ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

School of Computer and Communication Sciences

Foundations of Data Science
Fall 2026

Assignment date: Thursday, November 12th, 2026, 8:15
Due date: Thursday, November 12th, 2026, 9:45

Midterm Exam – Rooms SG 0211 and SG 0213

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

Only answers given on this handout count.

Name: _____

SCIPER: _____

Problem 1	/ 9
Problem 2	/ 10
Problem 3	/ 10
Total	/29

Problem 1. [**9 pts**] (Tensorization of KL Divergence) Let $n \geq 2$, and let P and Q be joint distributions of n finite-valued random variables. Denote their respective marginal distributions by $P_1, \ldots, P_n$ and $Q_1, \ldots, Q_n$. In class, we have seen that for entropy, it holds that

$$H(P) \leq \sum_{i=1}^{n} H(P_i). \tag{1}$$

In this problem, you show that a similar ordering is **not possible** between $D(P\|Q)$ and $\sum_{i=1}^{n} D(P_i\|Q_i)$.

(a) [**3 pts**] Give an example of distributions P and Q with $P \neq Q$ and $D(P\|Q) < \infty$ such that $D(P\|Q) = \sum_{i=1}^{n} D(P_i\|Q_i)$.

(b) [**3 pts**] Give an example of distributions P and Q such that $D(P\|Q) < \sum_{i=1}^{n} D(P_i\|Q_i)$.

(c) [**3 pts**] Give an example of distributions P and Q such that $D(P\|Q) > \sum_{i=1}^{n} D(P_i\|Q_i)$.

Solution 1. Many examples can be used here. In these solutions, we give one construction for each case, and then two general facts that explain which constructions can work.

(a) We have *equality* whenever *both* P and Q are product distributions, that is, $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P_i(x_i)$ and $Q(x_1, \ldots, x_n) = \prod_{i=1}^{n} Q_i(x_i)$. Indeed, we then obtain

$$\begin{aligned} D(P\|Q) &= \mathbb{E}_P \left[\log \frac{\prod_{i=1}^{n} P_i(X_i)}{\prod_{i=1}^{n} Q_i(X_i)} \right] \\ &= \sum_{i=1}^{n} \mathbb{E}_P \left[\log \frac{P_i(X_i)}{Q_i(X_i)} \right] \\ &= \sum_{i=1}^{n} D(P_i\|Q_i). \end{aligned}$$

To meet the two side conditions of the problem, choose the marginals such that $P_i \neq Q_i$ for at least one i (so that $P \neq Q$), and such that $Q_i(x) > 0$ whenever $P_i(x) > 0$, for every i (so that $D(P\|Q) < \infty$). For example, let the n random variables be independent and binary, uniform under P, and with $Q_i(0) = 1/3$, $Q_i(1) = 2/3$ under Q.

(b) Let us next tackle the case

$$D(P\|Q) < \sum_{i=1}^{n} D(P_i\|Q_i).$$

Choose distributions P_1 and Q_1 such that $0 < D(P_1\|Q_1) < \infty$. Under P, let the n random variables be equal with probability 1, with common distribution P_1. Under Q, let them be equal with probability 1, with common distribution Q_1. Both P and Q are then supported on the diagonal, with $P(x,\ldots,x) = P_1(x)$ and $Q(x,\ldots,x) = Q_1(x)$, and thus

$$D(P\|Q) = D(P_1\|Q_1).$$

All marginals of P are equal to P_1 and all marginals of Q are equal to Q_1, and thus

$$\sum_{i=1}^{n} D(P_i\|Q_i) = nD(P_1\|Q_1).$$

Since $n \geq 2$ and $0 < D(P_1\|Q_1) < \infty$, the inequality is strict.

