CSC_52081_EP Advanced Machine Learning and Autonomous Agents
Quizz week 9 (bandits)

Deadline: Wednesday March 19, 2025, 12h00 Paris time (midday, not midnight)

Instructions: Submit your answers by putting them at the end of the lab09's notebook, following the instructions that you see there. The deadline for the quiz is the same as the deadline for the code submission, there is nothing separate to do for the quiz than put the answers at the end of the notebook and submit the notebook by the deadline.

For each question, you need to check the correct statements. There may be 0, 1, or multiple correct statements; checked incorrect statements bring negative score.

Question 1

Consider a stochastic bandit with 3 Bernoulli arms of (unknown) means 0.3, 0.4, and 0.5 and a time horizon n.

a. The greedy strategy (or follow-the-leader strategy) has regret that grows logarithmically with n.

b. The uniform exploration strategy has regret that grows linearly with n.

c. On any given run, the regret at time n is higher with the greedy strategy (or follow-the-leader strategy) than with the uniform exploration strategy.

d. On any given run, the regret at time n is lower with the greedy strategy (or follow-the-leader strategy) than with the uniform exploration strategy.

e. The average of the regret at time n over many runs is lower with the greedy strategy (or follow-the-leader strategy) than with the uniform exploration strategy.

f. The variance of the regret at time n over many runs is lower with the greedy strategy (or follow-the-leader strategy) than with the uniform exploration strategy.

Question 2

Consider the same stochastic bandit problem as in the previous question.

a. The UCB algorithm has regret that grows sub-linearly with n.

b. The Thompson sampling algorithm has regret that grows sub-linearly with n.

c. If n is very large, the expected regret at time n will be lower for the UCB algorithm than for both the greedy and uniform exploration strategies.

d. The UCB algorithm always picks the arm that has the highest current empirical mean estimate.

e. The UCB algorithm always picks the arm that has been sampled the least so far.

f. The Thompson sampling algorithm always picks the arm that has the highest current empirical mean estimate.

Question 3

Consider an adversarial bandit problem with k arms and a time horizon n. Assume that the rewards can be chosen by the adversary in $[0, 1]$. At each time step, the learner receives bandit feedback, that is, observes the reward for the action chosen at that step.

a. The Exp3 algorithm with appropriate parameters has regret $O(\sqrt{nk \log(k)})$.

b. The UCB algorithm with appropriate parameters has regret $O(\sqrt{nk \log(k)})$.

c. No deterministic algorithm can achieve sublinear regret.

d. The Hedge algorithm with appropriate parameters has regret $O(\sqrt{n \log(k)})$.

Question 4

Check the correct inequalities. In all of them, x is a real number.

a. $\exp(x) \leq 1 + x + x^2$ for all $x \leq 1$

b. $\exp(x) \leq 1 + x + \frac{x^2}{2}$ for all $x \leq 1$

c. $\exp(x) \leq 1 + x + \frac{x^2}{2}$ for all $x \geq 0$

d. $\exp(x) \leq 1 + x$ for all x

e. The bound $\exp(x) \leq 1 + x + \frac{x^2}{2}$ is a tighter upper bound than $\exp(x) \leq 1 + x + x^2$ for $x \leq 0$.

Question 5

Consider an adversarial bandit problem with k arms and a time horizon n. Assume that the rewards can be chosen by the adversary in $[0, 1]$. Denote by $t = 1, 2, \cdots$ the current time step. At each time step t, the learner receives bandit feedback, that is, observes the reward for the action chosen at that step.

a. Assume that the learner uses the Exp3 algorithm with learning rate $\eta = \sqrt{\log(k)/(nk)}$. Then the regret at time n is upper bounded by $2\sqrt{nk \log(k)}$.

b. Assume that the learner uses the Exp3 algorithm with learning rate $\eta = \sqrt{\log(k)/(nk)}$. Then the regret at any time $m \geq 1$ is upper bounded by $2\sqrt{mk \log(k)}$.

c. Assume that the learner uses the Exp3 algorithm with time-dependent learning rate $\eta_t = \sqrt{\log(k)/(tk)}$. Then the regret at time n is upper bounded by $2\sqrt{nk \log(k)}$.

d. Assume that the learner uses the Exp3 algorithm with time-dependent learning rate $\eta_t = \sqrt{\log(k)/(tk)}$. Then the regret at time $m \geq 1$ is upper bounded by $2\sqrt{mk \log(k)}$.

Question 6

Let X be a uniform random variable on $[0, 1]$.

a. $E[exp(X)] \leq 1.87$

b. $E[exp(X)] \geq 1.86$

c. $E[exp(X)] \leq 1$

d. $E[exp(X)] \geq 0$

e. $E[exp(X)] \leq 2.7$

f. $E[exp(X)] \geq 2.7$

Question 7

Consider the setting of an adversarial bandit, where the adversary chooses for time t a reward x_{ti} for each arm i. We use the notation of the lecture (A_t, P_{ti}, X_t) and assume that $x_{ti} \in [0, 1]$.

Let $u \in [0, 1/2)$. Let $\hat{X}_{ti} = u - \frac{1_{A_t=i}}{P_{ti}}(u - X_t)$ for any arm i be an estimator of x_{ti}.

a. $\hat{X}_{ti}$ is an unbiased estimator of x_{ti}

b. $\hat{X}_{ti}$ is a biased estimator of x_{ti}

c. For a fixed P_{ti}, the variance of $\hat{X}_{ti}$ is minimal when $x_{ti} = u$.

d. For a fixed P_{ti}, the variance of $\hat{X}_{ti}$ is minimal when $x_{ti} = 0$.

e. For a fixed P_{ti}, the variance of $\hat{X}_{ti}$ is maximal when $x_{ti} = 1$.

f. For a fixed P_{ti}, when $x_{ti} = 0$, the variance of $\hat{X}_{ti}$ is higher than the variance of the loss-based estimator (i.e., the estimator $\hat{X}_{ti}$ for which we would set $u = 1$).