ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Foundations of Data Science
Fall 2020

Assignment date: Friday, January 29th, 2021, 8:15
Due date: Friday, January 29th, 2021, 11:15

Final Exam – CM1121

This exam is open book. But no electronic devices of any kind are allowed. There are six problems. We do not presume that you will finish all of them. Choose the ones you find easiest and collect as many points as possible. Good luck!

Name: _____

Problem 1	/ 10
Problem 2	/ 10
Problem 3	/ 15
Problem 4	/ 10
Problem 5	/ 10
Problem 6	/ 10
Total	/65

Problem 1. *(These Bandits – 10 pts)*

Consider an adversarial bandit setting with K bandits, where the rewards are arbitrary numbers $x_{t,k} \in [0,1]$ (t stands for the time index and runs from 1 to n and k is the index of the bandit, which goes from 1 to K). You are the adversary, in charge of designing the rewards. You know that the policy that is used is the Exp3 algorithm.

Your task is to fill in the numbers. You are given the constraint that the "average" value of all rewards must be $\frac{1}{2}$, where the "average" means the sum over all $n \times K$ entries divided by $n \times K$.

Your aim is to make the expected reward (not regret) of the player as small as possible.

(i) [5pts] What is the expected reward the player gets, when normalized by the time n? We are only interested in the first order term, i.e., the constant, and not higher terms that vanish with n.

(ii) [5pts] Explain how you fill in the numbers to minimize the expected reward and compute this reward.

Solution: We know that, using the Exp3 algorithm our expected regret (normalized by the time) is sublinear. In other words, we will do almost as good as if we knew the reward matrix and could choose that bandit that has the highest reward, i.e., that row of the matrix whose sum is maximal.

Therefore our task is to make sure that the maximum row sum is as small as possible.

By assumption the "average" value of all rewards is $\frac{1}{2}$. Note that we can compute this average in various ways. In particular we can first compute the reward for each arm and then average over the players. We conclude that the average reward, when averaged over the bandits must be $\frac{1}{2}$. And our aim is to minimize the maximum reward under this average constraint.

Therefore, if we want to minimize the expected reward that the player will get it is best if we make all bandits to have the same expected reward. Hence, fill in all rows (rewards of a bandit) in such a way that its expected value of this row is $\frac{1}{2}$. To first order, the expected reward that the player will get is then $\frac{1}{2}$.

Grading Notes:

Problem 2. *(Estimating Entropy – 15pts)*

You are given n iid samples of a Bernoulli random variable with parameter μ. The parameter is known to be in the range $[\kappa, 1-\kappa]$, where $0 < \kappa \leq \frac{1}{2}$. Let the samples be denoted by $S = \{X_1, X_2, \cdots, X_n\}$, $X_i \in \{0, 1\}$, $i = 1, \cdots n$.

Your task is to estimate the entropy of the underlying distribution accurately. Let h denote the true entropy of the distribution and $\hat{h} = \hat{h}(S)$ be your estimate.

(i) [5pts] Design a scheme to accurately estimate h. Give an explicit epression for $\hat{h}$ as a function of the samples S.

(ii) [5pts] Since the samples S are random your estimate $\hat{h}(S)$ is a random variable. Let $\delta, \epsilon > 0$. Derive a bound of the form

$$\mathbb{P}\{|\hat{h}(S) - h| \geq \epsilon\} \leq \delta.$$

(iii) [5pts] In the expression of (ii) assume that you set δ to some fixed constant. How does the gap ϵ behave as a function of n?

Hint: Simple does it.

Solution:

(i) [5pts] Let $\hat{\mu}(S) = \min\{\max\{\kappa, \frac{1}{n}\sum_{i=1}^{n} X_i\}, \frac{1}{2}\}$. This is essentially just the emperical mean. Since we know the range of the parameter it makes sense to incorporate this knowledge into the estimate. To estimate the entropy we use the plug-in estimator $\hat{h}(S) = h_2(\hat{\mu}(S))$, where $h_2(x) = -x\log_2(x) - (1-x)\log_2(1-x)$.

(ii) [5pts] Each $X_i - \mu$ is subgaussian with parameter $1/4$. Therefore

$$\mathbb{P}\{|\hat{\mu}(S) - \mu| \geq \epsilon\} \leq 2e^{2n\epsilon^2} \;\leftrightarrow\; \mathbb{P}\{|\hat{\mu}(S) - \mu| \geq \sqrt{\frac{\log(\delta/2)}{2n}}\} \leq \delta.$$

By direct computation, $h_2'(x) = \log_2((1-x)/x)$ and $h_2'(x) \leq c_\kappa = \log_2((1-\kappa)/\kappa)$ for $x \in [\kappa, \frac{1}{2}]$.

Therefore,

$$\mathbb{P}\{|\hat{h}(S) - h| \geq c_\kappa\sqrt{\frac{\log(\delta/2)}{2n}}\} \leq \delta.$$

(iii) [5pts] It decays as $1/\sqrt{n}$. Note: In the above bound we took the constant to be c_κ. But we can do better by observing that with probability $1-\delta$ the estimate is at most the given gap away from the true entropy. Therefore, rather than taking the derivative at κ we can take it at $\max\{\kappa, \hat{h} - \sqrt{\frac{\log(\delta/2)}{2n}}\}$. This does not change the speed of convergence but it will in general result in a smaller constant.