(c) For the final case,

$$D(P\|Q) > \sum_{i=1}^{n} D(P_i\|Q_i),$$

one option is to select P and Q to violate absolute continuity such that $D(P\|Q) = \infty$, but such that the sum of the marginal divergences remains finite. To accomplish this, choose distributions P_1 and Q_1 such that $D(P_1\|Q_1) < \infty$ and such that P_1 is *not* a point mass, that is, there exist $\alpha \neq \beta$ with $P_1(\alpha) > 0$ and $P_1(\beta) > 0$. (The simplest choice is $P_1 = Q_1$ uniform on $\{0,1\}$.) Under P, let the n random variables be independent and identically distributed according to P_1. Under Q, let the n random variables be equal (with probability 1), distributed according to Q_1. Then,

$$P\{X_1 = \alpha, X_2 = \beta, \cdots, X_n = \beta\} = P_1(\alpha)\, P_1(\beta)^{n-1} > 0,$$

but

$$Q\{X_1 = \alpha, X_2 = \beta, \cdots, X_n = \beta\} = 0,$$

and thus,

$$D(P\|Q) = \infty.$$

At the same time, all marginals of P are equal to P_1 and all marginals of Q are equal to Q_1, so that

$$\sum_{i=1}^{n} D(P_i\|Q_i) = nD(P_1\|Q_1) < \infty,$$

completing the example. Note that the condition that P_1 is not a point mass is needed: if P_1 is a point mass, then P is supported on the diagonal, $D(P\|Q) = D(P_1\|Q_1)$, and we do not obtain the desired strict inequality.

Alternative solution, with $D(P\|Q) < \infty$. It is not necessary to make $D(P\|Q)$ infinite. Let Q be *any product distribution*, $Q(x_1, \dots, x_n) = \prod_{i=1}^n Q_i(x_i)$, and let P be arbitrary with $D(P\|Q) < \infty$. Then

$$\begin{aligned} D(P\|Q) &= \mathbb{E}_P\left[\log \frac{P(X_1, \dots, X_n)}{\prod_{i=1}^n P_i(X_i)}\right] + \mathbb{E}_P\left[\log \frac{\prod_{i=1}^n P_i(X_i)}{\prod_{i=1}^n Q_i(X_i)}\right] \\ &= D\Big(P \,\Big\|\, \prod_{i=1}^n P_i\Big) + \sum_{i=1}^n D(P_i\|Q_i). \end{aligned}$$

The first term is non-negative, and it is zero if and only if P is itself a product distribution. Hence, *every* pair with Q a product distribution and P not a product distribution (and $D(P\|Q) < \infty$) is an example for this case. For instance, under Q let $X_1, \dots, X_n$ be independent and uniform on $\{0,1\}$, and under P let X_1 be uniform on $\{0,1\}$ and $X_1 = X_2 = \cdots = X_n$. Then $\sum_i D(P_i\|Q_i) = 0$ and $D(P\|Q) = (n-1)\log 2$.

Remarks. The following two general facts explain which constructions can work.

(i) *Each marginal divergence is at most the joint divergence.* By the chain rule and the non-negativity of (conditional) KL divergence, for every i,

$$D(P\|Q) = D(P_i\|Q_i) + D\big(P_{X_{\sim i}|X_i} \,\big\|\, Q_{X_{\sim i}|X_i} \,\big|\, P_i\big) \;\geq\; D(P_i\|Q_i),$$

where $X_{\sim i}$ denotes all variables except X_i. Consequently,

$$\max_i D(P_i\|Q_i) \;\leq\; D(P\|Q), \qquad \text{and thus} \qquad \sum_{i=1}^n D(P_i\|Q_i) \;\leq\; n\, D(P\|Q).$$

In particular, if one of the marginal divergences is infinite, then $D(P\|Q)$ is infinite as well. It is therefore *not* possible to solve Case (b) by making a marginal divergence infinite while keeping $D(P\|Q)$ finite. The example given in Case (b) attains the upper bound $n\, D(P\|Q)$ with equality.

(ii) *If* Q *is a product distribution, then* $D(P\|Q) \geq \sum_{i=1}^n D(P_i\|Q_i)$. This is the identity shown in the alternative solution of Case (c). Consequently, any example for Case (b) must use a distribution Q that is *not* a product distribution. In particular, if both P and Q are product distributions, we always have equality, as in Case (a).

Finally, recall that the marginals P_i and Q_i are determined by the joint distributions P and Q; they cannot be chosen separately.

