ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2023

Assignment date: Thursday, November 16th, 2023, 17:15
Due date: Thursday, November 16th, 2023, 19:00

Midterm Exam – INF2

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

Only answers given on this handout count.

Name: _____

Problem 1	/ 7
Problem 2	/ 8
Problem 3	/ 10
Problem 4	/ 10
Total	/35

Problem 1. *(Sum of binomials)*[7 pts]
You have seen in Homework 2 that the entropy function is related to the asymptotic value of the binomial coefficient:

$$\log_2 \binom{n}{np} = nh(p) + O(\log_2 n),$$

for $n \geq 1$ and $0 \leq p \leq 1$, where $h(p) \triangleq -p \log_2 p - (1-p)\log_2(1-p)$ is the binary entropy function. We want to derive a similar bound for the sum of binomial coefficients.

(a) [3 pts] Fix $0 \leq p \leq 1/2$ and let $\mathcal{C}$ be the set of all subsets of $\{1, 2, \ldots, n\}$ of size at most np. Let X be a random variable uniformly distributed over $\mathcal{C}$. Show that

$$H(X) \leq nh(p).$$

Hint: Let $(X_1, X_2, \ldots, X_n)$ be a random vector such that for every i, $X_i = 1$ if $i \in X$, and $X_i = 0$ otherwise. Argue that $H(X) = H(X_1, X_2, \ldots, X_n)$.

(b) [1 pts] Using part (a), conclude that

$$\sum_{i=0}^{\lfloor np \rfloor} \binom{n}{i} \leq 2^{nh(p)}.$$

(c) [3 pts] Using part (b), show that if $Z \sim \text{Binomial}(n, p = \frac{1}{2})$, then

$$\Pr\left(\left|Z - \frac{n}{2}\right| \geq c\sigma\right) \leq 2^{1-c^2/2}$$

for every $c \geq 0$, where $\sigma = \frac{\sqrt{n}}{2}$ is the standard deviation of Z.

Hint: you can use (without proving it) the bound $h(p) \leq 1 - 2\left(\frac{1}{2} - p\right)^2$.

Solution 1. (a) There is a one-to-one correspondence between X and $(X_1, X_2, \ldots, X_n)$: from the value of X we can uniquely determine the value of $(X_1, X_2, \ldots, X_n)$, and viceversa. Hence, $H(X) = H(X_1, X_2, \ldots, X_n)$. Then,

$$H(X) = H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) = nH(X_1)$$

where the last equality is due to symmetry. Now, $\Pr(X_1 = 1) \leq p \leq \frac{1}{2}$, and therefore $H(X_1) \leq h(p)$. Hence, $H(X) \leq nh(p)$.

(b)

$$H(X) = \log|\mathcal{C}| = \log \sum_{i=0}^{\lfloor np \rfloor} \binom{n}{i} \leq nh(p).$$

Hence,

$$\sum_{i=0}^{\lfloor np \rfloor} \binom{n}{i} \leq 2^{nh(p)}.$$

(c)

$$\begin{aligned}
\Pr\left(\left|Z - \frac{n}{2}\right| \geq c\frac{\sqrt{n}}{2}\right) &= 2\left(\frac{1}{2}\right)^n \sum_{i=0}^{\left\lfloor n\left(\frac{1}{2} - \frac{c}{2\sqrt{n}}\right)\right\rfloor} \binom{n}{i} \\
&\leq 2^{nh\left(\frac{1}{2} - \frac{c}{2\sqrt{n}}\right) - n + 1} \\
&\leq 2^{n\left(1 - \frac{c^2}{2n}\right) - n + 1} \\
&= 2^{1 - c^2/2}.
\end{aligned}$$

Problem 2. *(Geometrical interpretation of mutual information)*[8 pts]
In Homework 2 we introduced the conditional KL divergence between two probability kernels $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \to \mathcal{Y}$ given a distribution P_X over $\mathcal{X}$ as

$$D(P_{Y|X}\|Q_{Y|X}|P_X) \triangleq \sum_{x\in\mathcal{X}} P_X(x)D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x)),$$

where for every $x \in \mathcal{X}$, $D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x))$ is the standard KL divergence between the two distributions $P_{Y|X}(\cdot|x)$ and $Q_{Y|X}(\cdot|x)$ over $\mathcal{Y}$.

(a) [2 pts] Let X and Y be two random variables with joint distribution $P_{XY} = P_X P_{Y|X}$. Show that

$$I(X;Y) = \sum_{x\in\mathcal{X}} P_X(x)D(P_{Y|X}(\cdot|x)\|P_Y)$$

, where P_Y is the marginal distribution of Y. This formula shows that the mutual information can be interpreted as a weighted average of the distances between the conditional distributions $P_{Y|X}(\cdot|x)$ and the marginal distribution P_Y.

(b) [3 pts] Show that for any distribution Q_Y on $\mathcal{Y}$,

$$I(X;Y) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y).$$

You can think of this formula as a KL equivalent of the classical $I(X;Y) = H(Y) - H(Y|X)$.

(c) [3 pts] Show that

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X).$$

According to this formula, the minimizing Q_Y can be interpreted as the "center of gravity" of the conditional distributions $P_{Y|X}(\cdot|x)$, and the mutual information as its radius.

Solution 2. All the results can be proved working directly with the definitions of KL divergence and mutual information. The following is a simple solution that makes use of the results proved in Homework 2, Problem 3.

(a)

$$I(X;Y) = D(P_X P_{Y|X} \| P_X P_Y) = D(P_{Y|X} \| P_Y | P_X) = \sum_{x \in \mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x) \| P_Y),$$

where the second inequality is due to Homework 2, Problem 3(b).

(b)

$$\begin{aligned} D(P_Y \| Q_Y) + I(X;Y) &= D(P_Y \| Q_Y) + D(P_{X|Y} \| P_X | P_Y) \\ &= D(P_{XY} \| P_X Q_Y) \\ &= D(P_{Y|X} \| Q_Y | P_X) \end{aligned}$$

where the first inequality is due to part (a) by exchanging the roles of X and Y, the second equality is due to the chain rule of the KL divergence (Homework 2, Problem 3(a)), and the third inequality is again due to Homework 2, Problem 3(b).

(c) By part (b) we know that $I(X;Y) \leq D(P_{Y|X} \| Q_Y | P_X)$ for every Q_Y, since $D(P_Y \| Q_Y) \geq 0$. Hence, $I(X;Y) \leq \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X)$. The equality is achieved by picking $Q_Y = P_Y$, for which $D(P_{Y|X} \| Q_Y | P_X) = D(P_{Y|X} \| P_Y | P_X) = I(X;Y)$.

Problem 3. *(Lipschitz Bandits)*[10 pts]
Assume for the following that you have a bandit algorithm at your disposal that has an expected regret, call it R_n, bounded by $c\sqrt{Kn\log(n)}$, where K is the number of arms and n is the time horizon.

You have to design an algorithm for the following scenario. There are infinitely many bandits. More precisely the bandits are indexed by x, $x \in [0,1]$. Bandit x has mean $\mu(x)$ (which is unknown). But you do know that the various bandits are related in the sense that

$$|\mu(x) - \mu(y)| \leq L|x - y|, \tag{1}$$

where L is a known constant. This is known as the Lipschitz bandit problem due to the Lipschitz condition (1).

A natural approach to such a bandit problem is to discretize the space of bandits. I.e., assume that you pick K positions $0 \leq x_1 < x_2 < \cdots < x_K \leq 1$ and run your given bandit problem on these K bandits.

a) [5 pts] Bound the expected regret as a function of K, n, L and the placement of points.

b) [5 pts] For n and L fixed, minimize your expression with respect to K and the placement of points.

HINT: In order to simplify your computation, you might want to slightly loosen your bound.

Solution 3.

a) [5 pts] Let x^* be the position of the arm with highest reward and let $\mu^* = \mu(x^*)$. Let i^* be the discrete arm that is closest to x^*. Then by the Lipschitz condition

$$\begin{aligned}\mu^* &\leq \mu_{i^*} + L|x_{i^*} - x^*| \\ &\leq \max_i \mu_i + L|x_{i^*} - x^*| \\ &\leq \max_i \mu_i + \frac{1}{2}L \max_{i=1,\cdots,K-1} |x_{i+1} - x_i|.\end{aligned}$$

Hence

$$\begin{aligned}R_n &= \mu^* n - \mathbb{E}[\sum_{t=1}^{n} X_t] \\ &= (\mu^* - \max_i \mu_i)n + \max_i \mu_i n - \mathbb{E}[\sum_{t=1}^{n} X_t] \\ &\leq \frac{1}{2}nL\max_{i=1,\cdots,K-1}|x_{i+1} - x_i| + \sqrt{Kn\log(n)}.\end{aligned}$$

b) [5 pts] We get the tightest bound for $\max_{i=1,\cdots,K-1}|x_{i+1} - x_i|$ if we pick the positions uniform. This will give us $1/(K+1)$. However, to simplify the minimization, let us upper bound this by $1/K$. Hence, we have to take the derivative of $c\sqrt{Kn\log(n)} + \frac{1}{2}L/K$ wrt to K and then set the result to 0 and solve for K. We get $-((L - cK\sqrt{Kn\log(n)})/(2K^2)) = 0$ which gives us (ignoring integer constraints) $K = L^{(2/3)n^{(1/3)}}/(c^{(2/3)}n^{(1/3)}\log(n)^{(1/3)})$. If we plug this back into the expression we arrive at $3/2c^{(2/3)}L^{(1/3)}n^{(2/3)}\log(n)^{(1/3)}$.

Problem 4. *(Frames: Reconstruction Algorithm)*[10 pts]

Frames generalize the notion of orthonormal bases and have important applications in compression, noise reduction, frequency analysis, etc. Formally, a frame is defined as a set of vectors $V = \{v_i\}_{i=1}^m$ in an n-dimensional complex vector space such that there exist constants $0 < A \leq B < \infty$ such that for all vectors x,

$$A\|x\|_2^2 \leq \sum_{k=1}^{m} |\langle x, v_k\rangle|^2 \leq B\|x\|_2^2. \tag{2}$$

We refer to the numbers $\{\langle x, v_k\rangle\}_{i=1}^m$ as *frame coefficients*. Note that frames include orthonormal bases as the special case where $A = B = 1$.

