Deep Reinforcement Learning (Fall 2024)
Assignment 1

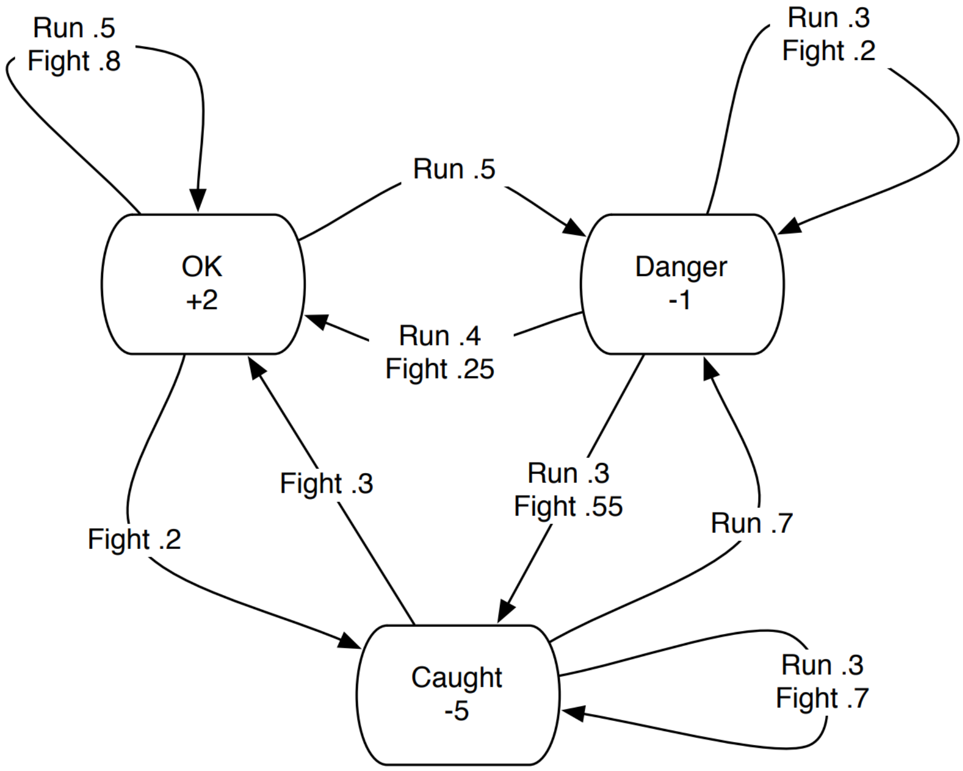
September 5, 2024

1 MDPs [15 pts]

A boy is being chased around the school yard by bullies and must choose whether to Fight or Run.

- There are three states:
  - Ok (O), where he is fine for the moment.
  - Danger (D), where the bullies are right on his heels.
  - Caught (C), where the bullies catch up with him and administer noogies.
- He begins in state O 75% of the time.
- He begins in state D 25% of the time.

The graph of the MDP is given here:



1. Fill out the table with the results of value iteration with a discount factor $\gamma = 0.9$ [9 pts]:

k	$J^k(O)$	$J^k(D)$	$J^k(C)$
1	2	-1	-5
2			
3			

2. At $k = 2$ with $\gamma = 0.9$ what policy would you select? Is it necessarily true that this is the optimal policy? At $k = 3$ what policy would you select? Is it necessarily true that this is the optimal policy? [6 pts]

2 Value Iteration Theorem [35 pts]

In this problem, we will deal with contractions and fixed points and prove an important result from the value iteration theorem. From lecture, we know that the Bellman backup operator B given below is a contraction with the fixed point as V^*, the optimal value function of the MDP. The symbols have their usual meanings. γ is the discount factor and $0 \leq \gamma < 1$. In all parts, $||v||$ is the infinity norm of the vector.

$$(BV)(s) = \max_a [R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \tag{1}$$

We also saw the contraction operator B_π which is the Bellman backup operator for a particular policy given below:

$$(B_\pi V)(s) = \mathbb{E}_{a \sim \pi}[R(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \tag{2}$$

(a) Recall that $||BV - BV'|| \leq \gamma ||V - V'||$ for two random value functions V and V'. Prove that B_π is also a contraction mapping: $||B_\pi V - B_\pi V'|| \leq \gamma ||V - V'||$. [5 pts]

(b) Prove that the fixed point for B_π is unique. What is the fixed point of B_π? [5 pts]

In value iteration, we repeatedly apply the Bellman backup operator B to improve our value function. At the end of value iteration, we can recover a greedy policy π from the value function using the equation below:

$$\pi(s) = \arg\max_a [r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a)V(s')] \tag{3}$$

Suppose we run value iteration for a finite number of steps to obtain a value function V (V has not necessarily converged to V^*). Say now that we evaluate our policy π obtained using the formula above to get V^π. **Note that here and for the rest of Q2, π refers to the greedy policy.**

(c) Is V^π always the same as V? Justify your answer. [5 pts]

