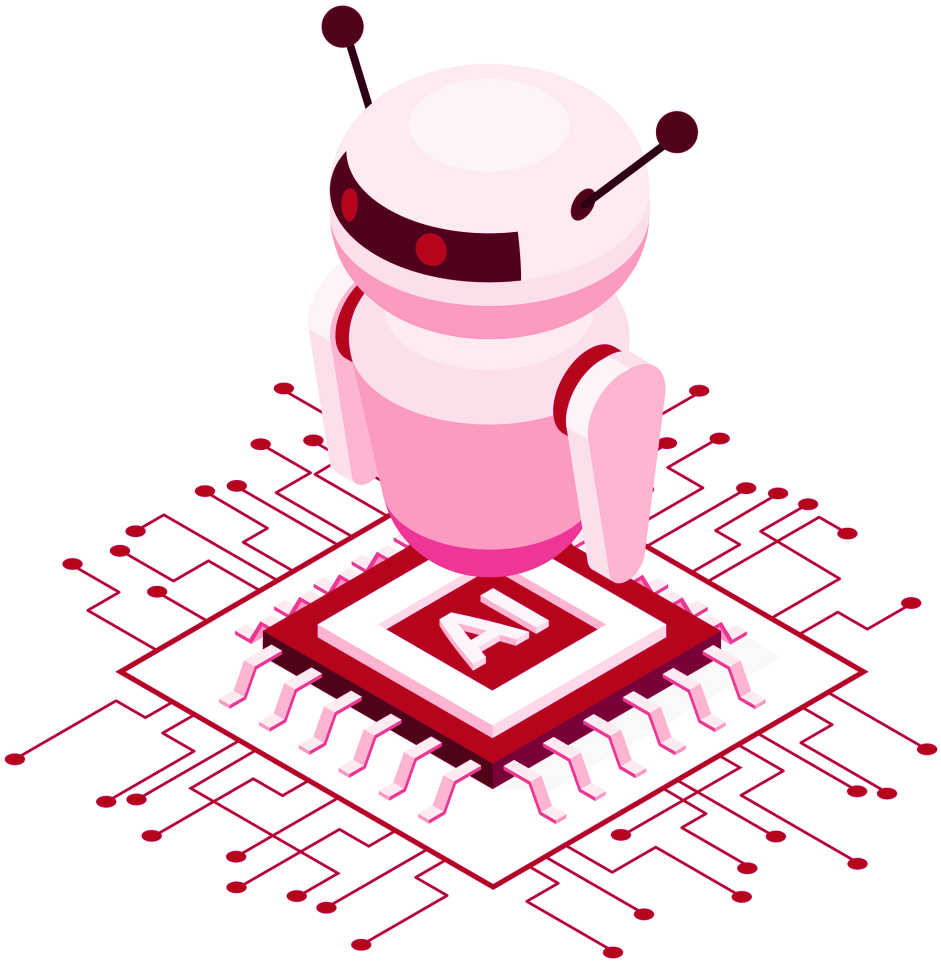


Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 7:

Value-Based Theory

By:

Ayeen Poostforoushan

401105742

Spring 2025

Contents

1 Iteration Family **1**

1.1 Positive Rewards 1

1.2 General Rewards 2

1.3 Policy Turn 3

2 Bellman or Bellwoman **6**

2.1 Bellman Operators 6

2.2 Bellman Residuals 7

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^*.

   As we know, the k-step optimal value function is the summed expected discounted rewards over the next k steps:

   $$V_k^*(s) \;=\; \max_{\pi} \mathbb{E}\Big[\sum_{t=0}^{k-1} \gamma^t R_{t+1} \;\Big|\; S_0 = s\Big].$$

   Using $R(s,a) \leq R_{\max}$, for any policy π we get

   $$\sum_{t=0}^{k-1} \gamma^t R_{t+1} \;\leq\; \sum_{t=0}^{k-1} \gamma^t R_{\max} = \; R_{\max} \frac{1-\gamma^k}{1-\gamma}.$$

   Therefore,

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

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.

   Let π_k be any policy that gets the maximum in the definition of V_k^*:

   $$V_k^*(s) = \mathbb{E}\Big[\sum_{t=0}^{k-1} \gamma^t R_{t+1} \;\Big|\; S_0 = s, \pi_k\Big].$$

   Construct a policy $\tilde{\pi}$ for the $(k+1)$-step problem by

   $$\tilde{\pi}(s_0, \ldots, s_k) = \begin{cases} \pi_k(s_0, \ldots, s_{k-1}) & \text{for } t < k, \\ \text{any action} & \text{at } t = k. \end{cases}$$

