ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2021

Assignment date: Thursday, January 27th, 2022, 8:15
Due date: Thursday, January 27th, 2022, 11:15

Final Exam – INM202, INM203

This exam is open book. No electronic devices of any kind are allowed. There are 5 problems. Choose the ones you find easiest and collect as many points as possible. We do not necessarily expect you to finish all of them. Good luck!

Name: _____

Problem 1	/ 10
Problem 2	/ 10
Problem 3	/ 10
Problem 4	/ 10
Problem 5	/ 10
Total	/50

(intentionally left blank)

Problem 1 (Angle Preservation in Johnson-Lindenstrauss Mapping)**.** [10pts]

Consider a set of K points, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^K \subset \mathbb{R}^n$. Recall that in the course we showed that the Johnson-Lindenstrauss mapping $f : \mathbb{R}^n \to \mathbb{R}^m$ preserves distances between points up to a small multiplicative factor.

In this problem, we will study how well the Johnson-Lindenstrauss mapping preserves angles between points in $\mathcal{X}$. As in class, we will consider a mapping of the form $f_H(\mathbf{x}) = \frac{1}{\sqrt{m}} H\mathbf{x}$, where each entry of H is drawn i.i.d from $\mathcal{N}(0, 1)$.

Let us denote by θ_{ijk} the angle between $\mathbf{x}_{ij} = \mathbf{x_i} - \mathbf{x_j}$ and $\mathbf{x}_{kj} = \mathbf{x_k} - \mathbf{x_j}$. We will study how this angle compares to the angle after the Johnson-Lindenstrauss mapping, $\tilde{\theta}_{ijk}$, i.e., to the angle between $f_H(\mathbf{x}_{ij})$ and $f_H(\mathbf{x}_{kj})$.

Ideally, for every triplet i, j, k, we want something of the form

$$(1 \mp \alpha_\epsilon) \cos\theta_{ijk} - \beta_\epsilon(|\mathbf{x}_{ij}|, |\mathbf{x}_{kj}|) \leq \cos\tilde{\theta}_{ijk} \leq (1 \pm \alpha_\epsilon) \cos\theta_{ijk} + \beta_\epsilon(|\mathbf{x}_{ij}|, |\mathbf{x}_{kj}|)$$

where $\alpha_\epsilon \approx 0$ and $\beta_\epsilon(|\mathbf{x}_{ij}|, |\mathbf{x}_{kj}|) \to 0$ as $\epsilon \to 0$ ($\mp$ and $\pm$ because $cos(\theta_{ijk})$ can be positive or negative). There are several ways of proving such a result, but we will derive this bound by controlling the difference between the inner products $\langle f_H(\mathbf{x}_{ij}), f_H(\mathbf{x}_{kj})\rangle$ and $\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle$.

(i) [3pts] Find $g(k)$ such that, if $m > g(k)$, then the probability that H is such that for every triplet i, j, k we have

$$(1-\epsilon)||\mathbf{x}_{ij} \pm \mathbf{x}_{kj}||_2^2 \leq ||f_H(\mathbf{x}_{ij}) \pm f_H(\mathbf{x}_{kj})||_2^2 \leq (1+\epsilon)||\mathbf{x}_{ij} \pm \mathbf{x}_{kj}||_2^2$$

and for every pair i, j we have

$$(1-\epsilon)||\mathbf{x}_{ij}||_2^2 \leq ||f_H(\mathbf{x}_{ij})||_2^2 \leq (1+\epsilon)||\mathbf{x}_{ij}||_2^2$$

is strictly positive. What is the order of $g(k)$?

[Hint : Check Johnson-Lindenstrauss proof if you don't know where to start.]

Solution :
From the lecture notes

$$\mathbb{P}((1-\epsilon)||\mathbf{x} - \mathbf{x}'||_2^2 \leq ||f_H(\mathbf{x}) - f_H(\mathbf{x}')||_2^2 \leq (1+\epsilon)||\mathbf{x} - \mathbf{x}'||_2^2) \geq 1 - 2e^{-\frac{m\epsilon^2}{8}}$$

but instead of having $\binom{k}{2}$ conditions to be fulfilled, now we have $2\binom{k}{3} + \binom{k}{2}$ conditions to be fulfilled. Therefore using the union bound, all the conditions are fulfilled with a probability larger than

$$1 - \left(2\binom{k}{3} + \binom{k}{2}\right) 2e^{-\frac{m\epsilon^2}{8}}$$

Equating the above equation to 0, we conclude that

$$m > \frac{8}{\epsilon^2} \log\left(2\binom{k}{3} + \binom{k}{2}\right)$$

$2\binom{k}{3} + \binom{k}{2}$ is polynomial in k. Therefore, $g(k) \propto \log(k)$.

