CSE 250a. Assignment 7

Out: *Tue Dec 01*
Due: *Tue Dec 08* — to box outside room 3214 before noon — **NO EXTENSIONS**
Reading: *Sutton & Barto*, Chapters 1-4.

7.1 Policy improvement

Consider the Markov decision process (MDP) with two states $s \in \{0, 1\}$, two actions $a \in \{0, 1\}$, discount factor $\gamma = \frac{2}{3}$, and rewards and transition matrices as shown below:

s	$R(s)$
0	-2
1	4

s	s'	$P(s'\|s, a=0)$
0	0	$\frac{3}{4}$
0	1	$\frac{1}{4}$
1	0	$\frac{1}{4}$
1	1	$\frac{3}{4}$

s	s'	$P(s'\|s, a=1)$
0	0	$\frac{1}{2}$
0	1	$\frac{1}{2}$
1	0	$\frac{1}{2}$
1	1	$\frac{1}{2}$

(a) Consider the policy π that chooses the action $a = 0$ in each state. For this policy, solve the linear system of Bellman equations to compute the state-value function $V^\pi(s)$ for $s \in \{0, 1\}$. Your answers should complete the following table.

s	$\pi(s)$	$V^\pi(s)$
0	0	
1	0	

(b) Compute the greedy policy $\pi'(s)$ with respect to the state-value function $V^\pi(s)$ from part (a). Your answers should complete the following table.

s	$\pi(s)$	$\pi'(s)$
0	0	
1	0	

7.2 Value and policy iteration

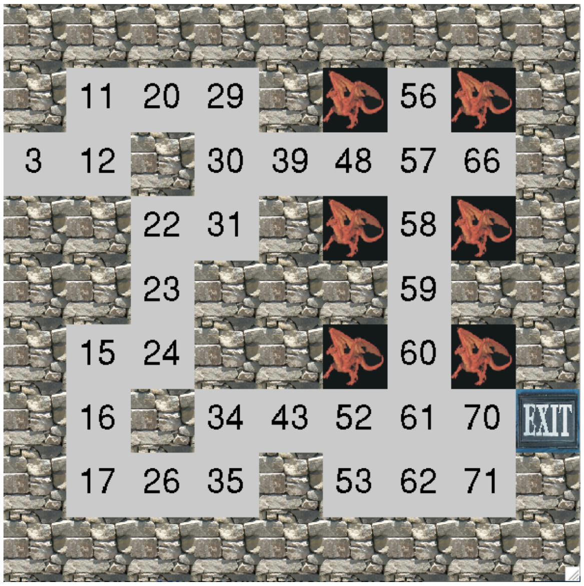
In this problem, you will use value and policy iteration to find the optimal policy of the MDP demonstrated in class. This MDP has $|\mathcal{S}| = 81$ states, $|\mathcal{A}| = 4$ actions, and discount factor $\gamma = 0.99$. Download the ASCII files on the course web site that store the transition matrices and reward function for this MDP. The transition matrices are stored in a sparse format, listing only the row and column indices with non-zero values; if loaded correctly, the rows of these matrices should sum to one.

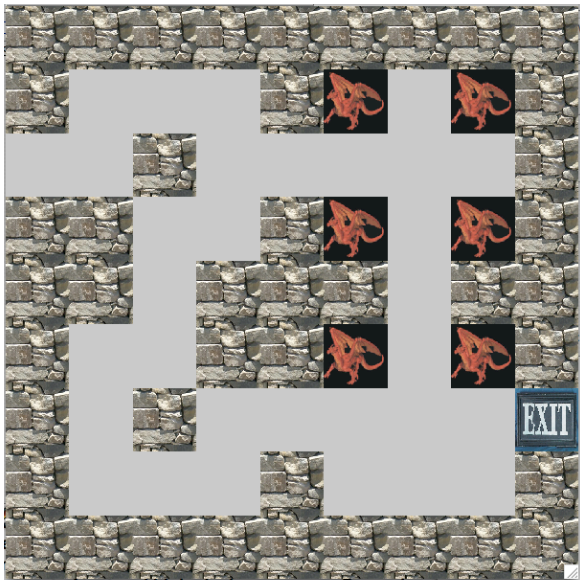
(a) Compute the optimal state value function $V^*(s)$ using the method of value iteration. Print out a list of the non-zero values of $V^*(s)$. Compare your answer to the numbered maze shown below. The correct value function will have positive values at all the numbered squares and negative values at the all squares with dragons.

(b) Compute the optimal policy $\pi^*(s)$ from your answer in part (a). Interpret the four actions in this MDP as (probable) moves to the WEST, NORTH, EAST, and SOUTH. Fill in the correspondingly numbered squares of the maze with arrows that point in the directions prescribed by the optimal policy. Turn in a copy of your solution for the optimal policy, as visualized in this way.

(c) Compute the optimal policy $\pi^*(s)$ using the method of policy iteration. For the numbered squares in the maze, does it agree with your result from part (b)?

(d) **Turn in your source code to all the above questions.** As usual, you may program in the language of your choice.





7.3 Effective horizon time

Consider a Markov decision process (MDP) whose rewards $r_t \in [0,1]$ are bounded between zero and one. Let $h = (1-\gamma)^{-1}$ define an *effective* horizon time in terms of the discount factor $0 \leq \gamma < 1$. Consider the approximation to the (infinite horizon) discounted return,

$$r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + \gamma^4 r_4 + \ldots,$$

obtained by neglecting rewards from some time t and beyond. Recalling that $\log \gamma \leq \gamma - 1$, show that the error from such an approximation decays exponentially as:

$$\sum_{n \geq t} \gamma^n r_n \;\leq\; h e^{-t/h}.$$

Thus, we can view MDPs with discounted returns as similar to MDPs with finite horizons, where the finite horizon $h = (1-\gamma)^{-1}$ grows as $\gamma \to 1$. This is a useful intuition for proving the convergence of many algorithms in reinforcement learning.

7.4 Convergence of iterative policy evaluation

Consider an MDP with transition matrices $P(s'|s,a)$ and reward function $R(s)$. In class, we showed that the state value function $V^\pi(s)$ for a fixed policy π satisfies the Bellman equation:

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

For $\gamma < 1$, one method to solve this set of linear equations is by iteration. Initialize the state value function by $V_0(s) = 0$ for all states s. The update rule at the k^{th} iteration is given by:

$$V_{k+1}(s) \;\leftarrow\; R(s) + \gamma \sum_{s'} P(s'|s,\pi(s)) V_k(s').$$

Use a contraction mapping to derive an upper bound on the error

$$\Delta_k = \max_s |V_k(s) - V^\pi(s)|$$

after k iterations of the update rule. Your result should show that the error Δ_k decays exponentially fast in the number of iterations, k, and hence that $\lim_{k\to\infty} V_k(s) = V^\pi(s)$ for all states s.

7.5 Value function for a random walk

Consider a Markov decision process (MDP) with discrete states $s \in \{0,1,2,\ldots,\infty\}$ and rewards $R(s) = s$ that grow linearly as a function of the state. Also, consider a policy π whose action in each state either leaves the state unchanged or yields a transition to the next highest state:

$$P(s'|s,\pi(s)) \;=\; \begin{cases} \frac{3}{4} & \text{if} \quad s' = s \\[1em] \frac{1}{4} & \text{if} \quad s' = s+1 \\ 0 & \text{otherwise} \end{cases}$$

Intuitively, this policy can be viewed as a right-drifting random walk. As usual, the value function for this policy, $V^\pi(s) = \mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t) \middle|\, s_0 = s\right]$, is defined as the expected sum of discounted rewards starting from state s.

(a) Assume that the value function $V^\pi(s)$ satisfies a Bellman equation analogous to the one in MDPs with finite state spaces. Write down the Bellman equation satisfied by $V^\pi(s)$.

(b) Show that one possible solution to the Bellman equation in part (a) is given by the linear form $V^\pi(s) = as + b$, where a and b are coefficients that you should express in terms of the discount factor γ. (*Hint*: substitute this solution into both sides of the Bellman equation, and solve for a and b by requiring that both sides are equal for all values of s.)

(*) *Challenge (optional, no credit):* justify that the value function $V^\pi(s)$ has this linear form. In other words, rule out other possible solutions to the Bellman equation for this policy.

