Advanced topics in machine learning: Final Exam

Instructor: Morteza H. Chehreghani

Due: See Canvas

- NOTE 1. You explain your solutions, for any question that the calculations are needed you must show the steps (for anything more advanced than +-*/ which you can do with a calculator). The exceptions are the mulitple-choice questions.
- NOTE 2. You must submit your solution to Canvas, in the same way as the assignments.
- NOTE 3. The exam must be done individually. You may not receive help from anyone else.
- NOTE 4. Your submission must be in pdf format. You may either type your solutions in latex/word and submit a pdf file, or take the photo/scanning of the handwritten solutions and upload the pdf file. If you take photos, make sure that it is easy to read and that you combine photos into a single pdf file such that each page appears in the right order. There are both command line and online tools to do this.
- NOTE 5. Read the questions carefully such that you do not miss any question and ensure you clearly give the answer required for each (sub)question.
- NOTE 6. You do not need to write code for any question. Your submitted solution should not include any code.
- See Canvas for the grading formula.
- For questions contact:
  - Q1: Emilio, Niklas and Morteza
  - Q2: Niklas and Morteza
  - Q3: Emilio and Morteza
  - Q4: Emilio and Morteza
  - Q5: Niklas and Morteza

1. (16 points) Multiple choice questions

   These are multiple choice questions belonging to the exam, found on the quiz page on canvas. For each question you have to select **all options that are true for full score**.

2. (8 points) Regret of stochastic bandit algorithms

   Consider the stochastic multi-armed bandit setting with IID rewards.

   (a) (5 points)It is easy to see that a greedy algorithm applied to a stochastic bandit problem has linear regret in the horizon T. Consider the ϵ-greedy algorithm with *fixed* ϵ. In other words, in each round the algorithm selects an arm uniformly at random with probability ϵ and otherwise greedily the best arm according to the current estimated expected rewards. Explain why this algorithm also has linear regret, even though it performs both exploration and exploitation.

(b) (3 points) Explain how the UCB1 algorithm balances exploration and exploitation.

3. (12 points) Policy evaluation and value iteration
Consider an MDP with 2 actions (0 and 1) and 3 states (0, 1 and 2).
If the agent takes action 0, the agent moves from state $s = i$ to state $s = \min(i + 1, 2)$.
If the agent takes action 1, the agent moves from state $s = i$ to state $s = 0$.
The agent obtains the reward 1 when taking any action in state $s = 0$ and 5 if taking action 0 in state $s = 2$, the reward is zero everywhere else. Let the discount factor $\gamma = 1$.

   (a) (8 points) Calculate the state values when taking two steps according to the policy π, action 0 with probability 0.25 and action 1 otherwise, starting from each state (i.e., perform policy evaluation for two steps).

   (b) (4 points) Explain how you could use value iteration to obtain the optimal policy for this task.

4. (10 points) Mixture of policies
Consider an RL problem with an arbitrary set of states and two actions 0 and 1. Also, consider policy π^a which always selects action 0 and π^b which always takes action 1. They have corresponding state values $V^a(s)$ and $V^b(s)$ for the respective reinforcement learning task.
Now, consider a mixed policy as $\pi^c = \alpha\pi^a + (1 - \alpha)\pi^b, \alpha \in (0, 1)$.

(a) (5 points) Give an example (draw, illustrate or write down) of an MDP that shows that π^c will visit states that neither π^a or π^b can reach.

(b) (5 points) Use your example to show/argue why $V^c(s)$ is not bounded by the values of $V^a(s)$ and $V^b(s)$ (its values does not have to lay between $V^a(s)$ and $V^b(s)$). Are there any limits to what values $V^c(s)$ can take?

5. (14 points) Regret of Thompson Sampling for Bayesian Bandits
Consider the K-armed Bayesian bandit problem with a mean vector $\mu \in [0, 1]^K$ drawn from some known prior distribution, as well as arm rewards in $\{0, 1\}$. You are given the following confidence radius and bounds (where T is the horizon, $n_t(a)$ is the number of times arm a has been played until round t, and $\bar{\mu}_t(a)$ is the average reward observed for arm a until round t):

$$r_t(a) = \sqrt{\frac{2\log(T)}{n_t(a)}}$$
$$\mathrm{UCB}_t(a) = \bar{\mu}_t(a) + r_t(a)$$
$$\mathrm{LCB}_t(a) = \bar{\mu}_t(a) - r_t(a)$$

   (a) (6 points) In the proof for Lemma 5.11 in the [Bandits] book, we can see that the Bayesian regret BR_t for Thompson Sampling suffered in round t can be upper bounded such that:

$$\mathrm{BR}_t \leq \mathbb{E}[\mathrm{UCB}_t(a_t) - \mathrm{LCB}_t(a_t)] + \mathbb{E}\left[[\mathrm{LCB}_t(a_t) - \mu(a_t)]^+\right] + \mathbb{E}\left[[\mu(a^*) - \mathrm{UCB}_t(a^*)]^+\right]$$

   Remember that $x^+ = 0$ if $x \leq 0$, and $x^+ = |x|$ otherwise. Also note that $a^* = \max_a \mu(a)$. Explain each of the three terms in the inequality above.

   (b) (8 points) Assume that the following inequality holds for any fixed arm a and round t. Which one of the three terms on the right side of the inequality in (a) can be bounded by using the following inequality? Show how that term can be bounded.

$$\mathbb{E}\left[\mu(a) - \mathrm{UCB}_t(a) \,|\, \mu(a) \geq \mathrm{UCB}_t(a)\right] \cdot \Pr\left\{\mu(a) \geq \mathrm{UCB}_t(a)\right\} \leq \frac{2}{TK}$$

   Hint: For a random variable X, an event E and the complement of the event E^c, we have $\mathbb{E}[X] = \mathbb{E}[X \,|\, E] \cdot \Pr\{E\} + \mathbb{E}[X \,|\, E^c] \cdot \Pr\{E^c\}$. Also, remember that a^* and a_t are both considered random variables, while the above inequality only holds for a fixed arm a.