Problem 2. [**10 pts**] (Heavy-Tailed Bandits)

Consider a two-armed stochastic bandit with distinct means $\mu_1, \mu_2 \in [0,1]$ and variances at most 1. Rewards need not be bounded or sub-Gaussian. The rewards from each arm are i.i.d., and the two reward sequences are independent.

Write $\Delta = |\mu_1 - \mu_2|$ and recall that

$$R_n = \Delta\, \mathbb{E}[T_{\text{bad}}(n)],$$

where $T_{\text{bad}}(n)$ counts the pulls of the inferior arm. All logarithms are natural.

(a) **[2 pts]** Arm 1 always returns 1. Arm 2 returns a nonnegative reward X satisfying

$$\mathbb{P}(X > x) = \frac{1}{(1+x)^3}, \qquad x \geq 0.$$

You may use $\mathbb{E}[X] = 1/2$ and $\text{Var}(X) = 3/4$. Thus both arms satisfy the variance assumption, although the reward distribution of arm 2 is not sub-Gaussian.

A learner samples each arm m times, then plays the arm with the larger empirical mean for the remaining $n - 2m$ rounds, where $m \geq 1$ is an integer and $n \geq 2m$. Show that

$$R_n \geq \frac{(n-2m)m}{4(1+m)^3}.$$

Hint: $1 - (1-p)^m \geq mp/2$ whenever $p \in [0,1]$ and $mp \leq 1$.

Choosing $m = \lceil \log n \rceil$ gives $R_n = \Omega(n/(\log n)^2)$ for sufficiently large n. In comparison, recall that under the sub-Gaussian assumptions studied in class, an exploration budget of order $\log n$ gives logarithmic regret for a fixed gap. We now investigate whether replacing the empirical mean can recover this guarantee under the variance assumption alone.

Let $X_1, \ldots, X_{b\ell}$ be i.i.d. with mean μ and variance at most 1, where ℓ, b are positive integers and b is odd. Define the block means and their *median-of-means estimator* by

$$Y_j = \frac{1}{\ell} \sum_{i=(j-1)\ell+1}^{j\ell} X_i, \qquad j = 1, \ldots, b,$$

$$\widehat{\mu} = \text{median}(Y_1, \ldots, Y_b).$$

Recall that, since b is odd, the median is the $(b+1)/2$-th value when $Y_1, \ldots, Y_b$ are arranged in nondecreasing order.

(b) [**2 pts**] Show that

$$\mathbb{P}\left\{ |Y_j - \mu| \geq \frac{2}{\sqrt{\ell}} \right\} \leq \frac{1}{4}.$$

(c) [**2 pts**] Show that

$$\mathbb{P}\left\{|\widehat{\mu}-\mu| \geq \frac{2}{\sqrt{\ell}}\right\} \leq e^{-b/8}.$$

Hint: Consider $Z_j = \mathbf{1}_{\{|Y_j-\mu|\geq 2/\sqrt{\ell}\}}$.

For the remaining parts, return to the general two-arm setting. Assume that n and Δ are known, but the better arm is unknown.

Consider an explore-then-exploit strategy that samples each arm $b\ell$ times, computes its median-of-means estimate using b blocks of size ℓ, and plays the arm with the larger estimate for the remaining $n-2b\ell$ rounds. Ties are broken according to a fixed rule. Assume that b is odd and $2b\ell \leq n$.

(d) [**2 pts**] Show that, if $\ell \geq 16/\Delta^2$, the expected regret satisfies

$$R_n \leq b\ell\Delta + 2(n-2b\ell)\Delta e^{-b/8}.$$

(e) [**2 pts**] Set

$$\ell = \left\lceil \frac{16}{\Delta^2} \right\rceil, \qquad b = \text{the smallest odd integer at least } 8\log(2n).$$

For horizons satisfying $2b\ell \leq n$, show that this strategy achieves

$$R_n = O\left(\frac{\log n}{\Delta}\right).$$

Solution 2. (a) First, note that arm 1 is optimal and the gap is $\Delta = 1/2$ since arm 1 has mean $\mu_1 = 1$, while arm 2 has mean $\mu_2 = 1/2$.

