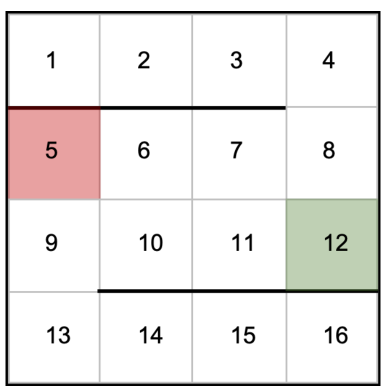
CS 234 Winter 2020
Assignment 1
Due: January 22 at 11:59 pm

For submission instructions please refer to website. For all problems, if you use an existing result from either the literature or a textbook to solve the exercise, you need to cite the source.

1 Gridworld [15 pts]

Consider the following grid environment. Starting from any unshaded square, you can move up, down, left, or right. Actions are deterministic and always succeed (e.g. going left from state 16 goes to state 15) unless they will cause the agent to run into a wall. The thicker edges indicate walls, and attempting to move in the direction of a wall results in staying in the same square (e.g. going in any direction other than left from state 16 stays in 16). Taking any action from the green target square (no. 12) earns a reward of r_g (so $r(12, a) = r_g \ \forall a$) and ends the episode . Taking any action from the red square of death (no. 5) earns a reward of r_r (so $r(5, a) = r_r \ \forall a$) and ends the episode. Otherwise, from every other square, taking any action is associated with a reward $r_s \in \{-1, 0, +1\}$ (even if the action results in the agent staying in the same square). Assume the discount factor $\gamma = 1$, $r_g = +5$, and $r_r = -5$ unless otherwise specified.



(a) (3pts) Define the value of r_s that would cause the optimal policy to return the shortest path to the green target square (no. 12). Using this r_s, find the optimal value for each square.

Solution

r_s = -1.

0	1	2	3
-5	2	3	4
2	3	4	5
1	0	-1	-2

(b) (3pts) Lets refer to the value function derived in (a) as $V_{old}^{\pi_g}$ and the policy as π_g. Suppose we are now in a new gridworld where all the rewards (r_s, r_g, and r_r) have +2 added to them. Consider still following π_g of the original gridworld, what will the new values $V_{new}^{\pi_g}$ be in this second gridworld?

Solution

12	11	10	9
-3	10	9	8
10	9	8	7
11	12	13	14

(c) (3pts) Consider a general MDP with rewards, and transitions. Consider a discount factor of γ. For this case assume that the horizon is infinite (so there is no termination). A policy π in this MDP induces a value function V^π (lets refer to this as V_{old}^π). Now suppose we have a new MDP where the only difference is that all rewards have a constant c added to them. Can you come up with an expression for the new value function V_{new}^π induced by π in this second MDP in terms of V_{old}^π, c, and γ?

Solution

$$\begin{aligned}
V_{old}^\pi(s) &= E_\pi[\sum_{T=0}^{\infty} \gamma^T r_{t+T} | s_t = s] \\
V_{new}^\pi(s) &= E_\pi[\sum_{T=0}^{\infty} \gamma^T (r_{t+T} + c) | s_t = s] \\
&= E_\pi[\sum_{T=0}^{\infty} \gamma^T r_{t+T} | s_t = s] + c \sum_{T=0}^{\infty} \gamma^T
\end{aligned}$$

$$= V^{\pi}_{old}(s) + \frac{c}{1-\gamma}$$

(d) (2pts) Lets go back to our gridworld from (a) with the default values for r_g, r_r, γ and with the value you specified for r_s. Suppose we now derived a second gridworld by adding a constant c to all rewards (r_s, r_g, and r_r) such that $r_s = +2$. How does the optimal policy change (Just give a one or two sentence description)? What do the values of the unshaded squares become?

Solution

The optimal policy becomes a policy that would just wander around forever, never reaching either target. The value of all unshaded squares (and the green target square) become $+\infty$. The value of the red square of death is -2.

(e) (2pts) Lets take the second gridworld from part (d) and change γ such that $\gamma < 1$. What happens to the optimal policy now?

