Assignment on Dynamic Programming and Operator Theory

Instructions. This short assignment is meant to reinforce core concepts from the first week of lecture. It should take 1–2 hours to complete. All questions can be answered based on the lecture notes—no additional material is required. Please show your work clearly. At the end, we include a few additional exercises that are more challenging or involve programming. These are optional and intended for further practice or discussion. Optional exercises are clearly marked with *.

Problem 1: Value Functions – Definitions and Relationship

(a) Definitions: Provide definitions for the *state-value function* $v^\pi(s)$ and the *action-value function* $q^\pi(s, a)$ for a policy π. Explain clearly in words their meanings, the meaning of $\mathbb{E}_\pi$, γ, s, a.

(b) Relationship: Clearly derive or explain the relationship between $v^\pi(s)$ and $q^\pi(s, a)$.

Problem 2: Computing v^π via Bellman Equations

Consider an MDP with states S_1, S_2, actions a_1, a_2, rewards and transitions given by:

$$S_1, a_1 \to S_2, R = 0 \qquad\qquad S_2, a_2 \to S_1, R = 3$$

with discount $\gamma = 0.5$.

(a) Write down Bellman equations for $v^\pi(S_1)$ and $v^\pi(S_2)$.

(b) Solve for v^π.

Problem 3: Contraction and Banach Fixed-Point Theorem

(a) Define a contraction mapping.

(b) Show the Bellman expectation operator is a contraction under $\|\cdot\|_\infty$.

(c) Consequences of the Banach fixed-point theorem.
Show from basic arguments that if an operator T is a γ-contraction it cannot have multiple fixed points.

Problem 4: Monotonicity and Policy Iteration

(a) Monotonic policy evaluation. Show that if $r(s, a) \geq 0$ for all (s, a), then, by initializing $v_0(s) = 0$, for all s, the value iteration algorithm for policy evaluation, i.e.,

$$v_{k+1} = T_\pi v_k$$

gives a sequence satisfying

$$v_0 \leq v_1 \leq \cdots \leq v_k \leq \cdots \leq v^\pi.$$

(b) Outline Policy Iteration.

(c) Show that PI provides monotonic improvement.

(d) Explain finite termination. Why does the algorithm terminate?, and the total number of sets is at most $|\mathcal{A}|^{|\mathcal{S}|}$?

Problem 5*: Policy Evaluation in a Simple Gambling MDP

You are in a casino! You start with \$10 and play until you either lose all your money or reach \$30. On each turn, you may choose one of two slot machines:

- **Slot Machine A:** costs \$10 and pays \$20 with probability 0.1, and \$0 otherwise.
- **Slot Machine B:** costs \$20 and pays \$30 with probability 0.4, and \$0 otherwise.

(a) Compute the expected reward of a single play from each machine.

(b) Model this scenario as an MDP:

- Define the state space and action space.
- Identify terminal states.
- Sketch a diagram of the MDP (e.g., similar to Example 3.3 in Sutton & Barto).

(c) Suppose you follow a policy π_β that chooses machine A with probability β and machine B with probability $1-\beta$ when you have \$20. When you have \$10, only machine A is available. Derive and solve the Bellman equations for $v^{\pi_\beta}(\$10)$ and $v^{\pi_\beta}(\$20)$.

(d) What is the optimal policy? Justify your answer based on the expressions from (c).

(e) Bonus: Generalize machine B to return \$30 with probability η (instead of 0.4). What condition on η ensures that all policies π_β are equally good?

Problem 6*: Equivalence of Discounted and Undiscounted MDPs with Geometric Horizon

Consider an undiscounted MDP $\mathcal{M}$ with action space $\mathcal{A}$ and state space $\mathcal{S} \cup \{z\}$ where z is an absorbing, terminal state satisfying:

$$p(z \mid z, a) = 1 \quad \forall a \in \mathcal{A}, \qquad r(z, a) = 0 \quad \forall a \in \mathcal{A}.$$

Assume that each step has a constant probability of transitioning to the terminal state:

$$P(z \mid s, a) = 1 - \gamma \qquad \text{for all } (s, a) \in \mathcal{S} \times \mathcal{A}.$$

The goal is to maximize the cumulative undiscounted reward:

$$\mathbb{E}_\pi \left[\sum_{t=0}^{\infty} R_{t+1} \,\middle|\, S_0 = s \right], \quad s \in \mathcal{S}.$$

(a) Let T denote the time step at which the process transitions to the absorbing state z, starting from $S_0 = s$. Show that T follows a geometric distribution and compute its parameter.

(b) Define the Bellman operators for this MDP as:

$$(T_\pi v)(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left(r(s, a) + \sum_{s' \in \mathcal{S} \cup \{z\}} P(s' \mid s, a) v(s') \right),$$

$$(T^* v)(s) = \max_{a \in \mathcal{A}} \left(r(s, a) + \sum_{s' \in \mathcal{S} \cup \{z\}} P(s' \mid s, a) v(s') \right).$$

Show that these operators are equivalent to the Bellman operators for a discounted MDP $\widetilde{\mathcal{M}}$ (with no terminal state and discount factor γ), with adjusted transition probabilities:

$$\widetilde{P}(s' \mid s, a) = \frac{1}{\gamma} P(s' \mid s, a).$$

Problem 7*: Robustness of Value Functions to Reward Perturbations

Let $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p, \gamma)$ be an MDP with discount factor $\gamma \in (0,1)$. Consider a perturbed MDP $\hat{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \hat{p}, \gamma)$, where:

$$\hat{p}(s' \mid s, a) = p(s' \mid s, a), \quad \text{for all } s, a, s',$$

and the reward distributions differ by at most ϵ in expectation:

$$\left| \sum_r r\hat{p}(r \mid s, a) - \sum_r rp(r \mid s, a) \right| \leq \epsilon \quad \forall s, a.$$

(a) Let $v^*(s)$ and $\hat{v}^*(s)$ be the optimal value functions of M and $\hat{M}$, respectively. Show that:

$$\|v^* - \hat{v}^*\|_\infty \leq \frac{\epsilon}{1-\gamma}.$$

(Hint: Let $\mathcal{T}_{\max}$ and $\hat{\mathcal{T}}_{\max}$ be the Bellman optimality operators of M and $\hat{M}$, respectively. Use the contraction property of $\mathcal{T}_{\max}$ and compare the fixed points.)

(b) Suppose instead that the rewards in $\hat{M}$ are deterministically shifted by ϵ:

$$\hat{r}(s, a) = r(s, a) + \epsilon, \quad \forall s, a.$$

Show that:

$$\hat{v}^*(s) = v^*(s) + \frac{\epsilon}{1-\gamma}, \quad \forall s.$$