   Then under $\tilde{\pi}$,

   $$V_{k+1}^{\tilde{\pi}}(s) = \mathbb{E}\Big[\sum_{t=0}^{k-1} \gamma^t R_{t+1} + \gamma^k R_{k+1} \;\Big|\; S_0 = s, \tilde{\pi}\Big].$$

   Since $R_{k+1} \geq 0$, we have

   $$\sum_{t=0}^{k-1} \gamma^t R_{t+1} + \gamma^k R_{k+1} \;\geq\; \sum_{t=0}^{k-1} \gamma^t R_{t+1},$$

so

$$V_{k+1}^{\tilde{\pi}}(s) \;\geq\; \mathbb{E}\Big[\sum_{t=0}^{k-1}\gamma^t R_{t+1} \;\Big|\; S_0 = s, \pi_k\Big] = V_k^*(s).$$

Finally, since $V_{k+1}^*(s)$ is the maximum over all policies,

$$V_{k+1}^*(s) \;\geq\; V_{k+1}^{\tilde{\pi}}(s) \;\geq\; V_k^*(s),$$

proving V_k^* is non–decreasing in k. Also, from part (1) we know V_k^* is bounded above by $\frac{R_{\max}}{1-\gamma}$, so the increasing bounded sequence converges to a fixed point of the bellman optimality equation.

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

   From the definition of the bellman equation for V_{k+1}^* we have

   $$V_{k+1}^*(s) = \max_a \Big[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V_k^*(s')\Big].$$

   Take $k \to \infty$ on both sides. The max and the finite sum are continuous, so

   $$V_\infty(s) = \max_a \Big[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V_\infty(s')\Big].$$

   This is exactly the Bellman optimality equation. So the limit of the k-step values is the optimal value function.

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^*.

   It seems like for the $\hat{V}_k$, we are summing the extra r_0 for all the k steps so the geometric sum of r_0 should be added to the normal V_k. We prove this by induction on k.

   Base case ($k = 0$).

   $$\hat{V}_0^*(s) = 0 = V_0^*(s) + r_0 \sum_{t=0}^{-1} \gamma^t.$$

   Inductive step. Assume for some $k \geq 0$,

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

   Then

   $$\hat{V}_{k+1}^*(s) = \max_a \mathbb{E}\big[\hat{R}(s,a) + \gamma\, \hat{V}_k^*(S')\big] = \max_a \mathbb{E}\Big[R(s,a) + r_0 + \gamma\big(V_k^*(S') + r_0 \frac{1-\gamma^k}{1-\gamma}\big)\Big].$$

Split off the r_0 terms and use the inductive hypothesis:

$$\hat{V}^*_{k+1}(s) = \max_a \mathbb{E}[R(s,a) + \gamma V^*_k(S')] + r_0\Big(1 + \gamma\frac{1-\gamma^k}{1-\gamma}\Big) = V^*_{k+1}(s) + r_0\frac{1-\gamma^{k+1}}{1-\gamma}.$$

Also note that the optimal policy that derives that $\hat{V}_k$ is the same as V_k because the $\max$ is still taken over the original MDP value function. So the optimal policy is the same too.

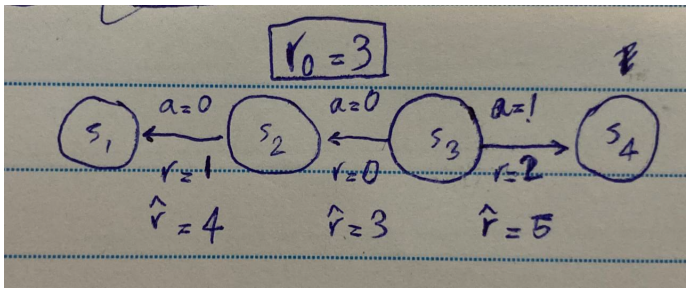
Finally, for the limit,

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

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

   It is necessary because the arguments in the last part hold only if assuming that the constant r0 value is added to all the k steps. If there is a terminating step, there will be no more r0 added for the next steps until k, and the proof is invalid.

   For example in this scenario where in the original MDP. The optimal value function for S2 is 2 by choosing a=1. But in the new MDP, the optimal value function for S2 is obtained by selecting a=0. Because it gives the value of $4 * 0.9 + 3 = 6.6 > 5$ comparing to a=1. So the optimal policy changed.



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 for the root.

   We prove this using the backup diagram way. Assume we currently have policy π_k. Start at an arbitrary root node s. In the root node, replace the π_k-action with the greedy action defining $\pi_{k+1}(s)$, but retain the old policy π_k and its value function V^{π_k} from the second layer onward. Surely $V(s)$ does not decrease, since we are taking

$$\max_a\big[r(s,a) + \gamma\sum_{s'} P(s'|s,a)\,V^{\pi_k}(s')\big] \;\geq\; \sum_a \pi_k(a \mid s)\big[r(s,a) + \gamma\sum_{s'} P(s'|s,a)\,V^{\pi_k}(s')\big],$$

and a maximum is at least as large as the convex combination under π_k.