Solution

For γ close to one the optimal policy is still to wander around forever. However as gamma decreases, there will be a point where the optimal policy switches to reach the green target square in the shortest time.

(f) (2pts) Lets go back to our gridworld from (a) with the default values for r_g, r_r, γ and with the value you specified for r_s. In this gridworld, our optimal policy from any unshaded square never terminates in the red square. Now suppose r_s can take on any real, non-infinite value and is not restricted to $\{+1, 0, -1\}$ anymore. Give a value of r_s such that there are unshaded squares starting from which following the optimal policy results in termination in the red square.

Solution

Any value of $r_s \leq -5$ is correct.

2 Value of Different Policies [35 pts]

In many situations such as healthcare or education, we cannot run any arbitrary policy and collect data from running those policies for evaluation. In these cases, we may need to use data collected from following one policy and to evaluate the value of a different policy. The equality proved in the following exercise can be an important tool for achieving this. The purpose of this exercise is to get familiar on how to compare the value of different policies, π_1 and π_2, on a fixed horizon MDP. A fixed horizon MDP is an MDP where the agent's state is reset after H timesteps; H is called the *horizon* of the MDP. There is no discount (i.e., $\gamma = 1$) and policies are allowed to be non-stationary, i.e., the action identified by a policy depends on the timestep in addition to the state. Let $x_t \sim \pi$ denote the distribution over states at timestep t (for $1 \le t \le H$) upon following policy π and $V_t^\pi(x_t)$ denote the value function of policy π in state x_t and timestep t, and $Q_t^\pi(x_t, a)$ denote the corresponding Q value associated to action a. As a clarifying example, $\mathbb{E}_{x_t \sim \pi_1} V(x_t)$ is the average value of the value function $V(\cdot)$ over the states at timestep t encountered upon following policy π_1. Please show the following:

$$V_1^{\pi_1}(x_1) - V_1^{\pi_2}(x_1) = \sum_{t=1}^{H} \mathbb{E}_{x_t \sim \pi_2} \Big(Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t, t)) \Big)$$

Intuition: The above expression can be interpreted in the following way. For concreteness, assume that π_1 is the better policy, i.e., achieving $V_1^{\pi_1}(x_1) \ge V_1^{\pi_2}(x_1)$. Suppose you're following policy π_2 and you are at timestep t in state x_t. You have the option to follow π_1 (the better policy) until the end of the episode, totalling $Q_t^{\pi_t}(x_t, \pi_1(x_t, t))$ return from the current state-timestep; or you have the option to follow π_2 for one timestep and then follow π_1 instead until the end of the episode (you can follow many other policies of course). This would you give you a "loss" of $Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t, t))$ that originates from following the worse policy π_2 instead of π_1 in that timestep. Then the equation above means that the value difference of the two policies is the sum of all the losses induced by following the suboptimal policy for every timestep, weighted by the expected trajectory of the policy you're following.

Solution We proceed by induction. For a generic state x_t at timestep t consider the value function difference:

$$V_t^{\pi_1}(x_t) - V_t^{\pi_2}(x_t) = Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) \overbrace{- Q_t^{\pi_1}(x_t, \pi_2(x_t, t)) + Q_t^{\pi_1}(x_t, \pi_2(x_t, t))}^{=0} - Q_t^{\pi_2}(x_t, \pi_2(x_t, t)).$$

We focus on the second difference and recall the definition of Q values:

$$\begin{aligned}
&Q_t^{\pi_1}(x_t, \pi_2(x_t, t)) - Q_t^{\pi_2}(x_t, \pi_2(x_t, t)) = \\
&= r_t(x_t, \pi_2(x_t, t)) + \mathbb{E}_{s' \sim p(x_t, \pi_2(x_t, t))} V_{t+1}^{\pi_1}(s') - r_t(x_t, \pi_2(x_t, t)) - \mathbb{E}_{s' \sim p(x_t, \pi_2(x_t, t))} V_{t+1}^{\pi_2}(s') \\
&= \mathbb{E}_{s' \sim p(x_t, \pi_2(x_t, t))} (V_{t+1}^{\pi_1}(s') - V_{t+1}^{\pi_2}(s'))
\end{aligned}$$

Plugging back into the prior display we obtain:

$$V_t^{\pi_1}(x_t) - V_t^{\pi_2}(x_t) = Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t, t)) + \mathbb{E}_{s' \sim p(x_t, \pi_2(x_t, t))} (V_{t+1}^{\pi_1}(s') - V_{t+1}^{\pi_2}(s'))$$

The first difference above looks like one term of the summation in the expression we need to prove; the second difference (the expectation) by induction gives the remaining terms. To see this, take expectation over the trajectories induced by π_2 at time t of the above expression; the tower property of expectation gives $\mathbb{E}_{x_t \sim \pi_2} \mathbb{E}_{s' \sim p(x_t, \pi_2(x_t, t))} = \mathbb{E}_{x_{t+1} \sim \pi_2}$:

$$\mathbb{E}_{x_t \sim \pi_2}(V_t^{\pi_1}(x_t) - V_t^{\pi_2}(x_t)) = \mathbb{E}_{x_t \sim \pi_2}(Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t, t))) + \mathbb{E}_{x_{t+1} \sim \pi_2}(V_{t+1}^{\pi_1}(x_{t+1}) - V_{t+1}^{\pi_2}(x_{t+1}))$$

The second term is equal to the left hand side but written for $t + 1$; we can use the inductive hypothesis there to obtain:

$$\begin{aligned}
&\mathbb{E}_{x_t \sim \pi_2}(V_t^{\pi_1}(x_t) - V_t^{\pi_2}(x_t)) \\
&= \mathbb{E}_{x_t \sim \pi_2}(Q_t^{\pi_1}(x_t, \pi_1(x_t, t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t, t))) + \sum_{\tau=t+1}^{H} (\mathbb{E}_{x_\tau \sim \pi_2}(Q_\tau^{\pi_1}(x_\tau, \pi_1(x_\tau, \tau)) - Q_\tau^{\pi_1}(x_\tau, \pi_2(x_\tau, \tau)))) \\
&= \sum_{\tau=t}^{H} (\mathbb{E}_{x_\tau \sim \pi_2}(Q_\tau^{\pi_1}(x_\tau, \pi_1(x_\tau, \tau)) - Q_\tau^{\pi_1}(x_\tau, \pi_2(x_\tau, \tau)))).
\end{aligned}$$

By induction, the above expression is valid for every t; evaluation at $t = 1$ gives the thesis.

3 Fixed Point [25 pts]

In this exercise we will use Cauchy sequences to prove that value iteration will converge to a unique fixed point (in this case, a value function V) regardless of the starting point. An element V is a fixed point for an operator B (in this case the Bellman operator) if performance of B on V returns V, i.e., $BV = V$. Recall that the Bellman backup operator B is defined as (in lecture 2):

$$V_{k+1} \stackrel{def}{=} BV_k = \max_a [R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V_k^{\pi}(s')].$$

Additionally, in lecture 2, we proved that this Bellman backup is a contraction for $\gamma < 1$ on the infinity norm

$$\|BV' - BV''\|_\infty \leq \gamma \|V' - V''\|_\infty$$

for any two value functions V' and V'', meaning if we apply it to two different value functions, the distance between value functions (in the ∞ norm) shrinks after application of the operator to each element.

(a) (5pts) Prove by induction that $\|V_{n+1} - V_n\|_\infty \leq \gamma^n \|V_1 - V_0\|_\infty$

Solution

Base case $n = 1$: We want to show $\|V_2 - V_1\|_\infty = \|V_1 - V_0\|_\infty$
$\|V_2 - V_1\|_\infty = \|BV_1 - BV_0\|_\infty \leq \gamma \|V_1 - V_0\|_\infty$.

Inductive case: Assume $\|V_{k+1} - V_k\|_\infty \leq \gamma^k ||V_1 - V_0||_\infty$
We want to show $\|V_{k+2} - V_{k+1}\|_\infty \leq \gamma^{k+1} \|V_1 - V_0\|_\infty$.

