ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Foundations of Data Science	Assignment date: Thursday, November 17th, 2022, 17:15
Fall 2022	Due date: Thursday, November 17th, 2022, 19:00

Midterm Exam – INF1

This exam is open book. No electronic devices of any kind are allowed. There are five problems. Good luck!

Only answers given on this handout count.

Name: ______________________________

Problem 1	/ 2
Problem 2	/ 5
Problem 3	/ 10
Problem 4	/ 10
Problem 5	/ 8
Total	/35

Problem 1 (Subgaussian RV)**.** [2pts; 0.5pts per question] Let X and Y be two independent subgaussian random variables. Which of the following are always subgaussian?

a) $X+Y$ b) X^2 c) XY d) $\begin{cases} 0, & \text{if } |X| \leq 1 \\ \frac{1}{|X|}, & \text{otherwise} \end{cases}$

Solution 1. $X+Y$, i.e., option a), is always subgaussian as we saw in the course. Option d) is bounded in $[0,1]$ and hence it is subgaussian.
Options b) and c), i.e., X^2 and XY are subexponential but not necessarily subgaussian. $XY = (\frac{X+Y}{2})^2 - (\frac{X-Y}{2})^2$ which is the sum of two subexponential random variables.

In class we defined subgaussian random variables to have zero mean. If you wrote either argument you get points here. However, in general, when talking about subgaussianity of a random variable, one looks at the zero mean variable $X - \mathbb{E}[X]$.

Problem 2 (Property Testing: Variance)**.** [5pts] A colleague claims to have implemented an algorithm which outputs i.i.d. samples distributed according to a discrete distribution P that has unit variance. Your task is to design a statistic to test whether this is indeed true.

Let Δ_k be the set of probability distributions on the alphabet $\mathcal{X} = \{1, \cdots, k\}$. Assume that $P \in \mathcal{P} \cup \mathcal{Q}$ with $\mathcal{P} := \{P \in \Delta_k : P \text{ has variance } 1\}$ and $\mathcal{Q} := \{P \in \Delta_k : P \text{ has variance } \in [0, 1-\epsilon] \cup [1+\epsilon, \infty)\}$, where $0 < \epsilon < 1$. You are given n samples $\{X_i\}_{i=1}^n$, where the X_i are independent copies sampled according to P.

Remark: For the following three questions we do not ask you to write down a proof (or explicit calculation) that your proposed solution works.

a) [1pt] We say that an estimator $e : S^n \mapsto \Pi$ on a sample S^n of length n (ϵ, δ)-learns a parameter $p \in \Pi$ if for any $(\epsilon, \delta) \in (0,1)^2$, given sufficiently many samples n, we have that $\mathbb{P}(\{|e(S^n) - p| > \epsilon\}) < \delta$. Give a brief explanation (one sentence, no calculations) why the empirical estimator of the second moment $\hat{\mu}_{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ can (ϵ, δ)-learn the second moment in our setting.

b) [2pts] First, assume that a genie tells you that X has zero mean. Design a simple test statistic and give a threshold in order to check for the above mentioned unit variance property.

   Hint: Use the claim in a).

c) [2pts] Now consider the more general case where X can have arbitrary mean. Again, design a simple test statistic and give a threshold.

   Hint: You can assume that $\hat{\mu}_X^2$ (ϵ, δ)-learns $\mathbb{E}[X]^2$.

Solution 2. a) The empirical mean of squares $\hat{\mu}_{X^2}$ concentrates strongly around the true second moment due to the boundedness of X^2 together with Hoeffding's inequality.

