ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Information Theory and Signal Processing
Fall 2019

Assignment date: January 27th, 2020, 8:15
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.

Solution: For a simple fixed-horizon scheme consider the following. Take the d orthonormal unit vectors e_i, $1 \leq i \leq d$. Each of those corresponds to a bandit. Dedicate an ϵ fraction of the time, i.e., $n\epsilon$ steps to exploring. In these first $n\epsilon$ steps probe each of those d bandits $m = n\epsilon/d$ times.

Note that the unknown vector θ can be written as

$$\theta = \sum_{i=1}^{d} \langle e_i, \theta \rangle e_i = \sum_{i=1}^{d} \theta_i e_i,$$

where by some abuse of notation we introduced the scalars $\theta_i = \langle e_i, \theta \rangle$. From our discussion in class we get for each such constant θ_i, m noisy estimates $\theta_i + Z$, where Z is 1-sub-Gaussian. Therefore, $\hat{\theta}_i$ has the form $\hat{\theta}_i = \theta_i + \frac{1}{m} \sum_{k=1}^{m} Z_m$. Hence,

$$\text{Prob}\{|\hat{\theta}_i - \theta_i| \geq \delta\} = \text{Prob}\{|\frac{1}{m} \sum_{k=1}^{m} Z_m| \geq \delta\} \leq 2e^{-m\delta^2/2} = 2e^{-\frac{n\epsilon\delta^2}{2}}.$$

Now we argue as follows. For the first $n\epsilon$ steps we might have an expected regret of 2 per steps since at worst we get a reward of -1 in each step instead of 1. So in total, for this phase we have a regret upper bound of $2n\epsilon$.

Now consider the second phase, which lasts for $n(1-\epsilon)$ steps. Let us upper bound the regret during this phase per step.

If any of the coefficients is more than δ away from the true value let us count a regret of 2 per step also. This is clearly an upper bound. Since there are d coefficients and the probability

for any of them to be more than δ away is upper bounded by $2e^{-\frac{n\epsilon\delta^2}{2}}$ this gives an additional term $4de^{-\frac{n\epsilon\delta^2}{2}}$ per time step if we use the union bound.

For the final term we know that each coefficient is at most δ away from the true value. So the question is by how much the inner product (which should give us 1 in expectation) can be smaller.

Without loss of generality we can assume that $\theta = (1, 0, 0, \cdots, 0)$. Then the vector whose components are θ_i differs from this θ by at most δ in each component. Therefore, these two vectors differ in Euclidean distance by at most $\sqrt{d}\delta$. Draw a ball of radius $\sqrt{d}\delta$ around θ. Then the "worst" $\hat{\theta}$ is the one that is just tangent to this ball. It has length $\sqrt{1 - d\delta^2}$ (we can assume that δ is sufficiently small so that this makes sense). The angle between those vectors, call it α, is $\cos^{-1}\sqrt{1 - d\delta^2}$. And two unit norm vectors that form an angle of α have an inner product of $\cos(\alpha)$. Therefore, we get a lower bound on the inner product of $\sqrt{1 - d\delta^2}$ instead of 1. The regret is therefore at most $1 - \sqrt{1 - d\delta^2} \leq d\delta^2$.

The expected regret for n steps can therefore be upper bounded by

$$\begin{aligned} R_n &\leq 2n\epsilon + n(1-\epsilon)[d\delta^2 + 4de^{-\frac{n\epsilon\delta^2}{2}}] \\ &\leq n(2\epsilon + \delta^2 d + 4de^{-\frac{n\epsilon\delta^2}{2}}). \end{aligned}$$

By a proper choice of ϵ and δ we can make this sublinear. E.g., pick $\epsilon \sim \delta^2 \sim O(\log(n)/\sqrt{n})$.

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?

Solution:

1. Every specific person is sampled a Poisson number of times with mean $100/m$.

2. If we look at two specific people then their counts are independent due to the Poisson sampling and each count follows again a Poisson distribution with mean $100/m$.

3. In general, all counts are independent random variables and follow a Poisson distribution with mean $100/m$.

4. Now assume that we are again given the counts but do not know identities. Note that in general there are many ways to get the same count.

   By assumption, $K = 102$ people are sampled a non-zero number of times. Out of those 100 you saw exactly once, one person you saw twice, and one person you saw three times.

   Let us write down the likelihood of this observation. Let $\lambda = n/m = 100/m$. The likelihood of the observation of these multiplicities given a particular number m of participants is then equal to

$$\begin{aligned}
&\frac{m!}{100!(m-100)!}\left(\frac{\lambda^0 e^{-\lambda}}{0!}\right)^{m-102}\left(\frac{\lambda^1 e^{-\lambda}}{1!}\right)^{100}\frac{\lambda^2 e^{-\lambda}}{2!}\frac{\lambda^3 e^{-\lambda}}{3!}\\
=&\frac{m!}{100!(m-102)!}\frac{\lambda^{105}e^{-\lambda m}}{2!3!}\\
=&\frac{m(m-1)\cdots(m-100)(m-101)}{m^{105}}\frac{100^{105}e^{-100}}{2!3!100!}
\end{aligned}$$

   You now get the estimate by maximizing wrt to the parameter m. I.e., we need to maximize

$$\frac{m(m-1)\cdots(m-101)}{m^{105}}.$$

   This is in principle simple but does not give a nice compact solution.