In lecture, we learned that running value iteration until a certain tolerance can bring us close to recovering the optimal value function. Let V_n and V_{n+1} be the outputs of value iteration at the n^{th} and $n+1^{th}$ iterations respectively. Let $\epsilon > 0$ and consider the point in value iteration such that $||V_{n+1} - V_n|| < \frac{\epsilon(1-\gamma)}{2\gamma}$. Let π be the greedy policy given the value function V_{n+1}. You will now prove that this policy π is ϵ-optimal. This result justifies why halting value iteration when the difference between success iterations is sufficiently small, ensures the decision policy obtained by being greedy with respect to the value function, is near-optimal.
Precisely if

$$||V_{n+1} - V_n|| < \frac{\epsilon(1-\gamma)}{2\gamma} \tag{4}$$

then,

$$||V^\pi - V^*|| \leq \epsilon. \tag{5}$$

(d) When π is the greedy policy, what is the relationship between B and B_π? [2 pts]

(e) Prove that $||V^\pi - V_{n+1}|| \leq \epsilon/2$.
Hint: Introduce an in-between term and leverage the triangle inequality. [6 pts]

(f) Prove $||B^k V - B^k V'|| \leq \gamma^k ||V - V'||$ [3 pts]

(g) Prove that $||V^* - V_{n+1}|| \leq \epsilon/2$. [7pts]
Hints: Note that $||V^* - V_{n+1}|| = ||V^* + V_{n+2} - V_{n+2} - V_{n+1}||$ and you can repeatedly apply this trick. It may also be useful to leverage part (f) and recall that V^* is the fixed point of the contraction B.

(h) Use the results from parts (e) and (g), to show that $||V^\pi - V^*|| \leq \epsilon$ [2 pts]

3 Simulation Lemma and Model-based Learning [25 pts]

Consider two finite horizon MDPs $\mathcal{M} := \{\mathcal{S}, \mathcal{A}, H, r_h, P_h, s_0\}$ and $\tilde{\mathcal{M}} := \{\mathcal{S}, \mathcal{A}, H, r_h, \tilde{P}_h, s_0\}$. Consider an arbitrary stochastic stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. In finite horizon MDPs, state value is also a function of steps the agent have taken so far, and we denote $V_h^\pi(s)$ as the expected value following π starting from s at step h. We define $V^\pi = V_0^\pi(s_0)$ as the expected total reward of π under $\mathcal{M}$, and $\tilde{V}^\pi = \tilde{V}_0^\pi(s_0)$ as the expected total reward of π under $\tilde{\mathcal{M}}$. We denote $\mathbb{P}_h^\pi$ as the state-action distribution of π under $\mathcal{M}$ at step h.

(a) Prove the following equality:

$$V^\pi - \tilde{V}^\pi = \sum_{h=0}^{H-1} \mathbb{E}_{s,a\sim\mathbb{P}_h^\pi}\left[\mathbb{E}_{s'\sim P_h(\cdot|s,a)}\tilde{V}_{h+1}^\pi(s') - \mathbb{E}_{s'\sim\tilde{P}_h(\cdot|s,a)}\tilde{V}_{h+1}^\pi(s')\right]$$

Here $\tilde{V}_{h+1}^\pi(s')$ is the expected total reward of π under $\tilde{\mathcal{M}}$ from time $h+1$.[10 pts]

(b) Imagine the following situation: we have $\mathcal{M}$ as the true real MDP, and we have $\tilde{\mathcal{M}}$ as some learned model that supposes to approximate the real MDP $\mathcal{M}$. Given $\tilde{\mathcal{M}}$, the natural thing to do is to compute the optimal policy under $\tilde{\mathcal{M}}$, i.e.,

$$\tilde{\pi}^\star = \arg\max_{\pi \in \Pi} \tilde{V}^\pi,$$

where $\Pi \subset \{\pi : \mathcal{S} \to \Delta(\mathcal{A})\}$ is a pre-defined policy class. Let us also denote the true optimal policy $\pi^\star$ under the real model as:

$$\pi^\star = \arg\max_{\pi \in \Pi} V^\pi.$$

A natural question is that what is the performance of $\tilde{\pi}^\star$ under the real model $\mathcal{M}$, compared to $\pi^\star$ under $\mathcal{M}$? To answer this question, let's assume $r(s,a) \in [0,1]$ for all s, a and prove the following inequality:

$$V^{\pi^\star} - V^{\tilde{\pi}^\star} \leq H \sum_{h=0}^{H-1} \left[\mathbb{E}_{s,a \sim \mathbb{P}_h^{\pi^\star}} \|\tilde{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\|_1 + \mathbb{E}_{s,a \sim \mathbb{P}_h^{\tilde{\pi}^\star}} \|\tilde{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \right]$$

(Hint: $\int |f(x)g(x)|dx \leq \|f(\cdot)\|_1 \|g(\cdot)\|_\infty$) [15 pts]

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)** 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. [10 pts]

(b) **(coding)** 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. [10 pts]

(c) **(written)** 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? [5 pts]