Alternatively: finite expectation of distribution P + (weak) law of large numbers.

b) The approach is analogous to the property test for uniformity as presented in the lecture. We first note that the variance is equal to the second moment. Compute the empirical mean of X^2, i.e., $\hat{\mu}_{X^2} := \frac{1}{n}\sum_{i=1}^n X_i^2$ with sufficiently many samples such that we have that $|\hat{\mu}_{X^2} - var(X)| < \epsilon/2$ with high probability due to Hoeffding's inequality (which applies due to the boundedness of X^2). Then consider the statistic $T(S) := |\hat{\mu}_{X^2} - 1|$, compare it to the threshold $\tau = \epsilon/2$ (with ϵ as given in the definition of $\mathcal{Q}$) and choose $\mathcal{P}$ if we are below the threshold and $\mathcal{Q}$ else. If $P \in \mathcal{P}$, by assumption $|\hat{\mu}_{X^2} - 1| < \epsilon/2$ with high probability and we successfully pick $\mathcal{P}$. If $P \in \mathcal{Q}$, then by assumption $|var(X) - 1| > \epsilon$ and furthermore $|\hat{\mu}_{X^2} - var(X)| < \epsilon/2$. Due to the triangle inequality we have that $|\hat{\mu}_{X^2} - 1| \geq |var(X) - 1| - |\hat{\mu}_{X^2} - var(X)| > \epsilon/2$ with high probability and we correctly select $\mathcal{Q}$. (*Remark: we did not expect the above derivation in the case $P \in \mathcal{Q}$ from you.*)

c) We additionally compute the empirical mean of X, i.e., $\hat{\mu}_X := \frac{1}{n}\sum_{i=1}^n X_i$.

We consider the statistic $T(S) := |\hat{\mu}_{X^2} - \hat{\mu}_X^2 - 1|$, compare it to the threshold $\tau = \epsilon/2$ (with ϵ as given in the definition of $\mathcal{Q}$) and choose $\mathcal{P}$ if we are below the threshold and $\mathcal{Q}$ else. The proof that this strategy works is analogous to the on presented in b).

Remark: the following is a rigorous justification as to why learnability of the first two moments implies learnability of the variance. Again, we did not expect this from you. Assume that we pick sufficiently many samples such that we $(\epsilon/4, \delta/2)$-learn the first two moments. By applying the triangle inequality and union bound we get that

$$\begin{aligned}
\mathbb{P}(\{|\hat{\mu}_{X^2} - \hat{\mu}_X^2 - var(X)| > \epsilon/2\}) &\leq \mathbb{P}(\{|\hat{\mu}_{X^2} - \mu_{X^2}| + |\hat{\mu}_X^2 - \mu_X^2| > \epsilon/2\}) \quad (1)\\
&\leq \mathbb{P}(\{|\hat{\mu}_{X^2} - \mu_{X^2}| > \epsilon/4\} \cup \{|\hat{\mu}_X^2 - \mu_X^2| > \epsilon/4\}) \quad (2)\\
&\leq \mathbb{P}((\{|\hat{\mu}_{X^2} - \mu_{X^2}| > \epsilon/4\}) + \mathbb{P}(\{|\hat{\mu}_X^2 - \mu_X^2| > \epsilon/4\}) \quad (3)\\
&\leq \delta/2 + \delta/2 = \delta \quad (4)
\end{aligned}$$

hence $\hat{\mu}_{X^2} - \hat{\mu}_X^2$ $(\epsilon/2, \delta)$-learns $var(X)$.

Problem 3 (Perfect Secrecy)**.** [10pts]

Alice is trying to communicate a message $M \in \mathcal{M}$ securely to Bob with the help of a shared secret key K. Alice *encrypts* her message into what is called a ciphertext C by computing $C = E_K(M)$. Here, $E_K(M)$ is a known function given the key K that takes the message M as input and outputs C (which in general is in a different alphabet that the message M).

Bob receives the message. He then *decrypts* the ciphertext by computing $M = D_K(C)$. Just as E_K, D_K is a known function for a fixed key K.

A cryptosystem is said to be *perfectly secure* if M and C are independent.

(a) [2pts] What are the values of $H(C|M,K), H(M|C,K)$ and $I(M;C)$ in a perfectly secure cryptosystem?

Solution 3: *(a)* $H(C|M,K) = 0$ since $C = E_K(M)$; $H(M|C,K) = 0$ since $M = D_K(C)$; $I(M;C)$=0 by definition, since M and C are independent.

(b) [3pts] Show that $H(M|C) \leq H(K|C)$ for a perfectly secure cryptosystem.

