CS 747 (Autumn 2021): End-semester Examination

Instructor: Shivaram Kalyanakrishnan

To be submitted by 11.55 p.m., November 21, 2021

Note. Provide justifications/calculations/steps along with each answer to illustrate how you arrived at the answer. You will not receive credit for giving an answer without sufficient explanation.

Submission. Write down your answer by hand, then scan and upload to Moodle. Write clearly and legibly. Begin the answer to any question on a new page. Be sure to mention your roll number.

Question 0. Have you read the instructor's message with subject "End-semester Examination", announced through Moodle on November 7, 2020? Have you followed the rules laid out in that message, in letter and in spirit? Specify related observations or comments, if any. [It is mandatory for you to answer this question.]

Question 1. This question pertains to an abstraction of games such as football and hockey, specifically to work out under what circumstances a team must play aggressively or defensively. To simplify our analysis, we assume that our team, denoted "agent" $\mathcal{A}$, plays against a fixed, static "opponent" $\mathcal{O}$, whose behaviour does not depend on time. Concretely, assume that the game is played for H steps for some $H \geq 1$: this means $\mathcal{A}$ takes a total of H actions. The two actions available to $\mathcal{A}$ are a (aggressive) and d (defensive). At most one goal can be scored in each step.

- If $\mathcal{A}$ plays action a, it has a probability p^a_+ of scoring a goal, a probability p^a_- of conceding a goal, and a probability $p^a_=$ of neither team scoring, where $p^a_+, p^a_-, p^a_= \in (0,1)$, $p^a_+ + p^a_- + p^a_= = 1$.
- If $\mathcal{A}$ plays action d, it has a probability p^d_+ of scoring a goal, a probability p^d_- of conceding a goal, and a probability $p^d_=$ of neither team scoring, where $p^d_+, p^d_-, p^d_= \in (0,1)$, $p^d_+ + p^d_- + p^d_= = 1$.

At the end of H steps, $\mathcal{A}$ is awarded 2 points if it has scored more goals than $\mathcal{O}$, and 0 points if it has scored fewer goals than $\mathcal{O}$. In case of a draw (wherein $\mathcal{A}$ and $\mathcal{O}$ have an equal number of goals), $\mathcal{A}$ earns 1 point.

1a. In summary, the problem instance you are given is the tuple $(p^a_+, p^a_-, p^a_=, p^d_+, p^d_-, p^d_=, H)$. Given such an instance, your aim is to compute behaviour for $\mathcal{A}$ that maximises its expected score. To that end, formulate an MDP based on the problem instance, arguing that an optimal policy for the MDP will describe the optimal behaviour we seek for $\mathcal{A}$. [3 marks]

1b. Describe a procedure to compute an optimal policy for the MDP you have defined in part a. To obtain full marks, show that the number of bitwise operations performed by your procedure is upper-bounded by a polynomial in H (treating as *constants* the number of bits used to encode the probabilities associated with the actions). Bitwise operations include reading and writing a bit; performing addition, subtraction, or comparison on a pair of bits; negating a bit, and so on. [2 marks]

1c. Assume that, as is commonly observed in practice, aggressive play has a strictly higher chance of scoring, but also of conceding a goal. That is,

$$p^a_+ > p^d_+;$$
$$p^a_- > p^d_-.$$

Under this assumption, describe the conditions under which a is an optimal action when there is only one step left in the game (that is, after $H-1$ steps have elapsed). [3 marks]

Question 2. Agents $\mathcal{A}$ and $\mathcal{O}$ are engaged in a matrix game, with $\mathcal{A}$ having actions a_1, a_2, and $\mathcal{O}$ having actions o_1 and o_2. The agents play for T rounds, $T \geq 1$. In each round, both players simultaneously declare their actions; one player is given a reward of 1 and the other a reward of 0. The table below gives the probabilities of $\mathcal{A}$ receiving a 1-reward (equivalently, of $\mathcal{O}$ receiving a 0-reward) for different pairs of actions played.