(ii) [3pts] Show that for an H that fulfills the conditions above we have for every triplet i, j, k

$$|\langle f_H(\mathbf{x}_{ij}), f_H(\mathbf{x}_{kj})\rangle - \langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle| \leq \epsilon(||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2).$$

[Hint : $4\langle \mathbf{x}, \mathbf{x}'\rangle = ||\mathbf{x} + \mathbf{x}'||_2^2 - ||\mathbf{x} - \mathbf{x}'||_2^2$]

Solution : We have,

$$\begin{aligned}
&4\langle f_H(\mathbf{x}_{ij}), f_H(\mathbf{x}_{kj})\rangle \\
&= ||f_H(\mathbf{x}_{ij}) + f_H(\mathbf{x}_{kj})||_2^2 - ||f_H(\mathbf{x}_{ij}) - f_H(\mathbf{x}_{kj})||_2^2 \\
&\leq (1+\epsilon)||\mathbf{x}_{ij} + \mathbf{x}_{kj}||_2^2 - (1-\epsilon)||\mathbf{x}_{ij} - \mathbf{x}_{kj}||_2^2 \\
&= 4\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle + 2\epsilon(||\mathbf{x}_{ij} + \mathbf{x}_{kj}||_2^2 + ||\mathbf{x}_{ij} - \mathbf{x}_{kj}||_2^2) \\
&= 4\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle + 4\epsilon(||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2)
\end{aligned}$$

We can form a similar inequality for the lower bound.

(iii) [2pts] Show that

$$\cos\tilde{\theta}_{ijk} \leq \max\left\{\frac{1}{1-\epsilon}\cos\theta_{ijk} + \frac{\epsilon}{1-\epsilon}D_{ijk}, \frac{1}{1+\epsilon}\cos\theta_{ijk} + \frac{\epsilon}{1+\epsilon}D_{ijk}\right\}$$

and

$$\min\left\{\frac{1}{1-\epsilon}\cos\theta_{ijk} - \frac{\epsilon}{1-\epsilon}D_{ijk}, \frac{1}{1+\epsilon}\cos\theta_{ijk} - \frac{\epsilon}{1+\epsilon}D_{ijk}\right\} \leq \cos\tilde{\theta}_{ijk}.$$

where

$$D_{ijk} = \frac{||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2}{||\mathbf{x}_{ij}||_2||\mathbf{x}_{kj}||_2}.$$

[Hint : Is there a way to relate cosine and inner product?]

Solution : Continuing from the previous point, we have

$$\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle - \epsilon(||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2) \leq \langle f_H(\mathbf{x}_{ij}), f_H(\mathbf{x}_{kj})\rangle \leq \langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle + \epsilon(||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2).$$

Observe that $\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle = |\mathbf{x}_{ij}||\mathbf{x}_{kj}|\cos\theta_{ijk}$. Dividing the inequality by $|f_H(\mathbf{x}_{ij})||f_H(\mathbf{x}_{kj})|$ gives us

$$\frac{\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle - \epsilon(||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2)}{||f_H(\mathbf{x}_{ij})||_2||f_H(\mathbf{x}_{kj})||_2} \leq \cos\tilde{\theta}_{ijk} \leq \frac{\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle + \epsilon(||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2)}{||f_H(\mathbf{x}_{ij})||_2||f_H(\mathbf{x}_{kj})||_2}.$$