Solution 3: *(b)*

$$\begin{aligned} H(M|C) &= H(M|C) - H(M|C,K) = I(M;K|C) \\ &= H(K|C) - H(K|M,C) \leq H(K|C). \end{aligned}$$

The last step follows since conditioning decreases entropy.

(c) [3pts] Show that $H(K) \geq \log_2(|\mathcal{M}|))$ for a perfectly secure cryptosystem. *Hint:* First establish an inequality between $H(K)$ and $H(M)$.

Solution 3: We have $H(M) = H(M|C) \leq H(K|C) \leq H(K)$. In the first step we used that $I(M;C) = 0$, the second step uses the result in (b) and the third step follows since conditioning decreases entropy.

Since this has to hold for all distributions of M, taking M to be uniformly distributed, we get $H(K) \geq \log_2(|\mathcal{M}|)$.

(d) [2pts] During Roman times, the Caeser shift cipher was used to transmit messages securely. Every letter in the message was replaced by another letter some fixed number of positions away and the shift was decided by the key. For example, if the key was 3, the letter A would be replaced by D, B by E and so on. To describe this scheme mathematically map each letter to a number $A \to 0, B \to 1, \dots, Z \to 25$. Then,

$$E_K(M) = (M + K) \bmod 26, \quad \text{and} \tag{5}$$

$$D_K(C) = (C - K) \bmod 26. \tag{6}$$

Show that the Caeser shift cipher is perfectly secure if the key K is distributed uniformly over $\mathcal{K} = \{0, 1, \dots, 25\}$ and we are transmitting only a single letter.

Solution 3. *(d)* To show perfect secrecy, we need to show that M and C are independent. Note that M need not be distributed uniformly and could be any arbitrary distribution.

$$\begin{aligned}
P(C = c) &= \sum_{k=0}^{25} P(K = k)P(M = (c - k) \bmod 26) \\
&= \frac{1}{26}\sum_{k=0}^{25} P(M = (c - k) \bmod 26) = \frac{1}{26} \\
P(C = c|M = m) &= P(K = (c - m) \bmod 26|M = m) \\
&= P(K = (c - m) \bmod 26) = \frac{1}{26}
\end{aligned}$$

Problem 4 (Minimum-Norm Solutions)**.** [10pts] Let us consider an *underdetermined* system of linear equations $A\mathbf{x} = \mathbf{b}$, where $A \in \mathbb{R}^{m\times n}$ is a "fat" matrix $(m < n)$ of rank m and where $\mathbf{b}$ is chosen such that a solution exists. In this case, there exist infinitely many solutions. Further, we denote the compact SVD of A by $A = U_{A,\text{red}}\Sigma_{A,\text{red}}V_{A,\text{red}}^T$ where $U_{A,\text{red}} \in \mathbb{R}^{m\times m}, \Sigma_A \in \mathbb{R}^{m\times m}, V_A \in \mathbb{R}^{n\times m}$.

a) [2pts] We first construct a "completed" system of equations

$$\tilde{A}\mathbf{x} = \tilde{\mathbf{b}} \tag{7}$$

with $\tilde{A} := \begin{bmatrix} A \\ B \end{bmatrix} \in \mathbb{R}^{n\times n}$ by appending $n-m$ rows to A. How do we have to choose the rows of B in order to guarantee that (7) has exactly one solution? Which dimensions do the components of the reduced SVD of $B = U_{B,\text{red}}\Sigma_{B,\text{red}}V_{B,\text{red}}^T$ have?

b) [3pts] How do the columns of $V_{B,\text{red}}$ relate to those of $V_{A,\text{red}}$? Restate the SVD of $\tilde{A}$ in terms of $U_{A,\text{red}}, \Sigma_{A,\text{red}}, V_{A,\text{red}}$, $U_{B,\text{red}}, \Sigma_{B,\text{red}}, V_{B,\text{red}}$ and then write down (7) in terms of this expression, also making explicit how $\tilde{b}$ relates to b and B.