Grading Notes: It does not have to be this scheme they describe. As long as it is reasonable and correct we give them full points.

Problem 3. *(Compression: Fibonacci Coding – 15pts)*

Consider the following binary encoding of a positive integer n: $\mathcal{C}_F(n) = I_1 \dots I_r 1$, where $n = \sum_{i=1}^{r} I_i F_{i+1}$ and F_i is i-th Fibonacci number, $F_0 = 0$, $F_1 = 1$, $F_2 = F_0 + F_1 = 1$, $\cdots$, $F_i = F_{i-1} + F_{i-2}$, $i \geq 2$, and $I_i \in \{0, 1\}$. E.g., 1011 denotes the integer $1 \times 1 + 0 \times 2 + 1 \times 3 = 4$.

For every positive integer n such a representation exists. In order to make it unique, given an integer, find the largest Fibonacci number that it contains. Note it and remove its value from the integer. Proceed recursively to find the unique representation. E.g., for $n = 4$, $F_4 = 3$ is the largest Fibonacci number that is contained in 4 and $F_2 = 1$ is the largest Fibonanni number that is contained in $n - F_4 = 1$. This gives us the representation 1011.

Recall that besides a recursive description of the Fibonacci numbers there exists an explicit formula $F_i = \lfloor \frac{\phi^i}{\sqrt{5}} + \frac{1}{2} \rfloor$, where $\phi = \frac{1+\sqrt{5}}{2} \sim 1.618$ is the golden ratio.

(i) [5pts] What is the length of $\mathcal{C}_F(n)$?

(ii) [2pts] Show that the code is prefix-free. *Hint*: use the property of Fibonacci numbers

(iii) [3pts] Show that $\log_\phi(\sqrt{5}i) \leq 3 + 2\log_2 i$.

(iv) [5pts] Consider a random variable U that takes values on the positive integers s.t. $P(U = i)$ is decreasing. Show that $\mathbb{E}[\text{length}(\mathcal{C}_F(U))] \leq 3 + 2H(U)$. *Hint*: first show that $iP(U = i) \leq 1$

Solution:

(i) [5pts] We need to find the largest Fibonacci number that is contained in n. This leads us to consider the largest solution to the equation

$$n \geq \lfloor \frac{\phi^i}{\sqrt{5}} + \frac{1}{2} \rfloor$$

Some thought shows that this has the solution $i_{max} = \lfloor \log_\phi(n\sqrt{5} + \frac{1}{2}) \rfloor$, which gives the length of $\mathcal{C}_F(n)$ to be

$$\lfloor \log_\phi(n\sqrt{5} + \frac{1}{2}) \rfloor + 1.$$

(ii) [2pts] There cannot be two consecutive 1's in $I_1 \dots I_r$ (since $F_{i-1} + F_i = F_{i+1}$). Hence two consecutive 1's appear only at the end of $\mathcal{C}_F(n)$ and the code is prefix-free.

(iii) [3pts] $\log_\phi(\sqrt{5}i) = \frac{\log_2(\sqrt{5}) + \log_2(i)}{\log_2((1+\sqrt{5})/2)}$. From $(1 + \sqrt{5})/2 > \sqrt{2}$ and $\sqrt{5} < 2\sqrt{2}$ we get the desired expression.

(iv) [5pts] If $l_i = r+1$, i is at least F_{r+1} and $i \geq F_{r+1} > \phi^{l_i}/\sqrt{5}$. Therefore $l_i < \log_\phi(\sqrt{5}i) < 3 + 2\log_2 i$. Since $iP(U=i) \leq 1$ (similarly to homework), we have $P(U=i) \leq 1/i$ and therefore $\log_2 P(U=i) \leq -\log_2 i$ and

$$\begin{aligned}\mathbb{E}[\text{length}(\mathcal{C}_F(U))] = \sum_i P(U=i) l_i &\leq 3 + 2\sum_i P(U=i)\log_2 i \\ &\leq 3 - 2\sum_i P(U=i)\log_2 P(U=i) = 3 + 2H(U)\end{aligned}$$

Problem 4. *(Exponential families - 10pts)*

For $t > 0$, Consider a family of distributions supported on $[t, +\infty]$ such that $\mathbb{E}[\ln X] = \frac{1}{\alpha} + \ln t$, $\alpha > 0$.

(i) [5pts] What is the parametric form of a maximum entropy distribution satisfying the constraint on the support and the mean?

(ii) [5pts] Find the exact form of the distribution.

Solution:

(i) [5pts] The maximum entropy distribution has the parametric form $e^{\theta \ln x - A(\theta)} = x^\theta e^{-A(\theta)}$.

(ii) [5pts] Let us first find the value of $A(\theta)$ from the density constraint $\int_t^\infty x^\theta e^{-A(\theta)} dx = 1$. This gives $e^{-A(\theta)} = -\frac{\theta+1}{t^{\theta+1}}$.