The expression on the problem is obtained by choosing the appropriate direction of bound for $||f_H(\mathbf{x}_{ij})||_2||f_H(\mathbf{x}_{kj})||_2$. For example, if $\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle \geq 0$, then,

$$\begin{aligned}
\cos\tilde{\theta}_{ijk} &\leq \frac{1}{1-\epsilon}\frac{\langle \mathbf{x}_{ij}, \mathbf{x}_{kj}\rangle}{||\mathbf{x}_{ij}||_2||\mathbf{x}_{kj}||_2} + \frac{\epsilon}{1-\epsilon}\frac{||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2}{||\mathbf{x}_{ij}||_2||\mathbf{x}_{kj}||_2} \\
&= \frac{1}{1-\epsilon}cos(\theta_{ijk}) + \frac{\epsilon}{1-\epsilon}\frac{||\mathbf{x}_{ij}||_2^2 + ||\mathbf{x}_{kj}||_2^2}{||\mathbf{x}_{ij}||_2||\mathbf{x}_{kj}||_2}
\end{aligned}$$

(iv) [2pts] Find $\mathbf{x}, \mathbf{x}'$ such that,

$$\mathbb{P}\left((1-\epsilon)\langle \mathbf{x}, \mathbf{x}' \rangle \leq \langle f_H(\mathbf{x}), f_H(\mathbf{x}') \rangle \leq (1+\epsilon)\langle \mathbf{x}, \mathbf{x}' \rangle\right) = 0$$

Using this, explain why we cannot get a bound of the form

$$(1 \mp \alpha_\epsilon)\cos\theta_{ijk} \leq \cos\tilde{\theta}_{ijk} \leq (1 \pm \alpha_\epsilon)\cos\theta_{ijk}$$

Solution : Take $\mathbf{x}$ and $\mathbf{x}'$,both non-zero vectors, to be orthogonal to each other, hence the event that we consider is such that

$$\langle f_H(\mathbf{x}), f_H(\mathbf{x}') \rangle = 0.$$

This event is a null-event given how we sample the matrix H.
Therefore, setting x_i, x_j, x_k such that, $x_{ij} = x, x_{kj} = x'$ we get, $\theta_{ijk} = \pm\pi/2 \implies \cos\theta_{ijk} = 0$. Therefore, if we had a bound,

$$(1 \mp \alpha_\epsilon)\cos\theta_{ijk} \leq \cos\tilde{\theta}_{ijk} \leq (1 \pm \alpha_\epsilon)\cos\theta_{ijk},$$

then $\cos\tilde{\theta}_{ijk}$ must be 0. For that, it must be that,

$$\langle f_H(\mathbf{x}), f_H(\mathbf{x}') \rangle = 0.$$

However, this happens with probability 0 over the choice of matrix H.

Problem 2 (KL and its Fenchel-Legendre dual)**.** [10pts] In this problem, we study the Kullback-Leibler divergence $D(P\|Q)$ as a function of P, for fixed Q. Throughout this problem, the logarithm shall be taken as the *natural* logarithm.

(i) [5pts] Let us first assume that P and Q are probability density functions (that is, continuous random variables). Fix Q to be the normal distribution of mean zero and variance σ^2. Let P be arbitrary but with the same second moment as Q. Show that in this case, $D(P\|Q) = h(Q) - h(P)$, that is, the difference of the differential entropy of the normal distribution and the differential entropy of P.

Solution: Let $\mathcal{X} = \{x : p(x) > 0\}$. Then,

$$
\begin{aligned}
D(P\|Q) &= \int_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx && (1)\\
&= -h(P) - \int_{x\in\mathcal{X}} p(x) \log q(x) dx && (2)\\
&= -h(P) - \int_{x\in\mathcal{X}} p(x) \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \right) dx && (3)\\
&= -h(P) - \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \int_{x\in\mathcal{X}} p(x) \frac{x^2}{2\sigma^2} dx && (4)\\
&= -h(P) + \frac{1}{2} \log \left(2\pi\sigma^2\right) + \frac{\mathbb{E}_P[X^2]}{2\sigma^2} && (5)\\
&= -h(P) + \frac{1}{2} \log \left(2\pi\sigma^2\right) + \frac{1}{2} && (6)\\
&= -h(P) + \frac{1}{2} \log \left(2\pi\sigma^2\right) + \frac{1}{2} \log e && (7)\\
&= -h(P) + \frac{1}{2} \log \left(2\pi e\sigma^2\right), && (8)
\end{aligned}
$$

where we recognize the second summand to be exactly the differential entropy of the Gaussian distribution with variance σ^2.