c) [2pts] State a simple way to find *some* solution to the original problem through solving (7) for x.

d) [3pts] Finally, prove that the one solution $\mathbf{x}$ that has the minimum 2-norm can be expressed as

$$\mathbf{x}_{MN} \quad = V_{A,red}\Sigma^{-1}_{A,red}U^T_{A,red}\mathbf{b}. \tag{8}$$

Solution 4. a) The system of equations (7) has a unique solution if and only if $\tilde{A}$ has full rank. Since the dimensions of each row vector of A is n, it is possible to add $n-m$ more linearly independent row vectors to A. For the components of the compact SVD of B we have $U_{A,\text{red}} \in \mathbb{R}^{n-m\times n-m}$, $\Sigma_{B,\text{red}} \in \mathbb{R}^{n-m\times n-m}$ and $V_{B,\text{red}} \in \mathbb{R}^{n\times n-m}$.

b) The columns of $V_{B,\text{red}}$ are pairwise orthogonal the columns of $V_{A,\text{red}}$. From this it follows that we can write the SVD of $\tilde{A}$ as

$$\tilde{A} = \begin{bmatrix} U_{A,\text{red}}\Sigma_{A,\text{red}}V_{A,\text{red}}^T \\ U_{B,\text{red}}\Sigma_{B,\text{red}}V_{B,\text{red}}^T \end{bmatrix} = \begin{bmatrix} U_{A,\text{red}} & 0 \\ 0 & U_{B,\text{red}} \end{bmatrix} \begin{bmatrix} \Sigma_{A,\text{red}} & 0 \\ 0 & \Sigma_{B,\text{red}} \end{bmatrix} \begin{bmatrix} V_{A,\text{red}} \\ V_{B,\text{red}} \end{bmatrix}^T. \tag{9}$$

Let $\mathbf{b}_A = \mathbf{b}$ and $\mathbf{b}_B = B\mathbf{x}$, then

$$\begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x} = \begin{bmatrix} U_{A,\text{red}} & 0 \\ 0 & U_{B,\text{red}} \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \mathbf{x} = \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix}. \tag{10}$$

c) We can find a solution to (7) as follows:

1. Pick some arbitrary B such that $\tilde{A}$ has full rank.
2. Compute a solution to the system of equations in b) by choosing some arbitrary $\mathbf{b}_B$ and performing a simple matrix inversion/back-substitution:

$$\mathbf{x} = \tilde{A}^{-1} \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix}. \tag{11}$$

d) The square of the 2-norm of $\mathbf{x}$ is

$$\begin{aligned} ||\mathbf{x}||_2^2 &= \mathbf{x}^T\mathbf{x} && (12)\\ &= \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix}^T \begin{bmatrix} U_{A,\text{red}} & 0 \\ 0 & U_{B,\text{red}} \end{bmatrix} \begin{bmatrix} \Sigma_{A,red} & 0 \\ 0 & \Sigma_{B,red} \end{bmatrix}^{-1} \begin{bmatrix} V_{A,red} \\ V_{B,red} \end{bmatrix}^T && (13)\\ &\cdot \begin{bmatrix} V_{A,red} \\ V_{B,red} \end{bmatrix} \begin{bmatrix} \Sigma_{A,red} & 0 \\ 0 & \Sigma_{B,red} \end{bmatrix}^{-1} \begin{bmatrix} U_{A,\text{red}} & 0 \\ 0 & U_{B,\text{red}} \end{bmatrix}^T \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} && (14)\\ &= ||\Sigma_{A,red}^{-1} U_{A,red}^T \mathbf{b}_A||_2^2 + ||\Sigma_{B,red}^{-1} U_{B,red}^T \mathbf{b}_B||_2^2. && (15) \end{aligned}$$