7.6 Explore or exploit

Consider a Markov decision process (MDP) with discrete states $s \in \{1, 2, \ldots, n\}$, binary actions $a \in \{0, 1\}$, deterministic rewards $R(s)$, and transition probabilities:

$$\begin{aligned} P(s'|s, a\!=\!0) &= \frac{1}{n} \text{ for all } s', \\ P(s'|s, a\!=\!1) &= \begin{cases} 1 & \text{for } s\!=\!s', \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

In this MDP, the "explore" action $a = 0$ leads to uniformly random exploration of the state space, while the "exploit" action $a\!=\!1$ leaves the agent in its current state.

The simple structure of this MDP allows one to calculate its value functions, as well as the form of its optimal policy. This problem guides you through those calculations.

(a) **Local reward-mining**

As usual, the state-value function $V^\pi(s)$ is defined as the expected sum of discounted rewards starting from state s at time $t\!=\!0$ and taking action $\pi(s_t)$ at time t:

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

Consider the subset of states $\mathcal{S}_1 = \{s | \pi(s)\!=\!1\}$ in which the agent chooses the "exploit" action. Compute $V^\pi(s)$ for states $s \!\in\! \mathcal{S}_1$ by evaluating the expected sum of discounted rewards.

(b) **Bellman equation for exploration**

Consider the subset of states $\mathcal{S}_0 = \{s | \pi(s)\!=\!0\}$ in which the agent chooses the "explore" action. Write down the Bellman equation satisfied by the value function $V^\pi(s)$ for states $s \!\in\! \mathcal{S}_0$. Your answer should substitute in the appropriate transition probabilities from these states.

(c) **State-averaged value function**
Assume that the reward function, averaged over the state space, has zero mean: $\frac{1}{n}\sum_s R(s) = 0$.
Let $\mu = \frac{1}{n}|\mathcal{S}_0|$ denote the fraction of states in which the agent chooses to explore.
Let $r = \frac{1}{n}\sum_{s\in\mathcal{S}_1} R(s)$ denote the average reward in the remaining states.
Let $v = \frac{1}{n}\sum_s V^\pi(s)$ denote the expected return if the initial state is chosen uniformly at random.

Combining your results from parts (a-b), show that the state-averaged expected return v is given by:

$$\boxed{v \;=\; \frac{\gamma r}{(1-\gamma)(1-\gamma\mu)}}$$

(d) **Value of exploration**
Using the results from parts (b) and (c), express the state-value function $V^\pi(s)$ for states $s\in\mathcal{S}_0$ in terms of the reward function $R(s)$, the discount factor γ, and the quantities r and μ.

(e) **Optimal policy**
The form of this MDP suggests the following intuitive strategy: in states with low rewards, opt for random exploration at the next time step, while in states with high rewards, opt to remain in place, exploiting the current reward for all future time. In this problem, you will confirm this intuition.
Starting from the observation that an optimal policy is its own greedy policy, infer that the optimal policy π^* in this MDP takes the simple form:

$$\pi^*(s) \;=\; \begin{cases} 1 & \text{if} \;\; V^*(s) \;>\; \theta \\ 0 & \text{if} \;\; V^*(s) \;\leq\; \theta \end{cases}$$

where the threshold θ is a constant, independent of the state s. As part of your answer, express the threshold θ in terms of the state-averaged optimal value function $v^* = \frac{1}{n}\sum_s V^*(s)$.

7.7 Stochastic approximation

In this problem, you will analyze an incremental update rule for estimating the mean $\mu = \mathrm{E}[X]$ of a random variable X from samples $\{x_1, x_2, x_3, \ldots\}$. Consider the incremental update rule:

$$\mu_k \leftarrow \mu_{k-1} + \alpha_k(x_k - \mu_{k-1}),$$

with the initial condition $\mu_0 = 0$ and step sizes $\alpha_k = 1/k$.

(a) Show that the step sizes $\alpha_k = 1/k$ obey the conditions (i) $\sum_{k=1}^{\infty}\alpha_k = \infty$ and (ii) $\sum_{k=1}^{\infty}\alpha_k^2 < \infty$, thus guaranteeing that the update rule converges to the true mean. (You are not required to prove convergence.)

(b) Show that for this particular choice of step sizes, the incremental update rule yields the same result as the sample average: $\mu_k = (1/k)(x_1 + x_2 + \ldots + x_k)$.