Next we find θ from the mean constraint $\int_t^\infty x^\theta e^{-A(\theta)} \ln x \; dx = \frac{1}{\alpha} + \ln t$. This gives $\frac{t^{\theta+1}((\theta+1)\ln t - 1)}{t^{\theta+1}(\theta+1)} = \ln t - \frac{1}{\theta+1} = \frac{1}{\alpha} + \ln t$ and therefore $\theta = -(\alpha + 1)$. The resulting form of the distribution is

$$p(x) = \frac{\alpha t^\alpha}{x^{\alpha+1}}$$

.

Problem 5. *(KL Divergence and L1 – 10pts)*

(i) [5pts] Show that $-D(p||q) \leq \log \sum_i \min(p_i, q_i) + \log \sum_i \max(p_i, q_i)$ *Hint:* The key is to write $\log(q/p)$ in some clever form involving min and max. And if you do not know what to do next with sums involving logs, ask Jensen.

(ii) [5pts] Use the fact that $\min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$ and $\max(a, b) = \frac{a+b}{2} + \frac{|a-b|}{2}$ and (a) to show that $|p - q|_1 \leq 2\sqrt{1 - \exp(-D(p||q))}$

Solution:

(i) [5pts]

$$\begin{aligned}
-D(p\|q) &= \sum_i p_i \log \frac{q_i}{p_i} \\
&= \sum_i p_i \left(\log \left[\min \left(\frac{q_i}{p_i}, 1 \right) \right] + \log \left[\max \left(\frac{q_i}{p_i}, 1 \right) \right] \right) \\
&\leq \log \sum_i p_i \min \left(\frac{q_i}{p_i}, 1 \right) + \log \sum_i p_i \max \left(\frac{q_i}{p_i}, 1 \right) \\
&= \log \sum_i \min(p_i, q_i) + \log \sum_i \max(p_i, q_i)
\end{aligned}$$

(ii) [5pts]

$$\begin{aligned}
&\log \sum_i \min(p_i, q_i) + \log \sum_i \max(p_i, q_i) \\
=& \log \sum_i \left(\frac{p_i + q_i}{2} - \frac{|p_i - q_i|}{2} \right) \\
&+ \log \sum_i \left(\frac{p_i + q_i}{2} + \frac{|p_i - q_i|}{2} \right) \\
=& \log \left[1 - \sum_i \frac{|p_i - q_i|}{2} \right] + \log \left[1 + \sum_i \frac{|p_i - q_i|}{2} \right] \\
=& \log \left[\left(1 - \sum_i \frac{|p_i - q_i|}{2} \right) \cdot \left(1 + \sum_i \frac{|p_i - q_i|}{2} \right) \right] \\
=& \log \left[1 - \left(\frac{1}{2} \sum_i |p_i - q_i| \right)^2 \right]
\end{aligned}$$

The rest goes by rearranging the terms.

Problem 6. *(Signal Representations – 10 pts)* Assume that we get m $u_1, \ldots, u_m$ in $\mathbb{R}^d$. The dimension d is very large. Therefore we would like to compress the data. We fix $n < d$ and we would like to produce n-dimensional representations $\hat{u}_1, \ldots, \hat{u}_m$ that are close to the original ones. Assume that we collect our data samples into a $d \times m$ matrix U and the desired representations into a $n \times m$ matrix $\hat{U}$.

In the course we learned that two possible compression techniques for this scenario is either to use a PCA or random projections.

Recall that random projections are linear maps $f(u) : \mathbb{R}^d \to \mathbb{R}^n$, defined as $f(u) = \frac{1}{\sqrt{n}} Xu$, where X is a real-valued matrix with iid zero-mean unit-variance entries.

(i) [5pts] Assume that your "goodness" criterion is the spectral norm $||U^T U - \hat{U}^T \hat{U}||_2$. What guarantees to you get for both methods? You can assume that the smallest eigenvalue of $X^T X$ is 0.

(ii) [5pts] Assume your "goodness" criterion is $\max_{i,j} |||u_i - u_j||^2 - ||\hat{u}_i - \hat{u}_j||^2|$. What guarantees do you get for both methods? No need for complicated computations.

Solution: Let λ_i be the eigenvalues of matrix $U^T U$. Recall that they are the squared singular values of the matrix U

(i) [5pts] PCA gives the optimal solution by leaving the n largest eigenvalues of $U^T U$ and setting $d - n$ others to zero. For spectral norm this gives $||U^T U - \hat{U}^T \hat{U}||_2 = \lambda_{n+1}^2$. In case of random projections, we have $||U^T U - \hat{U}^T \hat{U}||_2 = ||U^T (I - \frac{1}{n} X^T X) U||_2 \leq ||U^T||_2 ||I - \frac{1}{n} X^T X||_2 ||U||_2 = \lambda_1^2 ||I - \frac{1}{n} X^T X||_2 \approx \lambda_1^2$

(ii) [5pts] For random projections we have a bound of type $\delta \max_{i,j} ||u_i - u_j||^2$. In case of PCA, we have no guarantees on the distance, so the best bound we can come up with is just $\max_{i,j} ||u_i - u_j||^2$.

Grading Notes: