CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #4: Dynamic Programming

Manuscript version: *#4a4cf0-dirty* - 2023-02-14 17:08:07-07:00

In these problems, unless otherwise specified, we consider a finite MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, p)$ with the discounted total expected reward criterion with discount factor $0 \leq \gamma < 1$.

Question 1. In class we showed that the optimal value function v^* is the fixed point of the Bellman optimality equation:

$$v^*(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s') \right\}, \qquad s \in \mathcal{S}\,. \tag{1}$$

Define $q^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ by

$$q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v^*(s')\,, \qquad s \in \mathcal{S}, a \in \mathcal{A}\,. \tag{2}$$

Show that q^* satisfies the following fixed-point equation:

$$q^*(s,a) = r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} q^*(s',a')\,, \qquad s \in \mathcal{S}\,, a \in \mathcal{A}\,. \tag{3}$$

Question 2. Let π be a memoryless policy, let q_π be its action-value function. Show that q_π is the fixed point of the operator $\tilde{T}_\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ that is defined as

$$(\tilde{T}_\pi q)(s,a) = \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \sum_{a' \in \mathcal{A}} \pi(a'|s') q(s',a') \right] \qquad (s \in \mathcal{S}, a \in \mathcal{A}, q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}) .$$

Question 3. Let π be a memoryless policy. The Bellman operator for evaluating π is defined to be the operator $T_\pi : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ such that

$$(T_\pi v)(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v_k(s') \right] .$$

Show that T_π is a contraction.

Note: An alternate way to prove this, as suggested in the lectures, would be to first split the operator T into multiple operators as

$$T = \Pi \circ R \circ G \circ P,$$

where $\Pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}}$ is defined $(\Pi x)(s) := \sum_{a\in\mathcal{A}} \pi(a|s)x(s,a)$, $R : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is defined $(Rx)(s,a) := r(s,a) + x(s,a)$, $G : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is defined $(Gx)(s,a) := \gamma x(s,a)$, and $P : \mathbb{R}^{\mathcal{S}} \to R^{\mathcal{S}\times\mathcal{A}}$ is defined $(Px)(s,a) := \sum_{s'\in\mathcal{S}} p(s'|s,a)x(s)$. And then to show that each of these individual operators is a contraction, which would imply that T is a contraction as well.

Question 4. Let π be a memoryless policy. Take the Bellman operator $\tilde{T}_\pi$ as defined in Question 2. Show that $\tilde{T}_\pi$ is a contraction with contraction factor γ with respect to the maximum norm.

Question 5. (****) Let π be a memoryless policy. Take the Bellman operator T_π as defined earlier. Show that T_π is a contraction with contraction factor γ with respect to the weighted-2-norm

$$\|v\|_\pi = \sqrt{\sum_{s\in\mathcal{S}} \mu(s)(v(s))^2}\,,$$

where $\mu \in \mathcal{M}_1(\mathcal{S})$ is a distribution such that for any $s' \in \mathcal{S}$,

$$\mu(s') = \sum_{s\in\mathcal{S}} \mu(s) \sum_{a\in\mathcal{A}} \pi(a|s)p(s'|s,a)\,.$$

The distribution μ is known as the stationary distribution of π.

Note: We now discuss the inequality we used in the solution: For $n \in \mathbb{N}$, given real numbers $\alpha_1, \alpha_2, \ldots, \alpha_n \in [0,1]$ that sum to one, and arbitrary real numbers $x_1, x_2, \ldots, x_n \in \mathbb{R}$, the following inequality holds

$$\left(\sum_{i=1}^n \alpha_i x_i\right)^2 \le \sum_{i=1}^n \alpha_i x_i^2.$$

This is a special case of the incredibly useful Jensen's inequality (pronounced as "Yensen"). For completeness (and for fun!), we give an elementary proof of this special case.

We use induction on n. The case $n = 1$ is trivial. For $n = 2$,

$$\begin{aligned}
&(\alpha_1 x_1^2 + \alpha_2 x_2^2) - (\alpha_1 x_1 + \alpha_2 x_2)^2 = \alpha_1 x_1^2 + \alpha_2 x_2^2 - (\alpha_1^2 x_1^2 + \alpha_2 x_2^2 + 2\alpha_1\alpha_2 x_1 x_2)\\
&= \alpha_1(1-\alpha_1)x_1^2 + \alpha_2(1-\alpha_2)x_2^2 - 2\alpha_1\alpha_2 x_1 x_2 = \alpha_1\alpha_2 x_1^2 + \alpha_1\alpha_2 x_2^2 - 2\alpha_1\alpha_2 x_1 x_2\\
&= \alpha_1\alpha_2(x_1 - x_2)^2 \ge 0,
\end{aligned}$$

gives us the desired result, $\alpha_1 x_1^2 + \alpha_2 x_2^2 \ge (\alpha_1 x_1 + \alpha_2 x_2)^2$. Now assume that there exists an m, such that the inequality holds for all $n \le m$. (Note that for each of the different values of n, we will have different sequences $(\alpha_i)_i$ and $(x_i)_i$). Now we will show that the inequality would also hold for $n+1$. Consider the appropriately defined sequences $(\alpha_i)_{i\in[n+1]}$ and $(x_i)_{i\in[n+1]}$. Then,

$$\begin{aligned}
\left(\sum_{i=1}^{n+1} \alpha_i x_i\right)^2 &= \left(\alpha_1 x_1 + \sum_{i=2}^{n+1} \alpha_i x_i\right)^2 = \left(\alpha_1 x_1 + (1-\alpha_1)\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} x_i\right)^2\\
&\overset{(a)}{\le} \alpha_1 x_1^2 + (1-\alpha_1)\left(\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} x_i\right)^2 \overset{(b)}{\le} \alpha_1 x_1^2 + (1-\alpha_1)\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} x_i^2\\
&= \sum_{i=1}^{n+1} \alpha_i x_i^2,
\end{aligned}$$

where in step (a) we used inequality for 2 variables; and in step (b), noting that $\sum_{i=2}^{n+1} \frac{\alpha_i}{1-\alpha_1} = 1$, we used the inequality for n variables. This completes the induction, and thereby also the proof.