$\|V_{k+2} - V_{k+1}\|_\infty = \|BV_{k+1} - BV_k\|_\infty \leq \gamma \|V_{k+1} - V_k\|_\infty \leq \gamma\gamma^k \|V_1 - V_0\|_\infty = \gamma^{k+1} \|V_1 - V_0\|_\infty$
Where the first inequality holds by definition of contraction, and the second holds by the inductive hypothesis.

(b) (10pts) Prove that for any $c > 0$, $\|V_{n+c} - V_n\|_\infty \leq \frac{\gamma^n}{1-\gamma} \|V_1 - V_0\|_\infty$

Solution

Let $m = n + c$

Note that $\|V_m - V_n\|_\infty = \|V_m - V_{n+1} + V_{n+1} - V_n\|_\infty \leq \|V_m - V_{n+1}\|_\infty + \|V_{n+1} - V_n\|_\infty$
Where the inequality results from the triangular inequality.

Therefore

$$\begin{aligned}
\|V_m - V_n\|_\infty &\leq \sum_{k=n}^{m-1} \|V_{k+1} - V_k\|_\infty \\
&\leq \sum_{k=n}^{m-1} \gamma^k \|V_1 - V_0\|_\infty \\
&= \frac{\gamma^n - \gamma^m}{1-\gamma} \|V_1 - V_0\|_\infty \\
&\leq \frac{\gamma^n}{1-\gamma} \|V_1 - V_0\|_\infty
\end{aligned}$$

A *Cauchy sequence* is a sequence whose elements become arbitrarily close to each other as the sequence progresses. Formally a sequence $\{a_n\}$ in metric space X with distance metric d is a Cauchy sequence if given an $\varepsilon > 0$ there exists k such that if m, n $>$ k then $d(a_m, a_n) < \varepsilon$. Real Cauchy sequences are convergent.

(c) (2pts) Using this information about Cauchy sequences, argue that the sequence $V_0, V_1, ...$ is a Cauchy sequence and is therefore convergent and must converge to some element V and this V is a fixed point

Solution

$\frac{\gamma^n}{1-\gamma}||V_1 - V_0||_\infty \to 0$ as $n \to \infty$. Therefore for every value of ε we can find a n such that $\frac{\gamma^n}{1-\gamma}||V_1 - V_0||_\infty < \varepsilon$. In our case n plays the role of k in the definition so our sequence V_n is a Cauchy sequence. Therefore V_n is convergent and will converge to a fixed point.

(d) (8pts) Show that this fixed point is unique.

Solution

Proof by contradiction: Assume there are two fixed points, V and V'.
$\|V - V'\|_\infty = \|TV - TV'\|_\infty \leq \gamma\|V - V'\|_\infty$. However $\gamma \neq 0$ and $\gamma \neq 1$ so $\|V - V'\|_\infty = 0$ and there is only one fixed point.

4 Frozen Lake MDP [25 pts]

Now you will implement value iteration and policy iteration for the Frozen Lake environment from OpenAI Gym. We have provided custom versions of this environment in the starter code.

(a) **(coding)** (10 pts) Read through `vi_and_pi.py` and implement `policy_evaluation`, `policy_improvement` and `policy_iteration`. The stopping tolerance (defined as $\max_s |V_{old}(s) - V_{new}(s)|$) is tol $= 10^{-3}$. Use $\gamma = 0.9$. Return the optimal value function and the optimal policy.

(b) **(coding)** (10 pts) Implement `value_iteration` in `vi_and_pi.py`. The stopping tolerance is tol $= 10^{-3}$. Use $\gamma = 0.9$. Return the optimal value function and the optimal policy.

(c) **(written)** (5 pts) Run both methods on the Deterministic-4x4-FrozenLake-v0 and Stochastic-4x4-FrozenLake-v0 environments. In the second environment, the dynamics of the world are stochastic. How does stochasticity affect the number of iterations required, and the resulting policy?