Let $X_1, \ldots, X_m$ be the observations from arm 2 during exploration. Since the empirical mean of arm 1 is always 1, the probability of selecting arm 2 is

$$\mathbb{P}(\text{select arm 2}) = \mathbb{P}\left\{\frac{1}{m}\sum_{i=1}^{m} X_i > 1\right\} \geq \mathbb{P}\left\{\max_{1\leq i\leq m} X_i > m\right\},$$

where we used that $\{\max_{1\leq i\leq m} X_i > m\} \subseteq \{\sum_{i=1}^{m} X_i > m\}$.

Set $p = \mathbb{P}(X > m) = (1+m)^{-3}$. By independence and the supplied inequality,

$$\mathbb{P}\left\{\max_{1\leq i\leq m} X_i > m\right\} = 1-(1-p)^m \geq \frac{mp}{2} = \frac{m}{2(1+m)^3},$$

where $mp = m/(1+m)^3 \leq 1$.

Exploration pulls the inferior arm exactly m times. During exploitation, the inferior arm is played in all remaining rounds if selected. Therefore,

$$\begin{aligned} R_n &= \frac{m}{2} + \frac{n-2m}{2}\,\mathbb{P}(\text{select arm 2}) \\ &\geq \frac{(n-2m)m}{4(1+m)^3}. \end{aligned}$$

(b) Linearity of expectation and independence within a block gives

$$\mathbb{E}[Y_j] = \frac{1}{\ell} \sum_{i=(j-1)\ell+1}^{j\ell} \mathbb{E}[X_i] = \mu, \qquad \mathrm{Var}(Y_j) = \frac{1}{\ell^2} \sum_{i=(j-1)\ell+1}^{j\ell} \mathrm{Var}(X_i) \le \frac{1}{\ell}.$$

Chebyshev's inequality yields

$$\mathbb{P}\left\{|Y_j - \mu| \ge \frac{2}{\sqrt{\ell}}\right\} \le \frac{\mathrm{Var}(Y_j)}{4/\ell} \le \frac{1/\ell}{4/\ell} = \frac{1}{4}.$$

(c) Define $Z_j = \mathbf{1}_{\{|Y_j - \mu| \ge 2/\sqrt{\ell}\}}$. These indicators are independent because the blocks contain disjoint sets of independent observations, and part (b) gives $\mathbb{E}[Z_j] \le 1/4$.

If the median is at least $\mu + 2/\sqrt{\ell}$, at least $(b+1)/2$ block means are at least this large. Similarly, if the median is at most $\mu - 2/\sqrt{\ell}$, at least $(b+1)/2$ block means are at most this small. Consequently,

$$\left\{|\widehat{\mu} - \mu| \ge \frac{2}{\sqrt{\ell}}\right\} \subseteq \left\{\sum_{j=1}^{b} Z_j \ge \frac{b}{2}\right\}.$$

Since $\sum_j \mathbb{E}[Z_j] \le b/4$, Hoeffding's inequality gives

$$\begin{aligned}
\mathbb{P}\left\{|\widehat{\mu} - \mu| \ge \frac{2}{\sqrt{\ell}}\right\} &\le \mathbb{P}\left\{\sum_{j=1}^{b} Z_j \ge \frac{b}{2}\right\} \\
&\le \mathbb{P}\left\{\sum_{j=1}^{b} (Z_j - \mathbb{E}[Z_j]) \ge \frac{b}{4}\right\} \\
&\le \exp\left(-\frac{2(b/4)^2}{b}\right) = e^{-b/8}.
\end{aligned}$$

(d) Let $\widehat{\mu}_1, \widehat{\mu}_2$ be the two median-of-means estimates. Assume without loss of generality that arm 1 is better, so $\mu_1 - \mu_2 = \Delta$.

