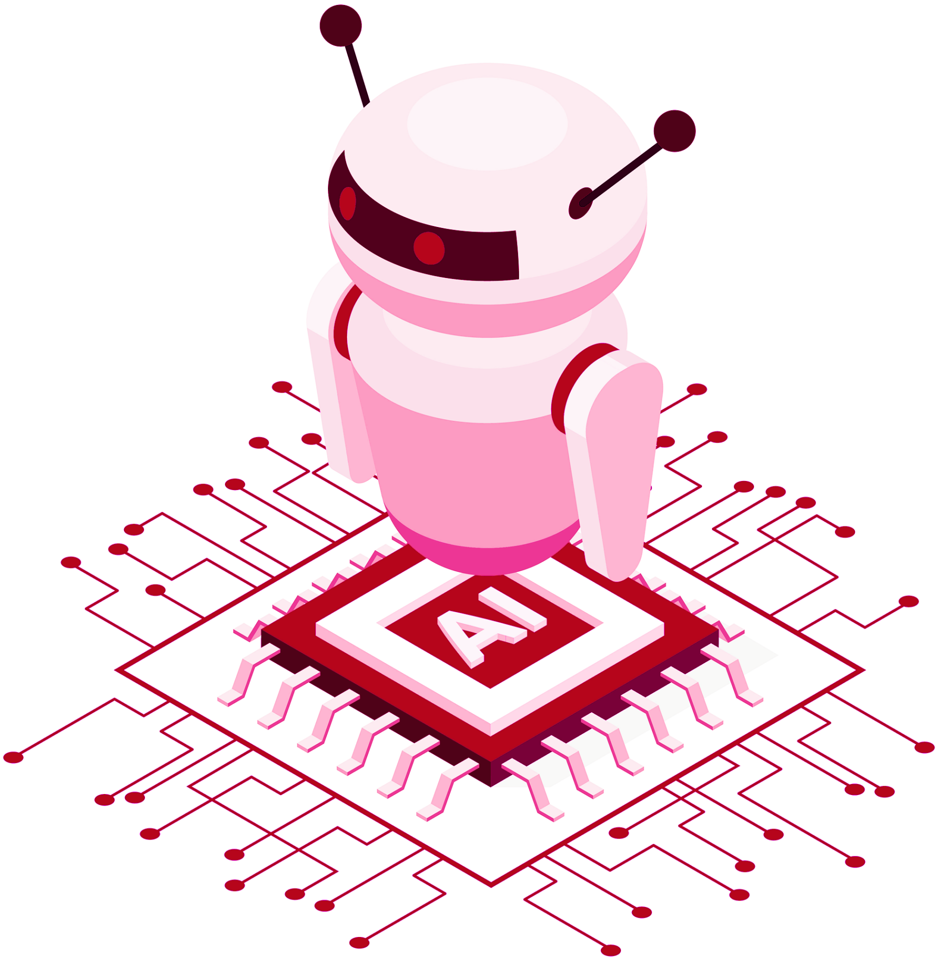

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Homework 7:

Value-Based Theory

By:

[Asemaneh Nafe]

[400105285]

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 $|R(s,a)| \leq R_{\text{max}}\ \forall(s,a)$, 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^*.

   We can write:

$$V_k^*(s) = \max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{k-1} \gamma^t R(s_t, a_t) \,\middle|\, s_0 = s\right]$$

   Since $R(s_t, a_t) \leq R_{\max}$ and $\gamma \in (0,1)$:

$$V_k^*(s) \leq \sum_{t=0}^{k-1} \gamma^t R_{\max} \leq k R_{\max}$$

   more precisely:

$$V_k^*(s) \leq \sum_{t=0}^{k-1} \gamma^t R_{\max} = R_{\max} \sum_{t=0}^{k-1} \gamma^t = R_{\max} \cdot \frac{1-\gamma^k}{1-\gamma}$$

   so:

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

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.