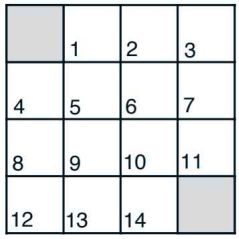
Question 6. Show that the operator $\tilde{T} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined using

$$(\tilde{T}q)(s,a) = \left[r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \max_{a' \in \mathcal{A}} q_k(s',a') \right] \qquad (s \in \mathcal{S}, a \in \mathcal{A}, q \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}})$$

is a contraction with contraction factor γ with respect to the maximum norm.

Question 7. Every deterministic memoryless policy corresponds to a stochastic memoryless policy (the reverse is not true). What probability will the stochastic memoryless policy assign to an action $a \in \mathcal{A}$ given a state $s \in \mathcal{S}$ that corresponds to the deterministic memoryless policy as given by a map $\pi : \mathcal{S} \to \mathcal{A}$.

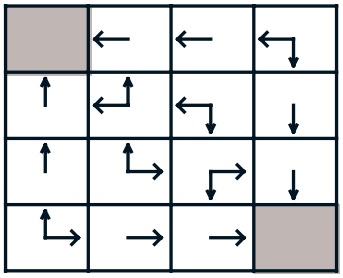
Question 8. (*Exercise 4.1 S&B)* Consider the 4x4 gridworld below, where actions that would take the agent off the grid leave the state unchanged. The first figure shows the number of nonterminal states. The task is episodic with $\gamma = 1$ and the terminal states are the shaded blocks. Using the precomputed values for the equiprobable policy below, what is $q_\pi(11, \text{down})$? What is $q_\pi(7, \text{down})$? (The figure in the bottom, left give the values of v_π.)



$R_t = -1$
on all transitions

$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0



Question 9. (*Exercise 4.1 S&B*) Consider the above gridworld again. But now assume that a new state 15 is added to the gridworld just below state 13, and its actions, left, up, right, and down, take the agent to the states 12, 13, 14, and 15, respectively. Assume that the transitions from the original states are unchanged. What, then is, $v_\pi(15)$ for the equiprobable random policy? Now suppose the dynamics of state 13 are also changed, such that action down from state 13 takes the agent to the new state 15. What is $v_\pi(15)$ for the equiprobable random policy in this case?

Question 10. (Challenge Question) A gambler has the opportunity to make bets on the outcomes of a sequence of coin flips. If the coin comes up heads, she wins as many dollars as she has staked on that flip; if it is tails, she loses her stake. The game ends when the gambler wins by reaching her goal of \$100, or loses by running out of money. On each flip, the gambler must decide what portion of her capital to stake, in integer numbers of dollars. The gambler is interested in maximizing the chance that she reaches her goal.

This problem can be formulated as an undiscounted, episodic, finite MDP. The state is the gambler's capital, $s \in \mathcal{S} = \{0, 1, 2, ..., 99, 100\}$ (with 0 and 100 being the terminal states), and the actions are the stakes $a \in \mathcal{A}(s) = \{1, ..., \min(s, 100 - s)\}$ for $s \in \mathcal{S} \setminus \{0, 100\}$. (Note that the action set depends on the state the agent is in, and we don't worry about taking any actions in the terminal states.) The reward is $+1$ when reaching the goal of 100 and zero on all other transitions. The probability of seeing heads is $p_h = 0.4$.

1. What does the value of a state mean in this problem? (For example, in a gridworld, modeled as an undiscounted finite-horizon MDP and where the agent receives a reward of 1 per step, the value represents the expected number of steps to the goal state.) To build some intuition, think about the minimum and maximum possible values, and think about the values of state 50 (which is 0.4) and state 99 (which is near 0.95). Also, to simplify things, only focus on $v_\pi(s)$ for $s \in \mathcal{S} \setminus \{0, 100\}$.

2. Modify the pseudocode for value iteration to more efficiently solve this specific problem, by exploiting your knowledge of the dynamics. *Hint: Not all states transition to every other state. For example, can you transition from state* 1 *to state* 99*?*

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(\mathit{terminal}) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $\quad v \leftarrow V(s)$
| $\quad V(s) \leftarrow \max_a \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$
| $\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r \,|\, s, a)\big[r + \gamma V(s')\big]$

Question 11. (Challenge Question) (*Exercise 4.4 S&B)* The policy iteration algorithm on page 80 has a subtle bug in that it may never terminate if the policy continually switches between two or more policies that are equally good (the lectures on the other hand were careful about this). This is okay for pedagogy, but not for actual use. Modify the pseudocode so that convergence is guaranteed. Note that there is more than one approach to solve this problem.

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
   $\Delta \leftarrow 0$
   Loop for each $s \in \mathcal{S}$:
   $v \leftarrow V(s)$
   $V(s) \leftarrow \sum_{s',r} p(s', r \,|\, s, \pi(s)) \big[r + \gamma V(s')\big]$
   $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
   old-action $\leftarrow \pi(s)$
   $\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r \,|\, s, a) \big[r + \gamma V(s')\big]$
   If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow false$
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2