ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Information Theory and Signal Processing	Assignment date: January 27th, 2020, 8:15
Fall 2019	Due date: January 27th, 2020, 11:15

Final Exam – CE2

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	/ 15
Total	/70

Problem 1. *(Bandits with Infinitely Many Arms)*

[10pts] In the course we considered bandits with a finite number of K arms. In this problem we will see that the same ideas apply if we have infinitely many arms as long as there is some additional structure.

Assume that there is an unknown unit-norm vector $\theta \in \mathbb{R}^d$. For every unit-norm vector $u \in \mathbb{R}^d$, there is a bandit. It gives the reward $X_u = \langle u, \theta \rangle + Z_u$, where Z_u is a zero-mean unit-variance Gaussian that is independent over time and independent with respect to different bandits. The nature of the reward is known to the player.

Find a policy, i.e., a strategy of what bandit to probe at any given point in time given a specific history, that has a sublinear regret as time tends to infinity. You can assume that you know the horizon, i.e., we are looking for fixed-horizon policies.

Hint: Start with the simplest thing you can think of. If you do not have time to do the math, describe in words the basic idea of your strategy and why it should give us a sublinear regret.

Problem 2. *(Estimating Support Size)*

[10pts] You are attending Balelec. You want to estimate how many people are attending. Let this number be m. Here is a very simple algorithm. You walk around randomly. Every 5 minutes you take a picture of the person who is right next to at this moment. Assume that 5 minutes is sufficiently long so that in this manner you sample participants at Balelec with uniform probability. Assume further that during the whole time you do your experiment no person joins or leaves Balelec.

You do this N times, where N is a Poisson random variable with mean $n = 100$. Once you are done you look at the photos. Assume that in total you have encountered $K = 102$ distinct people. Out of those 102, 100 you have seen only once, one you saw twice, and one you saw three times. Give an estimate of the number of people attending Balelec (the support size of the distribution). Call this number $\hat{m}$. We do not expect a number as answer since the estiate might involve an optimization step which might not be trivial to do by hand. Simplify as far as you can and then write down how you would get final answer.

Hint: Follow your own path or answer the question according to the following steps.

1. Assume that there are m people attending Balelec. Take a specific person at Balelec. Call this person "1". Given the procedure outlined above, what is the probability that this person appears c_1, $c_1 \geq 0$, times on your photos?
2. Now take two specific people. Call them "1" and "2". What is the probability that they appear $\{c_i\}_{i=1}^{2}$ times on your photos?
3. Now consider all people all Balelec together. Assume as before that each has a specific identity. What is the probability that the m people appear $\{c_i\}_{i=1}^{m}$ times on your photos?
4. Assume again that m people attend Balelec and also as before that we have the counts $\{c_i\}_{i=1}^{m}$. But this time we do not know who has what count, i.e., we do not know the identites of the people. All we know is the counts themselves. What is the probability of getting the counts $\{c_i\}_{i=1}^{m}$? [Note: What we see are the non-zero counts, but since we also assume that we know m, we know in fact all counts.]
5. How can you use the last expression to derive an estimate?

Problem 3. *(Conditional Independence and MMSE)*

[15pts] For simplicity, throughout this problem, **all random variables are assumed to be zero-mean.** *Remark:* You may directly skip to Part *(d)*, taking Equation (2) for granted (as a characterization of conditional independence for Gaussians).

(a) [3 Pts] Show that if X and Y are conditionally independent given Z, then

$$\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] = 0. \tag{1}$$

(b) [3 Pts] Recall that if X and Y are jointly Gaussian (zero-mean), then we have $Y = \alpha X + W$, for some constant α, where W is zero-mean Gaussian *independent of* X. Use this to prove the well-known fact that for jointly Gaussian X and Y, if $\mathbb{E}[XY] = 0$, then X and Y are independent. *Hint:* Simply plug in.

(c) [3 Pts] Let X, Y, Z be jointly Gaussian (and zero-mean, as throughout this problem). Prove that if

$$\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] = 0, \tag{2}$$

then X and Y are conditionally independent given Z. *Hint:* Make sure to solve Part *(b)* first. Recall that for three jointly Gaussians X, Y, Z, we can always write $Y = \gamma X + \delta Z + V$, for some constants γ and δ, where V is Gaussian and independent of X and Z.

(d) [3 Pts] Let X, Y, Z be jointly Gaussian (and zero-mean, as throughout this problem). Recall that we can write $Z = \alpha X + \beta Y + W$, for some constants α and β, where W is Gaussian of some appropriate variance σ_W^2, independent of X and Y. Formulate a necessary and sufficient condition on the triple $(\alpha, \beta, \sigma_W^2)$ such that X and Y are conditionally independent given Z.

(e) [3 Pts] Continuing from Part *(d),* let us now restrict to $\mathbb{E}[X^2] = \mathbb{E}[Y^2] = 1$, and use the notation $\rho = \mathbb{E}[XY]$. This means that we can restrict to $|\alpha| \leq 1$ and $|\beta| \leq 1$. Moreover, let us always select σ_W^2 such that $\mathbb{E}[Z^2] = 1$ (unique choice). Find the unique choice of (α, β) that attains the maximum in the estimation problem

$$\max_{\alpha,\beta} \min_{f} \mathbb{E}[(Z - f(X,Y))^2], \tag{3}$$

where the inner minimum is over all measurable functions $f(x, y)$.