   We will show that $V_{k+1}^*(s) \geq V_k^*(s)$ for all s. Let π be the policy that gives the optimal return in the k-step horizon, i.e. $V_k^{\pi} = V_k^*$. Now if we use this policy for $k+1$ steps, then since the rewards are non-negative, we deduce: (The indices start from zero)

$$V_{k+1}^{\pi} = V_k^{\pi} + \gamma^k r_k \geq V_k^{\pi} = V_k^*$$

   Therefore:

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

so the series $\{V_k^*\}_{k=1}^{\infty}$ is non-decreasing and bounded from above. so they can increase in non infinit steps and we have a K where $V_{k+1}^* = V_k^*$ an this means that base on value itteration we have

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

so $V_{k+1}^*(s)$ satisfy the Bellman optimality equation because the Bellman operator is a contraction mapping. Therefore, the fixed point of this process is the unique solution to the Bellman equation.

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

   We take the limit in the Bellman update equation:

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

   Taking the limit as $k \to \infty$:

   $$\lim_{k\to\infty} V_{k+1}^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \lim_{k\to\infty} V_k^*(s') \right]$$

   So:

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

   This is exactly the Bellman optimality equation. Since the Bellman operator is a contraction mapping (due to $\gamma < 1$), this equation has a unique fixed point. Therefore, V^* is the **unique optimal value function**.and in if we itterate value itteration algorithm long enough we will converge in the infinit to v*.

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^*.

   We now drop the non-negativity constraint on rewards $R(s,a)$ and assume that no terminating (absorbing) states exist.

   Let $r_0 \in \mathbb{R}$ and define a new MDP with reward function:

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

   Let V_k^* and $\hat{V}_k^*$ denote the k-step optimal value functions for the original and modified MDPs respectively.

We aim to derive a relationship between $\hat{V}_k^*$ and V_k^*. The k-step expected return under the modified reward function is:

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

$$= \max_\pi \left(\mathbb{E}_\pi \left[\sum_{t=0}^{k-1} \gamma^t R(s_t, a_t) \right] + r_0 \sum_{t=0}^{k-1} \gamma^t \right) = V_k^*(s) + r_0 \cdot \frac{1-\gamma^k}{1-\gamma}$$

Taking the limit as $k \to \infty$, we obtain:

$$\hat{V}^*(s) = \lim_{k\to\infty} \hat{V}_k^*(s) = \lim_{k\to\infty} \left(V_k^*(s) + r_0 \cdot \frac{1-\gamma^k}{1-\gamma} \right) = V^*(s) + \frac{r_0}{1-\gamma}$$

Conclusion: Shifting the reward by a constant r_0 leads to a uniform shift in all values by $\frac{r_0}{1-\gamma}$. The optimal policy remains unchanged because the ordering of Q-values is preserved:

$$\hat{Q}^*(s,a) = Q^*(s,a) + \frac{r_0}{1-\gamma} \Rightarrow \arg\max_a \hat{Q}^*(s,a) = \arg\max_a Q^*(s,a)$$

Thus, Value Iteration still converges to the same optimal policy even when rewards can be negative.

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

Suppose there is a terminating state s_T where, once entered, the agent remains forever and receives no further rewards ($R(s_T, a) = 0$ for all a).

Now consider adding $r_0 > 0$ to all rewards. Then entering s_T yields no future return:

$$\hat{V}(s_T) = 0$$

But staying in non-terminal states can yield an infinite cumulative return:

$$\sum_{t=0}^{\infty} \gamma^t r_0 = \frac{r_0}{1-\gamma} > 0$$

So under the modified rewards, the agent would always prefer to *avoid* the terminal state, even if in the original MDP it was optimal to terminate.

Counterexample:

Let an MDP have two states: s_0 and s_T (terminal). Let:

$$R(s_0, a_0) = -10 \quad \text{(stay in } s_0\text{)}, \quad R(s_0, a_1) = 0 \quad \text{(go to } s_T\text{)}$$

Original optimal policy: choose a_1 to terminate and avoid accumulating negative reward.

If we shift rewards by $r_0 = +20$:

$$\hat{R}(s_0, a_0) = 10, \quad \hat{R}(s_0, a_1) = 20$$

But after transition to s_T, reward remains 0 forever, so:

$$\hat{V}(s_T) = 0, \quad \hat{V}(s_0) \text{ (under } a_0) \; \approx \frac{10}{1-\gamma} > 0$$

Now the agent prefers looping in s_0 under the shifted reward, even though it was suboptimal in the original setting.