If both estimates satisfy $|\widehat{\mu}_k - \mu_k| < \Delta/2$, then

$$\widehat{\mu}_1 > \mu_1 - \frac{\Delta}{2} = \mu_2 + \frac{\Delta}{2} > \widehat{\mu}_2.$$

Thus selecting arm 2 requires at least one estimation error to be at least $\Delta/2$. Since $\ell \ge 16/\Delta^2$ implies $2/\sqrt{\ell} \le \Delta/2$, a union bound and part (c) give

$$\begin{aligned}
\mathbb{P}\{\text{select arm } 2\} &\le \sum_{k=1}^{2} \mathbb{P}\left\{|\widehat{\mu}_k - \mu_k| \ge \frac{\Delta}{2}\right\} \\
&\le \sum_{k=1}^{2} \mathbb{P}\left\{|\widehat{\mu}_k - \mu_k| \ge \frac{2}{\sqrt{\ell}}\right\} \\
&\le 2e^{-b/8}.
\end{aligned}$$

Exploration pulls the inferior arm exactly $b\ell$ times. Exploitation incurs regret $(n-2b\ell)\Delta$ if the inferior arm is selected, and zero otherwise. Hence

$$\begin{aligned} R_n &= b\ell\Delta + (n-2b\ell)\Delta\,\mathbb{P}(\text{select arm 2}) \\ &\le b\ell\Delta + 2(n-2b\ell)\Delta e^{-b/8}. \end{aligned}$$

(e) Since $b \ge 8\log(2n)$, we have $e^{-b/8} \le 1/(2n)$. Therefore,

$$R_n \le b\ell\Delta + \frac{n-2b\ell}{n}\Delta \le b\ell\Delta + \Delta.$$

Since $0 < \Delta \le 1$ and b is the smallest odd integer at least $8\log(2n)$,

$$\ell \le \frac{16}{\Delta^2} + 1 \le \frac{17}{\Delta^2}, \qquad b \le 8\log(2n) + 2.$$

Consequently,

$$R_n \le \frac{17\big(8\log(2n)+2\big)}{\Delta} + \Delta = O\left(\frac{\log n}{\Delta}\right).$$

Problem 3. [**10 pts**] (Known probabilities, unknown labels)

An advantage of the Bayesian approach to distribution estimation is that it allows us to incorporate prior knowledge about an unknown distribution. Here we look at a case where we know the probabilities, but not which probability belongs to which symbol.

Let $X_1, \ldots, X_n$ take values in $\{1, 2\}$. For a known constant $a \in (1/2, 1)$, the unknown distribution is either

$$p^{(1)} = (a, 1-a) \qquad \text{or} \qquad p^{(2)} = (1-a, a).$$

We assign equal prior probability to these two possibilities and we index them by $J \in \{1, 2\}$. I.e., J denotes the model index, with

$$\mathbb{P}\{J = 1\} = \mathbb{P}\{J = 2\} = \frac{1}{2}.$$

Conditional on $J = j$, the observations $X_1, X_2, \ldots$ are drawn independently from $p^{(j)}$.

For an observed sequence $x^n = (x_1, \ldots, x_n)$, let t_i denote the number of occurrences of symbol i, so that $t_1 + t_2 = n$.

(a) [**2 pts**] Write down the probability $Q_\pi(x^n)$ of the observed sequence under the Bayesian mixture.

(b) [**2 pts**] Find the ratio

$$\frac{\mathbb{P}\{J = 1 \mid X^n = x^n\}}{\mathbb{P}\{J = 2 \mid X^n = x^n\}},$$

and hence the posterior probability

$$\pi_1(x^n) := \mathbb{P}\{J = 1 \mid X^n = x^n\}.$$

Does the order of the observations affect the posterior? Explain.

(c) [**2 pts**] Find the posterior predictive distribution

$$q_i(x^n) := \mathbb{P}\{X_{n+1} = i \mid X^n = x^n\}, \qquad i \in \{1, 2\},$$

in terms of $\pi_1(x^n)$. What is this distribution when $t_1 = t_2$?

(d) [**2 pts**] Take $a = 3/4$ and $n = 3$. Let the observed sequence be $x^3 = (1, 1, 2)$. Compute the posterior probabilities of the two models and the predictive probability of observing symbol 1 next.