Alternatively, since we assume that second moments are equal, we could have observed that

$$
\begin{aligned}
D(P\|Q) &= -h(P) - \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \int_{x\in\mathcal{X}} p(x) \frac{x^2}{2\sigma^2} dx && (9)\\
&= -h(P) - \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \int_{x\in\mathcal{X}} q(x) \frac{x^2}{2\sigma^2} dx && (10)\\
&= -h(P) - \int_{x\in\mathcal{X}} q(x) \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \right) dx && (11)\\
&= -h(P) - \int_{x\in\mathcal{X}} q(x) \log q(x) dx, && (12)
\end{aligned}
$$

where the second summand is precisely the entropy of $q(x)$.

(ii) [5pts] In this part, for simplicity, we restrict to P and Q being probability mass functions (that is, discrete random variables). For a function $f(P)$, the Fenchel-Legendre dual (also called the convex dual) is defined as $f^*(R) = \sup_P \langle R, P\rangle - f(P)$. More concretely, if P is a probability mass function over a discrete and finite alphabet $\mathcal{X}$, the Fenchel-Legendre dual can be written as

$$f^*(R) = \max_P \left\{ \left(\sum_{x\in\mathcal{X}} R(x)P(x) \right) - f(P) \right\}, \qquad (*)$$

where $R(x)$ is an arbitrary function on the alphabet $\mathcal{X}$.

(a) [3pts] For the case where $f(P) = D(P\|Q)$, show that the optimizing distribution in Equation (**??**) is given by

$$P(x) = \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}}.$$

Hint: Write the Lagrangian for the optimization problem in Equation (**??**), including the constraint that $P(x)$ has to sum to one. Then take derivatives as usual, recalling that we are using the natural logarithm.

(b) [2pts] Show that the Fenchel-Legendre dual of $f(P) = D(P\|Q)$ (for fixed Q) is given by

$$f^*(R) = \log \left(\sum_{x\in\mathcal{X}} Q(x)e^{R(x)} \right),$$

which includes the logarithm of the moment-generating function of Q as a special case (select $R(x) = \lambda x$).

Solution: The Lagrangian is

$$L(\lambda, P) = \left(\sum_{x\in\mathcal{X}} R(x)P(x) \right) - \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} - \lambda \left(\sum_{x\in\mathcal{X}} P(x) - 1 \right) \qquad (13)$$

Taking the derivative with respect to $P(x)$ gives

$$\frac{d}{dP(x)} L(\lambda, P) = R(x) - \log \frac{P(x)}{Q(x)} - 1 - \lambda \qquad (14)$$

Setting this to zero, we find

$$P(x) = Q(x)e^{R(x)-(1+\lambda)}, \qquad (15)$$

where we observe that $P(x)$ is non-negative (which is good). Next, we have to select λ to make the $P(x)$ sum to one, that is

$$e^{-(1+\lambda)} = \frac{1}{\sum_x Q(x)e^{R(x)}}, \tag{16}$$

meaning that the optimizing $P(x)$ is given by

$$P(x) = \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}}. \tag{17}$$

Plugging this particular choice of $P(x)$ back into our main expression, we find

$$f^*(R) = \max_P \left\{ \left(\sum_{x\in\mathcal{X}} R(x)P(x) \right) - f(P) \right\} \tag{18}$$

$$= \sum_{x\in\mathcal{X}} R(x) \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}} - \sum_{x\in\mathcal{X}} \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}} \log \left(\frac{e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}} \right) \tag{19}$$

$$= \sum_{x\in\mathcal{X}} R(x) \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}} - \sum_{x\in\mathcal{X}} \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}} R(x)$$

$$+ \sum_{x\in\mathcal{X}} \frac{Q(x)e^{R(x)}}{\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})}} \log \left(\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})} \right) \tag{20}$$

$$= \log \left(\sum_{\tilde{x}} Q(\tilde{x})e^{R(\tilde{x})} \right). \tag{21}$$

Problem 3 (Tighter Generalization Bound)**.** [10pts] Let $D = X_1, ..., X_n$ iid from an unknown distribution P_X, let $\mathcal{H}$ be a hypothesis space, and $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ be a σ^2-subgaussian loss function for every h. In the lecture we have seen that the generalization error can be upper bounded using the mutual information.