	o_1	o_2
a_1	1/3	1/2
a_2	2/3	1/4

A game results in a history

$$a^0, o^0, r^0, a^1, o^1, r^1, \ldots, a^{T-1}, o^{T-1}, r^{T-1},$$

where for $0 \leq t \leq T-1$, $\mathcal{A}$ and $\mathcal{O}$ pick a^t and o^t, respectively, and r^t (taken as $\mathcal{A}$'s reward) is drawn from a Bernoulli distribution whose parameter is fixed by a^t and o^t as given in the table. The players may play strategies that depend on history, or are memoryless (but in general still *mixed* or stochastic).

2a. Suppose $\mathcal{O}$ plays a memoryless strategy $q \in [0,1]$, which is shorthand for saying that on each round, it plays o_1 with probability q and o_2 with probability $1-q$. What is the maximum expected reward that $\mathcal{A}$ can obtain by playing some fixed strategy (which could be memoryless or history-dependent) against $\mathcal{O}$'s q-strategy for T rounds? Your answer can be in terms of q; denote it $R_T^\star$ (to be used in part b). [2 marks]

2b. Now suppose $\mathcal{A}$ knows that $\mathcal{O}$ is playing a memoryless strategy, but it does not know which strategy (that is, it does not know q). Can $\mathcal{A}$ play a strategy $\pi_\mathcal{A}$ so as to converge to optimal play against $\mathcal{O}$'s q-strategy for any arbitrary choice $q \in [0,1]$? π_A can be history-dependent, but it must not depend on q. However, suppose it does play against $\mathcal{O}$'s q-strategy and obtains an expected aggregate reward of R_T in T rounds. The question is whether it is possible to achieve

$$\lim_{T\to\infty} \frac{R_T}{R_T^\star} = 1$$

for all $q \in [0,1]$. Prove either that there exists $\pi_\mathcal{A}$ satisfying this notion of optimality, or that there is no strategy $\pi_\mathcal{A}$ for which the result holds. [2 marks]

2c. If $\mathcal{A}$ plays a memoryless strategy p (picking a_1 with probability p and a_2 with probability $1-p$), and $\mathcal{O}$ plays a q-strategy, with $p, q \in [0,1]$, what is $\mathcal{A}$'s expected aggregate reward in T rounds? [1 mark]

2d. Let us denote your answer to part c as $f(p,q,T)$. Work out the value of

$$\max_{p\in[0,1]} \min_{q\in[0,1]} f(p,q,T).$$

What does this value signify? [3 marks]

Question 3. In this question, we investigate relationships between policies for $\mathcal{M}$, the family of MDPs of the form (S, A, T, R, γ) in which the set of actions $A = \{0, 1\}$. Assume that the MDPs in $\mathcal{M}$ encode continuing tasks, and have $\gamma < 1$.

For policy $\pi : S \to A$, let $\text{IS}(\pi)$ denote the set of improvable states of π (defined the Week 6 lecture). If $\text{IS}(\pi) \neq \emptyset$, a policy $\pi' : S \to A$ is said to be a *locally-improving* policy of π if

1. for $s \in S$, $\pi'(s) \neq \pi(s) \implies s \in \text{IS}(\pi)$, and
2. there exists $s \in S$ such that $\pi'(s) \neq \pi(s)$.

In alternative terms, π' is a locally-improving policy of π if any legal policy improvement operation (also presented in the Week 6 lecture) on π yields π'. For any policy π with $\text{IS}(\pi) \neq \emptyset$, let $\text{LI}(\pi)$ denote the set of all locally-improving policies of π.

3a. For this part, we restrict our attention to the family of MDPs $\mathcal{M}' \subset \mathcal{M}$ whose transitions are all deterministic (meaning T associates every state-action pair with a single next state). Now consider the following statement G_1.

