - V),$$

where π_V is any greedy policy w.r.t. V (so $TV = T^{\pi_V} V$). Again,

$$(1 - \gamma)\,\|V^* - V\|_\infty \leq \|V - TV\|_\infty, \qquad \Longrightarrow \qquad \boxed{\|V - V^*\|_\infty \leq \frac{\|V - BV\|_\infty}{1 - \gamma}}$$

Thus, for any value function V and any policy π, the distance to V^π or V^* is bounded by the corresponding Bellman residual scaled by $1/(1 - \gamma)$.

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.

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

   Answer : First observe that

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

Because the policy π is defined to choose the maximizing action in the inner expression, we also have

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

From the error bounds established earlier,

$$\frac{\varepsilon}{1-\gamma} = \frac{\|BV - V\|_\infty}{1-\gamma} \geq \|V^* - V\|_\infty \quad \text{and} \quad \frac{\varepsilon}{1-\gamma} = \frac{\|B^\pi V - V\|_\infty}{1-\gamma} \geq \|V - V^\pi\|_\infty.$$

Hence, for every state s,

$$V^*(s) - V^\pi(s) \leq |V^*(s) - V(s)| + |V(s) - V^\pi(s)| \leq \frac{2\varepsilon}{1-\gamma}.$$

Therefore,

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

which is exactly the desired result.

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

   Answer : A clear example arises in *robotic surgery*. Suppose the surgeon's assistive robot follows a policy π whose value function $V^\pi(s)$ measures the expected probability of task success minus a weighted penalty for tissue damage, starting from state s (e.g. current tool pose and patient anatomy). Before executing any motion, clinicians want a *provable lower bound* $\underline{V}(s) \leq V^\pi(s)$ to certify that—even in the worst case consistent with model uncertainty—the procedure's safety-adjusted success rate will exceed an acceptable threshold. Such a bound gives a formal guarantee that the robot will not enter states where patient risk outweighs potential benefit, enabling online vetoes or hand-offs to a human surgeon whenever $\underline{V}(s)$ falls below the safety margin.

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?

   Answer : No, that bound does not imply that $V^\pi = V^{\pi'}$. It simply guarantees that both V^π and $V^{\pi'}$ are at most a distance of $\frac{2\varepsilon}{1-\gamma}$ from the optimal value function V^*, in terms of the infinity norm. In other words, they are both close to V^*, but this does not mean they are necessarily close to each other, nor does it force them to be equal. There can still be a nonzero gap between V^π and $V^{\pi'}$ as long as each remains within the specified tolerance of the optimum.

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?)*

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

Answer : Suppose we are given a value function V such that $V \geq V^* \geq V^\pi$, which may arise from an approximation or intermediate estimate during an iterative algorithm. Then we can reuse the residual-based error bound from part 6:

$$\frac{\varepsilon}{1-\gamma} = \frac{\|B^\pi V - V\|_\infty}{1-\gamma} \geq \|V - V^\pi\|_\infty \geq V(s) - V^\pi(s) \geq V^*(s) - V^\pi(s)$$

for any $s \in S$. The first inequality uses the standard bound on the distance to the true value of policy π in terms of the Bellman residual, while the second and third follow directly from the assumed ordering $V \geq V^* \geq V^\pi$.

Therefore, we conclude:

$$\boxed{\frac{\varepsilon}{1-\gamma} \geq V^*(s) - V^\pi(s) \quad \forall s \in S}$$

which establishes the desired bound on the suboptimality of policy π in terms of the Bellman error of an over-approximating value function V.

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

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

Because B is monotone—that is, $U \leq W \;\Rightarrow\; BU \leq BW$—we can apply induction: $B^2V = B(BV) \leq BV \leq V$, $B^3V \leq B^2V$, and so on. Thus the sequence $\{B^nV\}_{n\geq 0}$ is monotonically non-increasing and bounded below. Contraction of B in the ℓ_∞ norm implies

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

where V^* is the unique fixed point of B (the optimal value function). Since each B^nV is dominated by V, the limit is as well: $V^* \leq V$ component-wise. Hence, the condition $BV \leq V$ is sufficient to guarantee that V is an *upper bound* on the optimal value function V^* without requiring prior knowledge of 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

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

Answer : Because $B^{\pi}V^{\pi} = V^{\pi}$ and $B^{\pi}V^{\pi} = BV^{\pi}$, the one–step contraction of B^{π} yields

$$\|B^{\pi}V - V^{\pi}\|_{\infty} \;\leq\; \gamma\,\|V - V^{\pi}\|_{\infty} \;\leq\; \frac{\gamma\varepsilon}{1-\gamma}. \tag{1}$$

Next set $V' := BV$. The residual bound $\|BV - V\|_{\infty} = \varepsilon$ implies

$$\frac{\varepsilon}{1-\gamma} \;\geq\; \|V - V'\|_{\infty} \;\geq\; \|V^{*} - V'\|_{\infty},$$

while the γ-contraction of B itself gives

$$\|V^{*} - V'\|_{\infty} \;\leq\; \frac{\gamma\varepsilon}{1-\gamma}. \tag{2}$$

Combining (1) and (2) we obtain

$$\|V^{*} - V^{\pi}\|_{\infty} \;\leq\; \|V^{*} - V'\|_{\infty} + \|V' - V^{\pi}\|_{\infty} \;\leq\; \frac{2\gamma\varepsilon}{1-\gamma},$$

and hence, for every state s,

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

For the second claim we reuse (1):

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

Since $V \geq V^{*}$ gives $BV \geq BV^{*} = V^{*}$, we deduce $BV^{\pi} \geq V^{*}$. Substituting this into the last term yields

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

which is the desired second inequality.

References

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

[2] Cover image designed by freepik