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \sqrt{\frac{2\sigma^2 I(D;H)}{n}}$$

(i) [4 pts] Modify the proof of the *Mutual Information Bound (11.2.2)* to show that if for all $h \in \mathcal{H}$, $\ell(h, X)$ is σ^2-subgaussian in X, then

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \sqrt{\frac{2\sigma^2 \sum_{i=1}^n I(X_i;H)}{n}}.$$

Hint: Recall from the lecture notes that

$$|\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]| \leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{P_{X_i H}} [\ell(H, X_i)] - \mathbb{E}_{P_{X_i} P_H} [\ell(H, X_i)] \right|.$$

Solution:

$$\begin{aligned}
||\mathbb{E}_{P_{DH}} [L_{P_X}(H) - L_D(H)]|| &\leq \frac{1}{n} \sum_{i=1}^n \left| \mathbb{E}_{P_{X_i H}} [\ell(H, X_i)] - \mathbb{E}_{P_{X_i} P_H} [\ell(H, X_i)] \right| \\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_H} \left[\left| \mathbb{E}_{P_{X_i|H}} [\ell(H, X_i)] - \mathbb{E}_{P_{X_i}} [\ell(H, X_i)] \right| \right] \qquad (11.14) \\
&\leq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{P_H} \left[\sqrt{2\sigma^2 D(P_{X_i|H} || P_{X_i})} \right] \qquad (11.12) \\
&\leq \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 \mathbb{E}_{P_H} \left[D(P_{X_i|H} || P_{X_i}) \right]} \qquad (11.15) \\
&= \frac{1}{n} \sum_{i=1}^n \sqrt{2\sigma^2 I(X_i;H)} \qquad (11.15) \\
&\leq \sqrt{\frac{2\sigma^2 \sum_{i=1}^n I(X_i;H)}{n}}
\end{aligned}$$

(ii) [3 pts] Show that, this new bound is never worse than the previous bound by showing that,

$$I(D;H) \geq \sum_{i=1}^n I(X_i;H).$$

Solution:

$$I(D;H) = I(X_1,...,X_n;H) = \sum_{i=1}^{n} I(X_i;H|X^{i-1}) \quad \text{(chain rule for MI)}$$
$$= \sum_{i=1}^{n} I(X_i;HX^{i-1}) \quad \text{(independence of } X_i\text{'s)}$$
$$\geq \sum_{i=1}^{n} I(X_i;H) \quad \text{(chain rule and non-negativity of MI)}$$

Therefore the new upper bound is never larger than the previous upper bound.

(iii) [3 pts] Let us consider an example. Assume that $D = X_1,..,X_n,\ n > 1$, are i.i.d. from $\mathcal{N}(\theta,1)$, and that we do not know θ. We want to learn θ assuming the loss $\ell(h,x) = \min(1,(h-x)^2)$ (which is bounded) and $\mathcal{H} = \mathbb{R}$. Our learning algorithm outputs $H = \frac{1}{n}\sum_{i=1}^{n} X_i$. Use the new bound to show that

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{1}{4(n-1)}}.$$

How does the old bound perform in this example?
Hint: Adding independent gaussian random variables, you get a gaussian random variable.

Solution: Note that the learning algorithm is a deterministic one, that is given a training set D, the learning algorithm outputs a deterministic number. Note also that by property of Gaussian, $H \sim \mathcal{N}(\theta, 1/n)$. Therefore,

$$I(D;H) = h(H) - h(H|D) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e0) = \infty \tag{22}$$

which gives a vacuous bound. Let us compute $I(X_1;H) = h(H) - h(H|X_1)$. Fix x_1, Then,

$$H = \frac{1}{n}x_1 + \frac{1}{n}\sum_{i=2}^{n} X_i \tag{23}$$

which is Gaussian around some mean (which we do not care about) and with variance $(n-1)/n^2$, and note that the variance does not depend on x_1. Therefore the mutual information can be computed as,

$$I(X_1;H) = h(H) - h(H|X_1) = \frac{1}{2}\log(2\pi e\frac{1}{n}) - \frac{1}{2}\log(2\pi e\frac{n-1}{n^2}) = \frac{1}{2}\log(\frac{n}{n-1}) \tag{24}$$