Hint: It may be useful to introduce the notation $a = \mathbb{E}[XZ]$ and $b = \mathbb{E}[YZ]$.

Problem 4. *(A Hilbert space of matrices)*

[10pts] In this problem, we consider the set of matrices $A \in \mathbb{R}^{m\times n}$ with standard matrix addition and multiplication by scalar.

(a) Briefly argue that this is indeed a vector space, using the definition given in class.

(b) Show that $\langle A, B\rangle = \mathrm{trace}(B^H A)$ is a valid inner product.

(c) Explicitly state the norm induced by this inner product. Is this a norm that you have encountered before?

(d) Consider as a further inner product candidate the form $\langle A, B\rangle = \mathrm{trace}(B^H W A)$, where W is a square $(m \times m)$ matrix. Give conditions on W such that this is a valid inner product. Explicit and detailed arguments are required for full credit.

Problem 5. *(Fisher Information and Divergence)*

[10pts] Suppose we are given a family of probability distributions $\{p(\,\cdot\,;\theta) : \theta \in \mathbb{R}\}$ on a set $\mathcal{X}$, parametrized by a real valued parameter θ. (Equivelently, a random variable X whose distribution depends on θ.) Assume that the parametrization is smooth, in the sense that

$$p'(x;\theta) := \frac{\partial}{\partial\theta} p(x;\theta) \quad \text{and} \quad p''(x;\theta) := \frac{\partial^2}{\partial\theta^2} p(x;\theta)$$

exist. (Note that the derivaties are with respect to the parameter θ, not with respect to x.) We will use the notation $E_{\theta_0}[\cdot]$ to denote expectations when the parameter is equal to a particular value θ_0, i.e., $E_\theta[g(X)] = \sum_x p(x;\theta) g(x)$.

Define the function $K(\theta,\theta') := D\big(p(\,\cdot\,;\theta) \| p(\,\cdot\,;\theta')\big)$.

(a) Show that for any θ_0, $\dfrac{\partial}{\partial\theta} K(\theta,\theta_0) = \displaystyle\sum_x p'(x;\theta) \log \frac{p(x;\theta)}{p(x;\theta_0)}$.

(b) Show that $\dfrac{\partial^2}{\partial\theta^2} K(\theta,\theta_0) = \displaystyle\sum_x p''(x;\theta_0) \log \frac{p(x;\theta)}{p(x;\theta_0)} + J(X;\theta)$ with

$$J(X;\theta) := E_\theta\Big[\big(p'(X;\theta)/p(X;\theta)\big)^2\Big].$$

(c) Show that when θ is close to θ_0

$$K(\theta,\theta_0) = \tfrac{1}{2} J(X;\theta_0)(\theta-\theta_0)^2 + o((\theta-\theta_0)^2)$$

(d) Show that $J(X;\theta) = -E_\theta\big[\frac{\partial^2}{\partial\theta^2} \log p(X;\theta)\big]$.

Problem 6. *(Universality via Typicality)*

[15pts] Given an alphabet $\mathcal{U}$, and a rate $0 \le R \le \log|\mathcal{U}|$, consider the sequence of sets

$$\mathcal{A}_n = \bigcup_{Q \in \Pi_n : H(Q) < R} T^n(Q), \quad n = 1, 2, \ldots$$

(i.e., $\mathcal{A}_n$ is the union of the typical sets of all empirical probability distributions with entropy at most R.)

(a) Find $\lim_{n\to\infty} \frac{1}{n} \log|\mathcal{A}_n|$.

Hint: For a lower bound, fix Q with $H(Q) < R$, and a sequence of types $Q_1, Q_2, \ldots$ with $\lim_{n\to\infty} Q_n = Q$. Now observe that for large n, A_n includes $T^n(Q_n)$.

Suppose $P \in \Pi$ with $H(P) < R$ (i.e., P is probability distribution on $\mathcal{U}$ with entropy strictly less than R.)

(b) With $\mathcal{A}_n^c$ denoting the complement of $\mathcal{A}^n$, find $\lim_{n\to\infty} P^n(\mathcal{A}_n^c)$.

(c) Show that there is a injective code $\mathcal{C}_n : \mathcal{U}^n \to \{0,1\}^*$ such that

$$\text{length}(\mathcal{C}_n(u^n)) = \begin{cases} 1 + \lceil \log|\mathcal{A}_n| \rceil & u^n \in \mathcal{A}^n \\ 1 + \lceil n \log|\mathcal{U}| \rceil & \text{else} \end{cases}$$

(d) Show that there is a sequence of injective codes $\mathcal{C}_n : \mathcal{U}^n \to \{0,1\}^*$ such that for any $P \in \Pi$ with $H(P) < R$ and any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\big(\text{length}(\mathcal{C}_n(U^n)) > n(R+\epsilon)\big) = 0.$$