Thus, the 2-norm of $\mathbf{x}$ achieves its minimum $||\mathbf{b}||_2 = ||\Sigma_{A,red}^{-1} U_{A,red}^T \mathbf{b}_A||_2$, when $||\mathbf{b}_B||_2 = 0$. Also, $||\mathbf{b}_B||_2 = 0$ requires that every entry of $\mathbf{b}_B$ is 0. Hence in such case,

$$\mathbf{x}_{MN} = \begin{bmatrix} V_{A,red} \\ V_{B,red} \end{bmatrix} \begin{bmatrix} \Sigma_{A,red} & 0 \\ 0 & \Sigma_{B,red} \end{bmatrix}^{-1} \begin{bmatrix} U_{A,\text{red}} & 0 \\ 0 & U_{B,\text{red}} \end{bmatrix}^T \begin{bmatrix} \mathbf{b}_A \\ 0 \end{bmatrix} = V_{A,red} \Sigma_{A,red}^{-1} U_{A,red}^T \mathbf{b}. \tag{16}$$

Problem 5. *(Bandits)*[8pts] You commute every single day (weekends included) by bike between EPFL and Lausanne center (we only consider this one direction to keep things simple). There are many different paths you can take: along the lake, bike roads, through the center etc. Let the number of possible paths be K. Further, let's assume that you have two different bicycles. A racing bike and a mountain bike. You are trying to figure out the 'best' combination of route and bike, let's say according to the time it takes you to commute.

Every day you can measure the time. This time of course depends on the route you take and the bike you pick. But lets say that it also depends on whether it rains or not (to keep things simple assume that it rains roughly half the time) and what day of the week it is. In addition many small random and unknown factors influence the time you measure.

a) [2pts] Formulate this as a bandit problem.

b) [2pts] What algorithm do you choose and what regret do you expect to see.

Consider next the following variation. Everything is as before but there are two changes. First, you have an archrival who knows your policy and every day may place various obstacles on some of the paths. Assume that this is the dominant factor that determines the commute time and that the obstacles may increase the time it takes to commute up to a factor 2. Secondly, each day you have only one of the two bikes available to you (perhaps it is a publi-bike or you share).

c) [2pts] Formulate this as a bandit problem.

d) [2pts] What algorithm do you choose and what regret do you expect to see.

Note: There are several reasonable answers and many possible considerations. Hence make sure to justify why you make the choices you make.

Solution 5. a) There are K paths and 2 bikes. Both are choices at your disposal. Hence, there are $2K$ arms. Further we have context. I.e., we have reason to believe that the time it takes for each route depends significantly on the weather (rain or not) and on the day of the week. Therefore, it makes sense to model the problem as a contextual bandit with 14 contexts, where each context is one pair of the form $(\text{weekday}, \text{rain/no rain})$. If we would like to reduce the context, we might want to quantize the days of the week into Sat-Sun and Mo-Fri.

b) We do not have a horizon given. If we know that the randomness for all cases can be captured by subgaussian random variables with a parameter no more than lets say σ^2 then we can run for each context an UCB algorithm. We know from the description that each context appears roughly the same number of times. And for each context the regret grows like $O(n \log(n))$, where n is roughly the number of days divided by 14.

More realistically, we will not know the parameter σ^2 a priori. We can then either find a reasonable upper bound by reasoning: e.g., bikes are never completely stuck in traffic and there are a finite number of paths each of a limited length. So the variance can reasonably be limited to a concrete value. Or we might have to estimate the variance for each branch and use this estimate in the UCB algorithm.

c) As before there are K paths. The bike is no longer your choice. Hence, there are only K arms. Further we have context. I.e., we have reason to believe that the time it takes for each route depends significantly on the weather (rain or not) and on the day of the week as well as on the bike we are given. Therefore, a reasonable choice is to choose contextual bandits with 28 contexts, where each context is a triple of the form $(\text{weekday}, \text{rain/no rain}, \text{bike type})$. If we would like to reduce the context, we might want to quantize the days of the week into Sat-Sun and Mo-Fri. We know that all the rewards are in a fixed range since even with obstacles the commute time never increases by more than a factor 2 and the obstacles are the dominant factors (and not other randomness). Hence we are in an adversarial setting rather than a stochastic setting (albeit still with context).

d) Since we are dealing with an adversarial setting and context it makes sense to split the data according to the context and to run the exp3 algorithm within each context.