This is true for all $I(X_i; H)$. Also, this loss function is bounded between $0-1$ therefore it is $1/4-$subgaussian. We get the bound,

$$|\mathbb{E}_{P_{DH}}\left[L_{P_X}(H) - L_D(H)\right]| \leq \sqrt{\frac{2\sigma^2 \sum_{i=1}^{n} I(X_i; H)}{n}} = \sqrt{\frac{2\sigma^2 n \frac{1}{2} \log(\frac{n}{n-1})}{n}} \tag{25}$$

$$= \sqrt{\frac{1}{4} \log(\frac{n}{n-1})} \tag{26}$$

$$= \sqrt{\frac{1}{4} \log(1 + \frac{1}{n-1})} \tag{27}$$

$$\leq \sqrt{\frac{1}{4} \frac{1}{n-1}} \tag{28}$$

Problem 4 (Bandits)**.** [10pts] In the course we mentioned *contextual bandits*. E.g., imagine that you suspect that the rewards that you get from the various arms depend on the day of the week (more generally, a particular feature of the input, aka the "context"). In this case it makes sense to run a separate bandit algorithm for each of the possible contexts. As we discussed, the downside of this approach is that we now have less data for each of the bandit algorithms and hence it will take us longer to learn.

Let c be the context, where $c \in \mathcal{C}$. We run a separate bandit algorithm for each of the $|\mathcal{C}|$ possible contexts and the total number of steps we take is n. We further assume that the number of arms is K and the K does not depend on the context. We use the same bandit algorithm for each of the contexts and we assume that this bandit algorithm has an expected regret of $O(\sqrt{Kn\log(n)})$ (when run over n time steps for a fixed context) in the stochastic setting.

Consider the following setup. For each context there is a different stochastic bandit with K arms. At each time a context is sampled independently at random, the player sees the context, chooses and arm, and receives a reward according to the distribution P_i^c (reward distribution for context c and arm i).

Define the expected regret for this setup as the difference between the expected reward that we could have gotten if for each context we would have chosen the arm with the largest mean minus the expected reward we get using our scheme.

(i) [5pts] Show that the expected regret at time n is $O(\sqrt{Kn|\mathcal{C}|\ln(n)})$.

Discussion: Compared to the case of a single context note that you "only" pay a factor of $\sqrt{|\mathcal{C}|}$. If the number of contexts is small, this might be acceptable. Note further, that this bound is valid, regardless how often each context appears. *Hint:* Start by assuming that each context $c \in \mathcal{C}$ appears n_c times.

Solution: In the sequel let $\{n_c\}_{c\in\mathcal{C}}$ denote the number of times we see context c. The regret is defined as,

$$R \triangleq \mathbb{E}[\sum_{c\in\mathcal{C}} \max_{i\in[k]} n_c\mu_i^{(c)} - \sum_{t:c_t=c} X_t]$$

Then, switching the expectation and summation over $\mathcal{C}$,

$$R = \sum_{c\in\mathcal{C}} R_c$$

where we define R_c as the regret for the times we see context c. Then, the expected regret for the contextual bandit algorithm is equal to

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}[\mathbb{E}[R|\{n_c\}_{c\in\mathcal{C}}]] \\
&\leq \sup_{\{n_c\}_{c\in\mathcal{C}}:\sum_c n_c=n} \sum_{c\in\mathcal{C}} \mathbb{E}[R_c|\{n_c\}_{c\in\mathcal{C}}] \\
&= \sup_{\{n_c\}_{c\in\mathcal{C}}:\sum_c n_c=n} \sum_{c\in\mathcal{C}} O(\sqrt{Kn_c\log(n_c)}) \\
&\leq \sup_{\{n_c\}_{c\in\mathcal{C}}:\sum_c n_c=n} \sum_{c\in\mathcal{C}} O(\sqrt{Kn_c\log(n)}) \\
&\leq O(\sqrt{Kn|\mathcal{C}|\log(n)}),
\end{aligned}$$

where in the last step we have used the fact that under the condition that $n = \sum_c n_c$ the bound is maximized by the choice $n_c = n/|\mathcal{C}|$. This can be seen by writing the Lagrangian $\sum_c \sqrt{n_c} + \lambda(\sum_c n_c - n)$ and taking the derivative with respect to the n_c. This gives $\frac{\partial L}{\partial n_c} = \frac{1}{2}\frac{1}{\sqrt{n_c}} + \lambda$. Setting those derivatives to 0 we see that all n_c should have the same size, i.e., $n_c = n/|\mathcal{C}|$.