Conclusion: Adding a constant to all rewards can change optimal policies if terminating states exist. Hence, the assumption of no terminal states is critical to preserve the optimality structure under constant reward shifts.

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.

I will proof in two ways 1. Proof via Backup Diagram Intuition.

Consider the Bellman backup diagram for a state s. In this diagram, the first layer corresponds to taking an action under the current policy π_k, and the second layer expands into the next states with their respective discounted values.

Now suppose instead of using π_k, we use the improved policy π_{k+1} at the first layer. By the policy improvement step, we know that:

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

which means that π_{k+1} selects the action that maximizes the expected return based on V^{π_k}. Therefore, for any state s, we have:

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

This means that in the backup diagram, replacing π_k with π_{k+1} in the first layer leads to a better or equal value at the root state s, since the improved policy puts more weight on better actions, which yield higher rewards and future values.

If we propagate this change to the thierd layer (i.e., recursively apply π_{k+1} at the children of s), we again replace the action at each child with the one maximizing the return under V^{π_k}. Doing so increases or maintains the value of each child node. This improved value then flows back up to the parent node (the original state s), leading to:

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

Applying this process across the entire state space, we systematically replace all uses of π_k with π_{k+1} throughout the backup diagram. This ensures that:

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

If for all states we have equality, i.e.,

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

$$V^{\pi_k}(s) = E[r + \gamma V^{\pi_k}(s')] = E[r + \gamma V^{\pi_{k+1}}(s')] = E[r + \gamma \max_a Q^{\pi_k}(s,a)], \quad \forall s \in S,$$

this is the bellman optimality equality so V^{π_k} is equal to v* which means π_k is optimal.

2. Proof in the source book. as I proofed (6), for all $s \in S$,

$$q_\pi(s, \pi'(s)) \geq v_\pi(s). \tag{1}$$

Then the policy π' must be as good as, or better than, π. That is, it must obtain greater or equal expected return from all states $s \in S$:

$$v_{\pi'}(s) \geq v_\pi(s). \tag{2}$$

Moreover, if there is strict inequality of (1) at any state, then there must be strict inequality of (2) at at least one state. This result applies in particular to the two policies that we considered in the previous paragraph, an original deterministic policy, π, and a changed policy, π', that is identical to π except that $\pi'(s) = a \neq \pi(s)$. Obviously, (1) holds at all states other than s. Thus, if $q_\pi(s,a) > v_\pi(s)$, then the changed policy is indeed better than π.

The idea behind the proof of the policy improvement theorem is easy to understand. Starting from (1), we keep expanding the q_π side and reapplying (1) until we get $v_{\pi'}(s)$:

$$\begin{aligned}
v_\pi(s) &\leq q_\pi(s, \pi'(s)) && (3)\\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] && (4)\\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s] && (5)\\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma \mathbb{E}_{\pi'}[R_{t+2} + \gamma v_\pi(S_{t+2})] \mid S_t = s] && (6)\\
&= \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 v_\pi(S_{t+2}) \mid S_t = s] && (7)\\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 v_\pi(S_{t+3}) \mid S_t = s] && (8)\\
&\vdots && (9)\\
&\leq \mathbb{E}_{\pi'}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \mid S_t = s] && (10)\\
&= v_{\pi'}(s). && (11)
\end{aligned}$$

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.

Proof of Convergence of Policy Iteration in Finite MDPs.

We aim to prove that Policy Iteration converges to the optimal policy π^* in a finite number of steps for a finite Markov Decision Process (MDP). That is, after finitely many iterations of evaluation and improvement, the algorithm reaches a policy that satisfies the Bellman optimality equation:

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

Assumptions: Let the MDP be defined by:

$$(S, A, P, R, \gamma),$$

where:

- S is a finite set of states,
- A is a finite set of actions,
- $P(s'|s,a)$ is the transition probability,
- $R(s,a)$ is the expected immediate reward,
- $\gamma \in [0,1)$ is the discount factor.

Step 1: Number of distinct policies is finite. Each deterministic policy maps every state $s \in S$ to a specific action $a \in A$. Since both S and A are finite, the total number of distinct deterministic policies is:

$$|A|^{|S|}.$$

This is a finite number.

Step 2: Policy improvement yields a strictly better policy unless optimal. From the previous result (Policy Improvement Theorem), we know:

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

with strict inequality for at least one state unless π_k is optimal. That is,

$$\pi_{k+1} = \pi_k \iff \pi_k \text{ is optimal.}$$

Step 3: No policy repeats unless convergence. Because the value function strictly improves (in at least one state) unless the policy is already optimal, and because there are only finitely many distinct policies, **no policy is repeated**. Otherwise, the process would loop and contradict the strict improvement of the value function.

Therefore, the sequence $\{\pi_k\}$ generated by Policy Iteration must terminate after a finite number of iterations with some policy π^*, which satisfies:

$$\pi_{k+1} = \pi_k = \pi^*.$$

Step 4: Final policy satisfies Bellman optimality equation. If π^* is not improved by the greedy policy improvement step, it must satisfy:

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

This is exactly the Bellman optimality condition, so π^* is an optimal policy.

Conclusion: Policy Iteration must converge in a finite number of steps to a policy π^* that satisfies the Bellman optimality equation. Hence, it always finds an optimal policy in finite MDPs.

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?

Convergence of Value Iteration and Policy Iteration to the Same Optimal Value Function V^*.

1. Value Iteration Converges to V^*. as shown earlier Let T be the Bellman optimality operator defined by:

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

This operator is known to be a *contraction mapping* under the sup-norm:

$$\|TV_1 - TV_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty.$$

By the Banach Fixed Point Theorem, T has a unique fixed point V^*, and for any initial guess V_0, the iterates:

$$V_{k+1} = TV_k,$$

converge to V^*.

2. Policy Iteration Converges to V^*. As shown earlier, the sequence V^{π_k} increases monotonically and converges to V^*, and the policy converges to an optimal policy π^* that satisfies the Bellman optimality equation:

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

3. V^* is Unique. The Bellman optimality equation has a unique solution V^*. Thus, both Value Iteration and Policy Iteration must converge to the same value function V^*, regardless of the path they take.

4. Different Policies with the Same Value Function. Even though the resulting policies from Value Iteration and Policy Iteration might differ in their action choices at some states, they are still optimal. This is because if a policy π satisfies:

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

then for all $s \in S$:

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

That is, π is greedy with respect to V^*, which makes it an optimal policy. Since there can be multiple actions that attain the maximum in the Bellman equation, multiple optimal policies may exist, all yielding the same optimal value function.

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.

1. Policy Iteration: Each iteration of Policy Iteration consists of two steps:

- **Policy Evaluation:** Solving the linear system:

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

  for all $s \in S$. This can be done either:

  - Exactly, via matrix inversion(s equation and s variabl): $\mathcal{O}(|S|^3)$,
  - Or approximately, using iterative methods (e.g., successive approximation) until convergence: $\mathcal{O}(N_{\text{eval}} \cdot |S|^2)$, where N_{eval} is the number of evaluation iterations.

- **Policy Improvement:** For each state, compute:

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

  This is $\mathcal{O}(|S||A|)$ and do this for all state will cost $\mathcal{O}(|S|^2|A|)$

Total cost per Policy Iteration: $\mathcal{O}(N_{\text{eval}} \cdot |S|^2 + |S|^2|A|)$.
If exact evaluation is used: $\mathcal{O}(|S|^3 + |S|^2|A|)$.

2. Value Iteration: Each iteration updates the value of every state using:

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

Cost per iteration: For each state and action, we compute a weighted sum over next states. Hence:

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

assuming a dense transition matrix.

3. Practical Comparison and Use Cases:

- **Value Iteration** performs cheap iterations but may require many to converge. It is preferred when:
  - Approximate solutions are acceptable,
  - Fast decisions are needed with limited computation budget,
  - Discount factor γ is small (faster convergence).
- **Policy Iteration** requires more expensive iterations (due to evaluation), but converges in fewer steps. It is preferred when:
  - The state space is small to medium,
  - Exact convergence is desired,
  - Efficient solvers for linear systems are available.

- In practice, **Modified Policy Iteration** (truncated evaluation steps) can provide a balance between the two methods.

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.

In an infinite-horizon Markov Decision Process (MDP), the discount factor $\gamma \in [0, 1)$ is typically used to ensure convergence of value functions by geometrically weighting future rewards. When $\gamma = 1$, future rewards are not discounted, and the objective becomes maximizing the total (undiscounted) expected reward:

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} R(s_t, a_t) \,\middle|\, s_0 = s \right].$$

1. Issues with $\gamma = 1$: When $\gamma = 1$, the total reward over an infinite horizon may:

- Diverge to infinity (e.g., in a reward loop),
- Be undefined (if rewards oscillate or have no limit),
- Make value functions unbounded.

This invalidates the standard contraction property used in proving convergence for both Value Iteration and Policy Iteration.

2. Value Iteration Behavior:

- The Bellman optimality operator:

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

  is *not* a contraction mapping when $\gamma = 1$.

- Therefore, Value Iteration:

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

  may **not converge**. Instead, the value function may oscillate or diverge unless additional assumptions are made (e.g., rewards are bounded and the MDP is transient or average-reward formulation is used).

3. Policy Iteration Behavior:

- Policy Evaluation step:

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

  becomes solving a linear system without guaranteed convergence when the system is not contractive.

- The matrix $(I - P^\pi)$ may not be invertible (if P^π has eigenvalue 1), which causes the solution to be non-unique or non-existent.

- Therefore, Policy Iteration may **fail to converge** or produce unstable results.

When $\gamma = 1$, a well-defined alternative is the **average-reward** (gain) formulation:

$$g^\pi = \lim_{N\to\infty} \frac{1}{N}\mathbb{E}_\pi\left[\sum_{t=0}^{N-1} R(s_t, a_t)\right].$$

In this setting, modified versions of Policy Iteration and Value Iteration (e.g., Relative Value Iteration) are used and can converge under ergodicity and unichain assumptions.

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.
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.
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.*

1. **Contraction property of B^π:**

   Let V and V' be two arbitrary value functions. We want to show that

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

   Proof:

   Recall the definition:

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

Then,

$$\begin{aligned}\|B^\pi V - B^\pi V'\|_\infty &= \gamma max_s \left|(\sum_{s'} p(s'|s,\pi(s))\,[V(s') - V'(s')])\right| \\ &\leq \gamma max_{s,s'}([V(s') - V'(s')]) \\ &= \gamma \|V - V'\|_\infty\end{aligned}$$

So B^π is a γ-contraction in the ∞-norm.

2. **Uniqueness of the fixed point for B^π:**

   Proof:

   Suppose, for contradiction, that there exist two fixed points V_1 and V_2 such that:

   $$V_1 = B^\pi V_1, \quad V_2 = B^\pi V_2, \quad \text{but} \quad V_1 \neq V_2$$

   Then,

   $$\|V_1 - V_2\|_\infty = \|B^\pi V_1 - B^\pi V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$$

   But since $0 \leq \gamma < 1$, this implies:

   $$\|V_1 - V_2\|_\infty < \|V_1 - V_2\|_\infty$$

   This is a contradiction unless $\|V_1 - V_2\|_\infty = 0$, which implies $V_1 = V_2$. Therefore, the fixed point is unique.

3. **Monotonicity of B^π:**

   Claim: If $V(s) \leq V'(s)$ for all s, then $(B^\pi V)(s) \leq (B^\pi V')(s)$ for all s.

   Proof:

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

   Since $V(s') \leq V'(s')$ for all s' and probabilities $p(s'|s,\pi(s)) \geq 0$, it follows that:

   $$\sum_{s'} p(s'|s,\pi(s)) V(s') \leq \sum_{s'} p(s'|s,\pi(s)) V'(s')$$

   Therefore,

   $$(B^\pi V)(s) \leq (B^\pi V')(s)$$

   So B^π is monotonic.

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? The Bellman error magnitude $\|BV - V\|$ equals 0 if and only if $V = BV$. That is, V is a fixed point of the Bellman optimality operator B.

   Since B is a γ-contraction and has a unique fixed point, this happens **only** when $V = V^*$, the optimal value function of the MDP.

   Therefore, $\|BV - V\| = 0$ if and only if $V = V^*$.

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

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

Answer:

Proof of the first inequality:

Define the Bellman operator for policy π as:

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

Then, note that V^π is the unique fixed point of B^π, i.e., $V^\pi = B^\pi V^\pi$.