(e) [**2 pts**] An alternative procedure selects the model with the largest posterior probability and uses that model's distribution to predict the next observation.

Which model does this procedure select when $t_1 > t_2$? Compare the probability it assigns to symbol 1 with the Bayesian predictive probability $q_1(x^n)$. Must the Bayesian predictive distribution equal one of the two candidate distributions? Explain what the difference represents.

Solution 3. (a) The likelihoods of the observed sequence are

$$\Pr\{X^n = x^n \mid J = 1\} = a^{t_1}(1-a)^{t_2}$$

and

$$\Pr\{X^n = x^n \mid J = 2\} = (1-a)^{t_1}a^{t_2}.$$

Averaging with the prior weights gives

$$Q_\pi(x^n) = \frac{1}{2}a^{t_1}(1-a)^{t_2} + \frac{1}{2}(1-a)^{t_1}a^{t_2}.$$

Since x^n is one particular ordered sequence, no binomial coefficient is needed.

(b) By Bayes' rule, the posterior odds equal the prior odds times the likelihood ratio. The prior odds are one, so

$$\begin{aligned}\frac{\Pr\{J=1 \mid X^n = x^n\}}{\Pr\{J=2 \mid X^n = x^n\}} &= \frac{a^{t_1}(1-a)^{t_2}}{(1-a)^{t_1}a^{t_2}} \\ &= \left(\frac{a}{1-a}\right)^{t_1-t_2}.\end{aligned}$$

Consequently,

$$\pi_1(x^n) = \frac{1}{1 + \left(\frac{1-a}{a}\right)^{t_1-t_2}}.$$

The other posterior probability is $\pi_2(x^n) = 1 - \pi_1(x^n)$.

The order does not matter: under either model, the likelihood depends only on the counts. In fact, the posterior depends only on $t_1 - t_2$. Each observation of symbol 1 multiplies the odds by $a/(1-a)$, while each observation of symbol 2 multiplies them by its reciprocal.

(c) By conditioning on J, we average the two model distributions using their posterior probabilities:

$$q(x^n) = \pi_1(x^n)p^{(1)} + \big(1 - \pi_1(x^n)\big)p^{(2)}.$$

Therefore,

$$\begin{aligned}q_1(x^n) &= a\pi_1(x^n) + (1-a)\big(1 - \pi_1(x^n)\big) \\ &= (1-a) + (2a-1)\pi_1(x^n),\end{aligned}$$

and

$$q_2(x^n) = 1 - q_1(x^n).$$

If $t_1 = t_2$, neither model is favored: $\pi_1(x^n) = \pi_2(x^n) = 1/2$. Hence

$$q(x^n) = (1/2, 1/2).$$

(d) Here $t_1 = 2$ and $t_2 = 1$. The posterior odds are

$$\left(\frac{3/4}{1/4}\right)^{2-1} = 3.$$

Thus

$$\pi_1(x^3) = \frac{3}{4}, \qquad \pi_2(x^3) = \frac{1}{4}.$$

The predictive probability of symbol 1 is

$$q_1(x^3) = \frac{3}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot \frac{1}{4} = \frac{5}{8}.$$

Therefore $q(x^3) = (5/8, 3/8)$.

(e) When $t_1 > t_2$, the posterior odds exceed one. The maximum a posteriori (MAP) procedure therefore selects $p^{(1)}$ and predicts symbol 1 with probability a.

Bayesian prediction retains uncertainty about the model. For a finite sample, both likelihoods are positive, so

$$\frac{1}{2} < \pi_1(x^n) < 1.$$

It follows that

$$\frac{1}{2} < q_1(x^n) < a.$$

In part (d), MAP predicts probability $3/4$, whereas Bayesian averaging predicts $5/8$.

The Bayesian predictive distribution need not equal either candidate. The source is assumed to be one of the two candidates, but the prediction averages over our uncertainty about which candidate is the true source. For example, after equal counts, the predictive distribution is uniform even though neither candidate is uniform.

As the count difference increasingly favors one model, its posterior probability approaches one, and the Bayesian prediction approaches that model's prediction.