> G_1: If π is a non-optimal policy for $M \in \mathcal{M}'$, then there exists an optimal policy $\pi^\star$ for M such that $\pi^\star \in \text{LI}(\pi)$.

In other words, G_1 says that in deterministic 2-action MDPs, one can always reach an optimal policy from a non-optimal policy with just one step of policy improvement (although in general, the corresponding choice of policy improvement might not be known). Is G_1 true or false? Provide a proof. [4 marks]

3b. For this part, we consider the entire family of 2-action MDPs $\mathcal{M}$ and the statement G_2.

> G_2: If π is a non-optimal policy for $M \in \mathcal{M}$, then there exists a policy $\pi' \in \text{LI}(\pi)$ such that for all $\pi'' \in \text{LI}(\pi), \pi' \succeq \pi''$.

G_2 says that among all the locally-improving policies of π, there exists one that dominates or is equal to every other. Is G_2 true or false? Provide a proof. [4 marks]

To prove that G_1 or G_2 is true, you must have a working that holds for all qualifying M and π. To show either of them false, a single counterexample (combination of M and π) will suffice.

Question 4. This question revisits REINFORCE, which was the subject of your most recent weekly quiz. An agent interacts with a deterministic episodic MDP (S, A, T, R). Indeed the transition function T is known to the agent. Hence, when it is at state $s \in S$, the agent can consider each possible action $a \in A$ and compute the next state $s' = T(s, a)$ that it will reach. No discounting is used in defining values.

The agent executes a stochastic policy π parameterised by a d-dimensional parameter vector $w \in \mathbb{R}^d$ for some $d \geq 1$. States are represented by features $\phi_i : S \to \mathbb{R}$, $1 \leq i \leq d$, and the parameter vector being optimised is $w = (w_1, w_2, \ldots, w_d)$. Following policy π_w, suppose the agent encounters the trajectory

$$s^0, a^0, r^0, s^1, a^1, r^1, \ldots, s^H,$$

where s^H is a terminal state. Also suppose that the agent updates its policy parameters from w to w' by performing a REINFORCE update based on this episode, with learning rate $\alpha > 0$. There is no baseline subtraction. The pseudocode below performs the update, but leaves it to you to fill out the steps to calculate the appropriate gradient g. The two parts of the question specify two different choices of π_w; for each you must provide the lines in the pseudocode to correctly set g_i.

```
Q ← 0.
For i = 1, 2, ..., d:
    Δ_i ← 0.
For t = H − 1, H − 2, ..., 0:
    Q ← Q + r^t.
    For i = 1, 2, ..., d:
        //Fill these lines any way you want,
        //using new variables if needed,
        //so that g_i is set correctly.
        //Use as many lines as needed.
        _____
        _____
        _____
        _____
        Δ_i ← Δ_i + g_i · Q.
For i = 1, 2, ..., d:
    w'_i ← w_i + αΔ_i.
```

For each part below, show your derivation to obtain g_i. Thereafter, you only need to provide pseudocode for the blanks above; no need to repeat the portions that are already filled out. In your pseudocode, you can use the states, actions, and rewards encountered, and also T, ϕ, and w.

4a. The probability of taking $a \in A$ from $s \in S$ is obtained by performing a soft-max operation on a linear evaluation of $T(s, a)$. Concretely,

$$\pi_w(s, a) = \frac{e^{w \cdot \phi(T(s,a))}}{\sum_{b \in A} e^{w \cdot \phi(T(s,b))}}.$$ [3 marks]

4b. For $x \in \mathbb{R}$, let $\sigma(x)$ denote the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. When in state $s \in S$, the agent employs the following random process to select $a \in A$. (1) It chooses $a \in A$ uniformly at random. (2) With probability $\sigma(w \cdot \phi(T(s, a)))$ it returns a as the action to take. (3) If no action is returned, the agent goes back to step 1. [4 marks]