Let us define the *synthesis operator* S associated with a frame V via the following linear mapping:

$$Sx = \sum_{k=1}^{m} \langle x, v_k\rangle v_k. \tag{3}$$

Now imagine that you are given only the frame coefficients (and of course the frame itself) and want to reconstruct the original signal x therefrom. In contrast to the synthesis step for orthonormal bases, it generally does not hold that $x = Sx$ if V is a frame, hence one needs to come up with a dedicated reconstrution algorithm.

One of the simplest such algorithms is the following:

Inputs: $Sx, \{v_i\}_{i=1}^m, A, B$

Initialize: $x_0 \leftarrow 0$

For $k = 1, \ldots, N$**:**

$x_k \leftarrow x_{k-1} + \frac{2}{A+B} S(x - x_{k-1})$

Output: x_N

a) [2 pts] Show that for the spectral norm of self adjoint matrices U (i.e., matrices such that $U^{\mathsf{H}} = U$), it holds that $\|U\| = \sup_{\|x\|_2=1} |\langle x, Ux\rangle|$.

Hint 1: the min-max Theorem states that for the spectrum $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ of $n \times n$ Hermitian matrices $A \in \mathbb{R}^{n\times n}$ it holds that $\lambda_k = \min_W \max_{\|x\|_2=1} \{\langle x, Ax\rangle \mid \dim(W) = n - k + 1\}$, where W are linear subspaces of $\mathbb{R}^n$.

Hint 2: start by using the min-max Theorem to control the spectrum of $\|U\|^2$.

Whenver you use the min-max Theorem, be explicit about how you apply it!

b) [2 pts] Show that $\langle (I - \frac{2}{A+B} S)x, x\rangle \leq \frac{B-A}{B+A}\|x\|_2^2$.

c) [2 pts] Similarly, show that $\langle (I - \frac{2}{A+B} S)x, x\rangle \geq -\frac{B-A}{B+A}\|x\|_2^2$ and show that this implies that $\|I - \frac{2}{A+B} S\| \leq \frac{B-A}{B+A}$.

d) [1 pts] Show that it holds

$$x - x_k = \left(I - \frac{2}{A+B}S\right)(x - x_{k-1}). \tag{4}$$

e) [3 pts] Derive an upper bound on the reconstruction error $\|x - x_N\|_2$ in terms of A, B, N and $\|x\|_2$ that decays geometrically in N.

Which kind of frames allow for to the most efficient signal reconstrution in terms of required iterations of the above algorithm?

Solution 4.

a) Denote by $Q^{\mathsf{H}}DQ$ the spectral decomposition of U. Then, $\|U\|^2 = \max_{\|x\|_2=1} x^{\mathsf{H}}U^{\mathsf{H}}Ux = \max_{\|x\|=1}(Qx)^{\mathsf{H}}D^2(Qx) = (\lambda^*)^2$ where λ^* denotes the eigenvalue of U with maximum modulus. This implies that $\|U\| = |\lambda^*|$, where $|\lambda^*|$ can be found to be $\max_{\|x\|=1} |\langle x, Ux\rangle|$ by combining the min-max Theorem variational descriptions of λ_1 for U and $-U$, respectively.

b) Using a), it follows that

$$\langle (I - \frac{2}{A+B}S)x, x\rangle = \|x\|_2^2 - \frac{2}{A+B}\sum_{k=1}^{m} |\langle x, v_k\rangle|^2, \quad \forall x$$

this implies together with the frame condition that

$$\langle (I - \frac{2}{A+B}S)x, x\rangle \leq \|x\|_2^2 - \frac{2A}{A+B}\|x\|_2^2 = \frac{B-A}{B+A}\|x\|_2^2$$

.

c) A calculation analogous to b) shows that

$$\langle (I - \frac{2}{A+B}S)x, x\rangle \geq -\frac{B-A}{B+A}\|x\|_2^2$$

. Combining this with a) and b) gives the desired result.

d)

$$x - x_k = x - x_{k-1} - \frac{2}{A+B}S(x - x_{k-1}) \tag{5}$$

$$= \left(I - \frac{2}{A+B}S\right)(x - x_{k-1}) \tag{6}$$

e) Repeating the step in d) N times, we obtain

$$x - x_k = \left(I - \frac{2}{A+B}S\right)^N (x - x_0).$$

Using c) together with sub-multiplicativity of the operator norm yields

$$\|x - x_k\|_2 = \|\left(I - \frac{2}{A+B}S\right)^N (x - x_0)\|_2 \tag{7}$$

$$\leq \|I - \frac{2}{A+B}S\|^N \|x - x_0\|_2 \leq \left(\frac{B-A}{B+A}\right)^N \|x\|_2. \tag{8}$$

We can perfectly reconstruct x with just a single iteration of the above algorithm for tight frames, i.e., frames with $A = B$.