Let us analyze the difference:

$$\|V - V^\pi\| = \|V - B^\pi V^\pi\|$$

Add and subtract $B^\pi V$:

$$\|V - V^\pi\| = \|V - B^\pi V + B^\pi V - B^\pi V^\pi\| \leq \|V - B^\pi V\| + \|B^\pi V - B^\pi V^\pi\|$$

Using the contraction property of B^π:

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

Therefore,

$$\|V - V^\pi\| \leq \|V - B^\pi V\| + \gamma \|V - V^\pi\|$$

Rearranging:

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

Proof of the second inequality:

Similarly, since $V^* = BV^*$ and B is a γ-contraction, we proceed analogously:

$$\|V-V^*\| = \|V-BV^*\| = \|V-BV+BV-BV^*\| \leq \|V-BV\| + \|BV-BV^*\|$$

Using contraction property of B:

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

So,

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

Therefore,

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

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.

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

Proof:

Let us define $\varepsilon = \|BV-V\|_{\infty}$, and let π be the greedy policy extracted from V, i.e.,

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

This implies that:

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

this holds for all s and π is greedy with respect to V, so:

$$BV = B^{\pi}V$$

Therefore,

$$\|V-B^{\pi}V\| = \|V-BV\| = \varepsilon$$

Now, from part (e), we already proved that:

$$\|V-V^{\pi}\| \leq \frac{\|V-B^{\pi}V\|}{1-\gamma} = \frac{\varepsilon}{1-\gamma} \quad (1)$$

Also, from the same part, we proved that:

$$\|V - V^*\| \leq \frac{\|V - BV\|}{1-\gamma} = \frac{\varepsilon}{1-\gamma} \quad (2)$$

Now, we use the triangle inequality:

$$\|V^\pi - V^*\| \leq \|V^\pi - V\| + \|V - V^*\| \leq \frac{\varepsilon}{1-\gamma} + \frac{\varepsilon}{1-\gamma} = \frac{2\varepsilon}{1-\gamma}$$

for all s we have:

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

scinse $V^\pi(s) \leq V^*(s)$ hence:

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

second way to proof:

From part (e), we know that for any policy π and any value function V:

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

Since π is the greedy policy with respect to V, we have:

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

and thus:

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

Now, by the definition of the infinity norm, the inequality above implies:

$$|V(s) - V^\pi(s)| \leq \frac{\varepsilon}{1-\gamma} \quad \text{for all } s \in S.$$

This gives the two-sided bound:

$$V^\pi(s) \geq V(s) - \frac{\varepsilon}{1-\gamma} \quad \text{and} \quad V^\pi(s) \leq V(s) + \frac{\varepsilon}{1-\gamma}. \qquad (1)$$

Also from part (e), we know:

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

which again implies:

$$|V(s) - V^*(s)| \leq \frac{\varepsilon}{1-\gamma} \quad \text{for all } s \in S,$$

so in particular:

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

Combining (1) and (2), we get:

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

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

   A domain where having a lower bound on $V^\pi(s)$ is useful is **autonomous driving**. In such systems, safety and performance guarantees are critical. If $V^\pi(s)$ represents the expected cumulative reward (e.g., safety, time efficiency, fuel consumption) starting from state s under policy π, then a guaranteed lower bound ensures that the policy will not perform worse than a certain level. This is essential for certification and deployment in safety-critical environments where catastrophic failures must be avoided.

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 does *not* imply that $V^\pi(s) = V^{\pi'}(s)$ for any s. The bound

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

only provides a worst-case guarantee on how far V^π can be from V^* in terms of performance. It says nothing about the exact value of $V^\pi(s)$ or how V^π compares to $V^{\pi'}$. Even if V and V' have the same Bellman residual magnitude ε, they can represent very different vectors, leading to different greedy policies π and π' and different value functions V^π and $V^{\pi'}$.

Therefore, the same residual error magnitude does not guarantee equality of the resulting value functions at any state.

