## CSE 250a. Assignment 9

**Out:** *Tue Nov 23*
**Due:** *Thu Dec 02* (by 3:30 PM, Pacific Time, via gradescope)
**Grace period:** *none*.
**Reading:** *Sutton & Barto*, Chapters 3.1-4.4

### 9.1 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.

### 9.2 Three-state, two-action MDP

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

| $s$ | $R(s)$ |
|---|---|
| 1 | -15 |
| 2 | 30 |
| 3 | -25 |

| $s$ | $s'$ | $P(s'\|s, a=\uparrow)$ |
|---|---|---|
| 1 | 1 | $\frac{3}{4}$ |
| 1 | 2 | $\frac{1}{4}$ |
| 1 | 3 | 0 |
| 2 | 1 | $\frac{1}{2}$ |
| 2 | 2 | $\frac{1}{2}$ |
| 2 | 3 | 0 |
| 3 | 1 | 0 |
| 3 | 2 | $\frac{3}{4}$ |
| 3 | 3 | $\frac{1}{4}$ |

| $s$ | $s'$ | $P(s'\|s, a=\downarrow)$ |
|---|---|---|
| 1 | 1 | $\frac{1}{4}$ |
| 1 | 2 | $\frac{3}{4}$ |
| 1 | 3 | 0 |
| 2 | 1 | 0 |
| 2 | 2 | $\frac{1}{2}$ |
| 2 | 3 | $\frac{1}{2}$ |
| 3 | 1 | 0 |
| 3 | 2 | $\frac{1}{4}$ |
| 3 | 3 | $\frac{3}{4}$ |

(a) **Policy evaluation**

Consider the policy $\pi$ that chooses the action shown in each state. For this policy, solve the linear system of Bellman equations (by hand) to compute the state-value function $V^\pi(s)$ for $s \in \{1, 2, 3\}$. Your answers should complete the following table. (*Hint:* the missing entries are integers.) **Show your work for full credit.**

| $s$ | $\pi(s)$ | $V^\pi(s)$ |
|---|---|---|
| 1 | $\uparrow$ | -18 |
| 2 | $\uparrow$ | |
| 3 | $\downarrow$ | |

(b) **Policy improvement**

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. **Show your work for full credit.**

| $s$ | $\pi(s)$ | $\pi'(s)$ |
|---|---|---|
| 1 | $\uparrow$ | |
| 2 | $\uparrow$ | |
| 3 | $\downarrow$ | |

---

## 9.3 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 $\pi$ whose action in each state either leaves the state unchanged or yields a transition to the next highest state:

$$P(s'|s, \pi(s)) = \begin{cases} \frac{2}{3} & \text{if} \quad s' = s \\ \frac{1}{3} & \text{if} \quad s' = s+1 \\ 0 & \text{otherwise} \end{cases}$$

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 $s$, with discount factor $0 \leq \gamma < 1$.

(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 $\gamma$. (*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 $s$.)

---

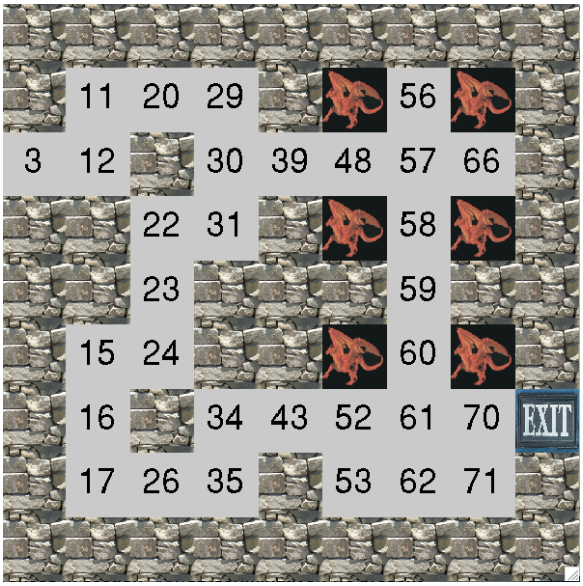
## 9.4 Policy and value iteration

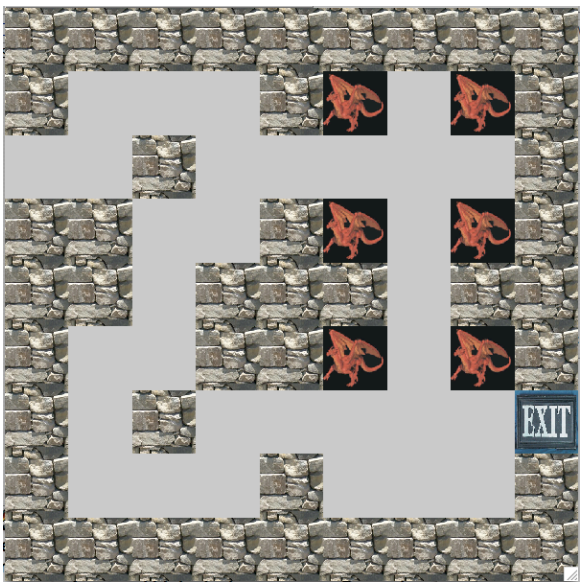
In this problem, you will use policy and value iteration to find the optimal policy of the MDP demonstrated in class. This MDP has $|\mathcal{S}| = 81$ states, $|\mathcal{A}| = 4$ actions, and discount factor $\gamma = 0.9925$. Download the ASCII files on the course web site that store the transition 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 policy $\pi^*(s)$ and optimal value function $V^*(s)$ of the MDP using the method of *policy iteration*. (i) Examine the non-zero values of $V^*(s)$, and compare your answer to the numbered maze shown below. The correct solution will have positive values at all the numbered squares and negative values at all the squares with dragons. Fill in the correspondingly numbered squares of the maze with your answers for the optimal value function. Turn in a copy of your solution for $V^*(s)$ as visualized in this way. (ii) Interpret the four actions in this MDP as (attempted) moves to the WEST, NORTH, EAST, and SOUTH. Fill in the correspondingly numbered squares of the maze (on a separate print-out) with arrows that point in the directions prescribed by the optimal policy. Turn in a copy of your solution for $\pi^*(s)$ as visualized in this way.

(b) Compute the optimal state value function $V^*(s)$ using the method of *value iteration*. For the numbered squares in the maze, does it agree with your result from part (a)? (It should.) Use this check to make sure that your answers from value iteration are correct to at least two decimal places.

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



## 9.5 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 $\pi$ 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 $k^{\text{th}}$ 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 $\Delta_k$ 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 $s$.

## 9.6 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 $\alpha_k = \frac{1}{k}$.

(a) Show that the step sizes $\alpha_k = \frac{1}{k}$ 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: $\mu_k = \frac{1}{k}(x_1 + x_2 + \ldots + x_k)$. *Hint:* use induction.

## 9.7 Course evaluation

Please complete the online course evaluation forms. The forms are completely anonymous, and the summaries of evaluations are not made available to instructors until after grades are submitted. The results will play an important role in the future of the course—whether it continues to be offered in the fall, whether it continues to be allocated resources for two sections, whether it continues to be open to students outside CSE, whether it continues to be taught by the same instructor, etc.