Next, move to the second layer; for each successor state s' reached under the transition from the root, again replace $\pi_k(s')$ by the greedy action $\pi_{k+1}(s')$, keeping π_k and V^{π_k} thereafter. The same argument shows $V^{\pi_k}(s')$ does not decrease at any s'. Since the root's value is a convex combination of these increased $V(s')$ values under the distribution of the transition function, the root's value increases again.

Continue this procedure through all layers until infinity. At each state, replacing π_k by its greedy action cannot decrease the backed-up value. In the limit, we have fully implemented policy π_{k+1} and obtain $V^{\pi_{k+1}}$ which satisfies

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

If at every state the inequality is actually equality, then no greedy change was possible anywhere, so π_k already satisfies the Bellman optimality equation and is optimal.

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.

   The total number of deterministic policies is finite,which is $|A|^{|S|}$. In each policy improvement step, we showed

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

   with strict inequality for at least one state unless π_k already satisfies the Bellman optimality equation. Hence each improvement produces a policy that is strictly better in value at some state, so it cannot cycle back to any previous policy. Since there are only finitely many policies, this process must terminate after at most $|A|^{|S|}$ iterations.

   When it terminates at π^*, no further improvement is possible, so

   $$\pi^*(s) = \arg\max_a \Big[r(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^{\pi^*}(s') \Big] \quad \forall s.$$

   That means the value V^{π^*} satisfies

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

   so V^{π^*} is a fixed point of the Bellman optimality operator. We know that this fixed point is unique and equals V^*. Therefore π^* is an optimal policy and Policy Iteration converges to it.

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?

   As we know, the Bellman optimality operator B has a unique fixed point V^*, i.e.

   $$B\,V^* = V^*.$$

   Value Iteration repeatedly applies $V_{k+1} = B\,V_k$, so by contraction it converges to that unique fixed point V^*.

Policy Iteration produces a final policy π^* whose value function satisfies

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

so V^{π^*} is also a fixed point of B. Uniqueness of the fixed point forces $V^{\pi^*} = V^*$.

Therefore, both algorithms converge to the same V^*. Different policies may achieve V^*, but any policy whose value satisfies the Bellman optimality equation is optimal.

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.

   Let $|S| = S$ and $|A| = A$.

   - Value Iteration: for each state (there are S) we loop over every action (A) and sum over all successor states (S), giving
   $$O\big(A \cdot S \cdot S\big) = O\big(A\, S^2\big).$$
   - Policy Iteration:
     (a) Policy Evaluation: each sweep over all S states requires summing over all S successors, so one sweep costs $O(S^2)$. To converge we need N_{iter} such sweeps, for
     $$O\big(S^2\, N_{\text{iter}}\big).$$
     (b) Policy Improvement: for each of the S states we evaluate each of the A actions by summing over S successors, giving
     $$O\big(A \cdot S \cdot S\big) = O\big(A\, S^2\big).$$
     Therefore one full Policy Iteration step costs
     $$O\big(S^2\, N_{\text{iter}} + A\, S^2\big).$$

   Value Iteration is often preferred when model dynamics are simple or when approximate backups are used (e.g., with function approximation), since each iteration is relatively cheap $O(A\, S^2)$ and no full policy evaluation is required. Policy Iteration is advantageous when the state–action space is moderate in size and fast convergence in very few iterations is desired, despite the higher per-iteration cost $O(S^2\, N_{\text{iter}} + A\, S^2)$; it quickly locks onto the optimal policy once policy evaluation completes.

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.

    - Value Iteration: We already showed :
    $$\|BV - BV'\|_\infty \le \gamma\, \|V - V'\|_\infty.$$
    If $\gamma = 1$, that becomes
    $$\|BV - BV'\|_\infty \le \|V - V'\|_\infty,$$
    so B stops being a contraction. That means Value iter. can have extreme positive values or become unstable and never converge.
    - Policy Iteration: When $\gamma = 1$, policy evaluation is solving $(I - P^\pi)V = r^\pi$. If P^π can cycle forever, $I - P^\pi$ the operator which is a matrix might be singular and you don't get a unique V^π. Without a well-defined evaluation, improving the policy doesn't reliably work.