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?)*

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

Suppose V is an arbitrary value function and π is the greedy policy extracted from V. Let $\varepsilon = \|BV - V\|$ be the Bellman residual magnitude. Assume that $V^* \leq V$, i.e., V overestimates the optimal value function element-wise. We want to show:

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

Proof: First, recall that for any policy π, the value function V^π is the unique fixed point of the Bellman operator B^π, i.e.,

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

Also recall that the greedy policy π satisfies:

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

So:

$$BV = B^\pi V$$

Hence,

$$\|BV - V\| = \|B^\pi V - V\| = \varepsilon$$

Now consider the error between V^π and V. Since B^π is a γ-contraction:

$$\|V - V^\pi\| = \|V - B^\pi V^\pi\| = \|V - B^\pi V + B^\pi V - B^\pi V^\pi\| \leq \|V - B^\pi V\| + \|B^\pi V - B^\pi V^\pi\| \leq \varepsilon + \gamma \|V - V^\pi\|$$

Rearranging:

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

This gives us a **pointwise bound** for any state s:

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

Since we assumed $V^*(s) \leq V(s)$ and also know $V^\pi(s) \leq V^*(s)$ for all s (because V^* is the optimal value function), we now have:

$$V(s) \geq V^*(s) \geq V^\pi(s)$$

Thus,

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

And since $V^*(s) \leq V(s)$, this gives:

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

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

    Proof: We will use induction on repeated applications of the Bellman optimality operator B. Recall that:

    $$V^{(0)} = V, \quad V^{(n+1)} = BV^{(n)} = B^{n+1}V$$

    and that the sequence $V^{(n)}$ converges to the unique fixed point V^*, i.e.,

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

    We will prove by induction that for all $n \geq 0$, we have:

    $$B^n V \leq V$$

    Base case: $n = 0$

    $$B^0 V = V \leq V \quad \text{(trivially true)}$$

    Inductive step: Assume that $B^n V \leq V$. Apply B (which is monotonic) to both sides:

    $$B^{n+1}V = B(B^n V) \leq B(V) \leq V \quad \text{(by assumption that } BV \leq V\text{)}$$

    Hence, by induction, $B^n V \leq V$ for all n. Taking the limit:

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

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

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

$B^{\pi^*}V^* = V^*$ and $B^\pi V = BV$. Thus, it follows that:

$$||B^\pi V - V^*|| = ||B^\pi V - B^{\pi^*}V^*|| \leq \gamma||V - V^*|| \leq \frac{\gamma\epsilon}{1-\gamma} \tag{12}$$

Now, defining $V' = BV$, base on the part 5 we have:

$$\frac{||BV' - BV||}{1-\gamma} = \frac{||BV' - V'||}{1-\gamma} \geq ||V^* - V'|| = ||V^* - BV|| = ||V^* - B^\pi V|| \tag{13}$$

by the contraction property:

$$\frac{||BV' - BV||}{1-\gamma} \leq \frac{\gamma||V' - V||}{1-\gamma} = \frac{\gamma\epsilon}{1-\gamma} \tag{14}$$

Concluding this chain of logic, we arrive at the statement:

$$||V^* - B^\pi V|| \leq \frac{\gamma\epsilon}{1-\gamma} \tag{15}$$

Integrating the prior results, we observe that:

$$||V^* - V^\pi|| = ||V^* - B^\pi V + B^\pi V - V^\pi|| \leq ||V^* - B^\pi V|| + ||B^\pi V - V^\pi|| \leq \frac{\gamma\epsilon}{1-\gamma} + \frac{\gamma\epsilon}{1-\gamma} = \frac{2\gamma\epsilon}{1-\gamma} \tag{16}$$

This chain of inequalities directly implies the desired first inequality:

$$\frac{2\gamma\epsilon}{1-\gamma} \geq ||V^* - V^\pi|| \geq V^*(s) - V^\pi(s) \tag{17}$$

so

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

To validate the second inequality, we employ the earlier bound:

$$\frac{\gamma\epsilon}{1-\gamma} \geq ||B^{\pi^*}V - V^*|| = ||BV - V^*|| \geq BV(s) - V^\pi(s) \tag{18}$$

using the fact that $V \geq V^*$, we infer that $BV \geq BV^* = V^*$. Substitution into the above expression then yields:

$$\frac{\gamma\epsilon}{1-\gamma} \geq V^*(s) - V^\pi(s) \tag{19}$$

Demonstrating the target result.

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

References

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

[2] Cover image designed by freepik