(ii) [5pts] Assume now that $\mathcal{C} = [0, 1)$, i.e., $\mathcal{C}$ is very large, and in fact uncountable. In this case we cannot use the strategy above. What would you do in such a case? Under what circumstances will such a scheme likely work? If we assume a fixed number of arms and a very large time horizon, how would you expect the regret to scale?

Solution: Assume that the means of all the K arms are Lipschitz functions of the context c. Hence, by slightly changing the context c, the means only change slightly. In this case it makes sense to quantize the domain $[0, 1)$, i.e., write $[0, 1) = \cup_{i=0}^{m}[\delta i, \delta(i+1))$, where $\delta = 1/m$ for some natural number m.

Since the function is Lipschitz, we have for each $1 \leq i \leq m$,

$$E[R|c \in [\delta(i-1), \delta i]] \leq E[R|c = \delta i - \delta/2] + O(n/m).$$

Plugging this upper bound to our derivation in the previous point gives us,

$$E[R] \leq O(\sqrt{Knm\log(n)}) + O(n/m).$$

This bound is optimized when $m \in O\left(\sqrt[3]{\frac{n}{k\log n}}\right)$. The regret due to this choice of m scales as

$$E[R] \leq O(\sqrt[3]{Kn^2\log(n)}),$$

which is still sub-linear.

Problem 5 (Testing against Poisson Distributions)**.** [10pts] Recall that in the course we talked about how to test against a uniform distribution. In a very similar manner one can test against any fixed distribution. In this problem we explore a slightly more interesting question, i.e., to what degree can we test against a whole class of distributions?

Let $\mathcal{X} = \{0, 1, 2, \cdots\} = \mathbb{N}_{\geq 0}$. Let $\mathcal{P}$ be the *set of Poisson distributions* and let $\mathcal{Q}$ be the set of distributions on $\mathcal{X}$ that have an ℓ_1 distance of at least $\epsilon > 0$ from every element in $\mathcal{P}$.

Let $\bar{X}(X^n) = \frac{1}{n}\sum_{i=1}^n X_i$, i.e., $\bar{X}(X^n)$ is the empirical mean given the samples X^n. Consider the following test statistics $T(X^n) \to \mathbb{R}$, defined by

$$T(X^n) = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\bar{X}} - 1 \right|.$$

The sum in this expression is called the *dispersion*.

(i) [4pts] Assume that the samples come from an element of $\mathcal{P}$ and assume further that the number of samples is large. What value do you expect $T(X^n)$ to take on and why? HINT: What is the expected value of $\bar{X}(X^n)$ and what is the expected value of $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

Solution: Recall that for a Poisson distribution the mean equals the variance. The dispersion is equal to the ratio of the empirical variance to the empirical mean. We therefore expect a value close to 0.

(ii) [4pts] Assume that the number of samples is very large and that $T(X^n)$ takes on a large value, lets say close to 1. What can you conclude?

Solution: As discussed in part (i), if the samples came from an actual Poisson distribution then the test statistic should give us a value close to 0 with high probability if the number of samples is large. We therefore can conclude that with high probability the samples do not come from an element of $\mathcal{P}$, i.e., not from a Poisson distribution.

(iii) [2pts] Assume that the number of samples is very large and that $T(X^n)$ takes on a value very close to the value that you determined in point (i). Can you conclude anything with high probability, and if so, what, or if not, why not?

Solution: Unfortunately we cannot conclude that the samples come from a distribution in $\mathcal{P}$. The test only checks if the samples comes from a distribution whose mean is close to its variance. Poisson distributions have this property but they are not the only distributiosn that have this property. E.g., pick a distribution on $\mathbb{N}_{\geq 0}$ that has weight 1/2 on 2 and 1/2 on 6. It has mean 4 and this is also its variance.