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.

$$\|B^\pi V - B^\pi V'\|_\infty = \max_{s \in S} \Big| r(s,\pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V(s') - \big[r(s,\pi(s)) + \gamma \sum_{s'} P(s' \mid s, \pi(s)) V'(s') \big] \Big|$$

$$= \max_{s \in S} \gamma \Big| \sum_{s'} P(s' \mid s, \pi(s)) \big[V(s') - V'(s') \big] \Big| \leq \max_{s \in S} \gamma \sum_{s'} P(s' \mid s, \pi(s)) \big| V(s') - V'(s') \big| \leq \gamma \, \|V - V'\|_\infty .$$

   When we take abs from each $|V(s) - V'(s)|$ instead of the whole $|\sum_{s'} P(s' \mid s, \pi(s)) \big[V(s') - V'(s') \big]|$ the value increases.

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.

   To prove uniqueness, assume there are two distinct fixed-points. (If they are not distinct the required property is ongoing.)

$$V = B^\pi V \quad \text{and} \quad W = B^\pi W,$$

with $V \neq W$. Then by the contraction property of part 1 we have:

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

So we have:

$$(1-\gamma)\,\|V - W\|_\infty \;\leq\; 0,$$

and since $1 - \gamma > 0$ this forces $\|V - W\|_\infty = 0$, i.e. $V = W$. So there is no different fixed points and it is unique.

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

   For any fixed s:

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

   So we proved $B^\pi V \leq B^\pi V'$ elementwise using the fact that each of the elements of V are smaller than elements of 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?

   This equality happens when all the elements of BV and V are equal. Therefore the V is a fixed point and as we proved before, $V = V^*$

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

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

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

   We add and subtract $B^\pi V$, creating the term $B^\pi V - V^\pi$, and then proceed as follows:

$$\|\,V - V^\pi\| = \big\|\,V - B^\pi V \;+\; (B^\pi V - V^\pi)\big\| \;\leq\; \|\,V - B^\pi V\| \;+\; \|\,B^\pi V - V^\pi\|.$$

Since V^π is the fixed point of B^π, we have $V^\pi = B^\pi V^\pi$, and by the contraction property,

$$\| B^\pi V - V^\pi \| = \| B^\pi V - B^\pi V^\pi \| \leq \gamma \| V - V^\pi \|.$$

Then we have:

$$\| V - V^\pi \| \leq \| V - B^\pi V \| + \gamma \| V - V^\pi \| \implies (1-\gamma) \| V - V^\pi \| \leq \| V - B^\pi V \|,$$

For the second equation, we do the similar thing with the optimal V:

We add and subtract BV, creating the term $BV - V^*$, then we have:

$$\| V - V^* \| = \| V - BV + (BV - V^*) \| \leq \| V - BV \| + \| BV - V^* \|.$$

Since $V^* = BV^*$ is the fixed point of B, and B is a contraction,

$$\| BV - V^* \| = \| BV - BV^* \| \leq \gamma \| V - V^* \|.$$

Combining these we have:

$$\| V - V^* \| \leq \| V - BV \| + \gamma \| V - V^* \| \implies (1-\gamma) \| V - V^* \| \leq \| V - BV \|,$$

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.

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

From question 5 we have two inequalities:

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

Now use the triangle inequality at state s:

$$\begin{aligned} V^*(s) - V^\pi(s) &= \big(V^*(s) - V(s)\big) + \big(V(s) - V^\pi(s)\big) \\ &\leq \big|V^*(s) - V(s)\big| + \big|V(s) - V^\pi(s)\big| \\ &\leq \frac{\varepsilon}{1-\gamma} + \frac{\varepsilon}{1-\gamma} \\ &= \frac{2\varepsilon}{1-\gamma}. \end{aligned}$$

Rearranging gives the desired bound:

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

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

In the real-world applications when we are trying to do policy evaluation, when we have a lower bound on our current value function of the policy derived from the residual bellman error magnitude, we can be confident about the value function. So in critical use cases like RL in surgery, if a value function's lower bound is higher than a threshold, we are confident that it's true optimal value is better too.

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?

   No. The lower bound for both the V and V' just means that their policy's value funciton has the same lower bounds but they are not necessarily the same. They just both can't have a lower value than the same lower bound.

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?)*

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

Since $V^*(s) = \max_{\mu} V^{\mu}(s)$, for every policy π and state s we have:

$$V^{\pi}(s) \leq V^*(s).$$

From part 5 we have:

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

Since $V^*(s) \leq V(s)$ for all s, it follows that

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

So we have:

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

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 $\lim_{n\to\infty} B^n V$?

We use induction on n to show $B^n V \leq V$ for all n. The base case $n = 1$ is just $BV \leq V$. Assume $B^n V \leq V$. Since B is monotone, applying B to both sides gives

$$B^{n+1}V = B\big(B^n V\big) \leq BV \leq V.$$

Hence $B^n V \leq V$ for every n. By the contraction property, $\lim_{n\to\infty} B^n V = V^*$. We take the limit in the inequality $B^n V \leq V$:

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

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

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

References

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

[2] Cover image designed by freepik