Stochastic Control
Sequential Learning Algorithms
End semester examination: Apr 2022

Instructions

1. There are nine questions which we will treat as homework+examination. Write down the solutions to all of these and upload them on Moodle before noon on Wednesday, 27 Apr.
2. If you need to make any additional assumptions, you may do so but state them explicitly.
3. You are not allowed to collaborate on any of the solutions with any other person.
4. You can consult your notes or any other text but not explicitly seek out solutions for these problems.

Questions

1. For this question, consider the worst case *learning from expert advice* setting. Specifically, there are k experts, and $y_{i,t} \in [0, 1]$ denotes the loss experienced by Expert i at time t. At each time $t \in [n]$,

   - The algorithm picks an expert $A_t \in [k]$ (using only information gathered until time t)
   - The loss of each expert i, $y_{i,t}$ is revealed,
   - The algorithm experiences loss $y_{A_t,t}$.

   Note that the problem instance is defined by the adversarially chosen table of costs $\boldsymbol{y} = (y_{i,t}, 1 \leq i \leq k,\ 1 \leq t \leq n)$.

   Now, suppose that we define the following stronger notion of regret associated with your (possibly randomized) algorithm A :

   $$\bar{R}_n(A) = \sup_{\boldsymbol{y}} \mathbb{E}\left[\sum_{t=1}^{n} y_{A_t,t} - \sum_{t=1}^{n}\left(\min_{1 \leq 1 \leq k} y_{i,t}\right)\right].$$

   Here the expectation is with respect to any randomization performed by the algorithm.

   (a) Why is $\bar{R}_n(A)$ a stronger notion of regret compared to the one analysed in class?

   (b) Prove that for any algorithm,

   $$\bar{R}_n(A) \geq n\left(1 - \frac{1}{k}\right).$$

   This means that sub-linear regret (under this stronger notion of regret) is impossible!

2. Define, for $p, q \in (0,1)$,

$$d(q,p) = q \log(q/p) + (1-q)\log[(1-q)/(1-p)].$$

Note that $d(q,p)$ is the KL divergence between two Bernoulli distributions having means q and p. It is also referred to as the *binary relative entropy* function.

Prove that for fixed q,

(a) $d(q,p)$ is a convex function of p,

(b) $d(q,p)$ attains its minimum (with respect to p) at $p = q$, with $d(q,q) = 0$,

(c) $\lim_{p \downarrow 0} d(q,p) = \lim_{p \uparrow 1} d(q,p) = \infty$.

3. Suppose that S_n is a $\text{Binomial}(n,p)$ random variable (i.e., it represents the number of heads seen over n independent tosses of a biased coin, where the probability of a heads on each toss equals p). For $a \in (0,p)$, show that

$$\mathsf{Prob}(S_n \leq na) \leq e^{-n\gamma},$$

where $\gamma = d(a,p)$. Here, $d(\cdot,\cdot)$ is the binary relative entropy function defined in Question 2.

4. Consider the setup of Question 3. Define $\hat{p} := \frac{S_n}{n}$. Prove that for $\gamma > 0$,

$$\mathsf{Prob}(d(\hat{p},p) \geq \gamma,\ \hat{p} \leq p) \leq e^{-n\gamma}.$$

Further, for $\gamma > 0$, defining $U(\gamma) = \max\{p' \in (0,1) \ :\ d(\hat{p},p') \leq \gamma\}$, show that

$$\mathsf{Prob}(p \geq U(\gamma)) \leq e^{-n\gamma}.$$

5. The goal of this question is specialize the UCB algorithm analysed in class to Bernoulli instances.

Specifically, suppose it is known that each arm of an MAB instance has a Bernoulli reward distribution. Adapt the UCB algorithm analysed in class to this specialized instance and derive the corresponding regret bound.

Hint: Use the result of Question 4.

6. You are given a (biased) coin; on each toss of this coin, the probability of Heads is known to be either μ_1 or μ_2. Your goal is to identify, with probability $\geq 1-\delta$, the true bias μ of the coin (either μ_1 or μ_2) using the minimum number of coin tosses. Note that the accuracy threshold δ and the bias possibilities μ_1 and μ_2 are given to you beforehand.

Using the tools learnt in this course, you devise a policy π for this purpose; note that the policy performs a random number of coint tosses T, and then stops (as per some stopping criterion) and reports its guess $\hat{\mu}$ (either μ_1 or μ_2) of the true coin bias. Let $\mathbb{P}_{\mu_1}$ and $\mathbb{P}_{\mu_2}$ denote, respectively, the probability measures on the observations corresponding to coin bias μ_1 and μ_2, under policy π. You may assume that your policy 'stops' with probability 1, i.e., $\mathbb{P}_{\mu_1}(T < \infty) = \mathbb{P}_{\mu_2}(T < \infty) = 1$. Moreover, your policy is 'sound,' i.e., $\mathbb{P}_{\mu_1}(\hat{\mu} \neq \mu_1) \leq \delta$ and $\mathbb{P}_{\mu_2}(\hat{\mu} \neq \mu_2) \leq \delta$.

(a) Prove that

$$D(\mathbb{P}_{\mu_1}, \mathbb{P}_{\mu_2}) = \mathbb{E}_{\mu_1}[T] d(\mu_1, \mu_2).$$

(b) Next, show that

$$\mathbb{E}_{\mu_1}[T] \geq \frac{1}{d(\mu_1, \mu_2)} \log\left(\frac{1}{4\delta}\right).$$

Similarly, show that

$$\mathbb{E}_{\mu_2}[T] \geq \frac{1}{d(\mu_2, \mu_1)} \log\left(\frac{1}{4\delta}\right).$$

(c) Interpret the above information theoretic lower bounds.

7. Consider a *conservative* variant of the UCB algorithm with rewards in $[0, 1]$, say CCB. In CCB, there is an initial round-robin phase. At each time t after this initial phase, CCB plays the arm with highest lower confidence bound on its mean reward, i.e.,

$$A_t = \arg\max\left(\hat{\mu}_i(t) - \sqrt{\frac{2\log t}{N_i(t)}}\right),$$

where $\hat{\mu}_i(t)$ and $N_i(t)$ are the observed reward sample mean and the number of plays from arm i upto (and not including) time t, respectively. Provide explicit arguments about the achievable regret for this policy over a horizon T.

8. Recall the notion of a conjugate prior: For a given likelihood function if the prior $\mathsf{Prob}(\theta)$ and the posterior $\mathsf{Prob}(\theta|x)$ have the same algebraic form, then $\mathsf{Prob}(\theta)$ is the conjugate prior for the given likelihood function. We have seen that the Beta distribution prior is a conjugate prior for a $\mathrm{Bernoulli}$ likelihood. Show explicitly the following conjugate priors for various likelihoods (sample distributions).

(a) Beta is a conjugate prior for $\mathrm{Binomial}$.

(b) Beta is a conjugate prior for $\mathrm{Geometric}$.

(c) Gamma is a conjugate prior for $\mathrm{Poisson}$. Recall that the Gamma distribution has the following form

$$f(x; \alpha, \beta) = \begin{cases} \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Here $\alpha, \beta > 0$ and the Gamma funnction is given by

$$\Gamma(\beta) = \int_0^{\infty} x^{\beta-1}e^{-x}dx$$

with $\Gamma(\beta) = \beta!$ if β is an integer.

(d) Pareto is a conjugate prior for (continuous) Uniform$(0, \theta)$, $\theta \geq 0$. The Pareto distribution has two parameters, a and $x_,$ and has the following form,

$$f_{RVX}(x) = \begin{cases} \frac{ax_m}{x^{a+1}} & x \geq x_m \\ 0 & x \leq x_m \end{cases}$$

9. Recall the Explore-Then-Commit bandit algorithm in which the algorithm explores all the arms in a round robin fashion, that we studied in class. Consider a the 2-armed bandit with $\mathrm{Bernoulli}$ distributed rewards and parameters (means) $\mu_1, \mu_2 \in [0,1]$. Let T be the time horizon and let ϵT, $0 < \epsilon < 1$ be exploration phase. Let $\Delta = \mu_1 - \mu_2 > 0$.

   (a) Show that there is a choice of ϵ, depending only on the T and not depending on D, under which the regret of the algorithm is bounded above by $c\left(\Delta + \frac{\log T}{\Delta}\right)$ where $c > 0$ is a universal constant

   (b) Now suppose the commitment time is allowed to be data-dependent. This means that the algorithm explores each arm alternately until some condition based on the observations is met, after which it commits to a single arm for the remainder. Design a condition such that the regret of the resulting algorithm can be bounded by $c_1\left(\Delta + \frac{\log T}{\Delta}\right)$ where c_1 is a universal constant. Note that this means thta your condition to end the exploration should only depend on the observed rewards and the time horizon, and not on μ_1, μ_2, Δ.