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 $b = \mathbb{E}[YZ]$.

Solution:

(a) Plug in, using the factorization of the pdf.

(b) Plug in, as in the hint.

(c) As in *(a)*, express $Y = \gamma X + \delta Z + W$, where W is independent of X and Z. Then, plugging in as in *(b)*,

$$\begin{aligned}
0 &= \mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] && (4)\\
&= \mathbb{E}[(X - \mathbb{E}[X|Z])(\gamma X + \delta Z + W - \gamma\mathbb{E}[X|Z]) - \delta Z)] && (5)\\
&= \mathbb{E}[(X - \mathbb{E}[X|Z])(\gamma X + W - \gamma\mathbb{E}[X|Z]))] && (6)\\
&= \mathbb{E}[(X - \mathbb{E}[X|Z])(\gamma(X - \mathbb{E}[X|Z]) + W)] && (7)\\
&= \gamma\mathbb{E}[(X - \mathbb{E}[X|Z])^2] + \mathbb{E}[(X - \mathbb{E}[X|Z])W]. && (8)
\end{aligned}$$

The last expectation is zero since W is independent of X and Z. Hence, we can satisfy the condition only if we select $\gamma = 0$. That is, we must have $Y = \delta Z + W$, where W is independent of X and Z. But this means that Y is conditionally independent of X, given Z.

(d) In class, we have seen that for jointly Gaussians, the conditional mean is equal to the linear MMSE, which is the key to this exercise:

$$\begin{aligned}
\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] &= \mathbb{E}[(X - \frac{\mathbb{E}[XZ]}{\mathbb{E}[Z^2]}Z)(Y - \frac{\mathbb{E}[YZ]}{\mathbb{E}[Z^2]}Z)] && (9)\\
&= \mathbb{E}[XY] - \frac{\mathbb{E}[XZ]\mathbb{E}[YZ]}{\mathbb{E}[Z^2]}. && (10)
\end{aligned}$$

Setting this to zero, the condition is

$$\mathbb{E}[XY]\mathbb{E}[Z^2] = \mathbb{E}[XZ]\mathbb{E}[YZ]. \tag{11}$$

Plugging in, we find

$$\begin{aligned}
\mathbb{E}[Z^2] &= \mathbb{E}[(\alpha X + \beta Y + W)^2] = \alpha^2\mathbb{E}[X^2] + \beta^2\mathbb{E}[Y^2] \ + 2\alpha\beta\mathbb{E}[XY] + \sigma_W^2, && (12)\\
\mathbb{E}[XZ] &= \mathbb{E}[X(\alpha X + \beta Y + W)] = \alpha\mathbb{E}[X^2] \ + \beta\mathbb{E}[XY], && (13)\\
\mathbb{E}[YZ] &= \mathbb{E}[Y(\alpha X + \beta Y + W)] = \alpha\mathbb{E}[XY] \ + \beta\mathbb{E}[Y^2]. && (14)
\end{aligned}$$

Combining, we have conditional independence if

$$\mathbb{E}[XY](\alpha^2\mathbb{E}[X^2] + \beta^2\mathbb{E}[Y^2] \ + 2\alpha\beta\mathbb{E}[XY] + \sigma_W^2) = (\alpha\mathbb{E}[X^2] \ + \beta\mathbb{E}[XY])(\alpha\mathbb{E}[XY] \ + \beta\mathbb{E}[Y^2]). \tag{15}$$

In my view, a nice way of rewriting this condition is (assuming X and Y are not independent)

$$\begin{aligned}
\sigma_W^2 &= \frac{(\alpha\mathbb{E}[X^2] \ + \beta\mathbb{E}[XY])(\alpha\mathbb{E}[XY] \ + \beta\mathbb{E}[Y^2]) - \mathbb{E}[XY](\alpha^2\mathbb{E}[X^2] + \beta^2\mathbb{E}[Y^2] \ + 2\alpha\beta\mathbb{E}[XY])}{\mathbb{E}[XY]}\\
&= \alpha\beta\frac{\mathbb{E}[X^2]\ \mathbb{E}[Y^2] - (\mathbb{E}[XY])^2}{\mathbb{E}[XY]}. && (16)
\end{aligned}$$

This shows, for example, that we can select α and β completely arbitrarily (non-zero) as long as the sign of their product $\alpha\beta$ is the same as the sign of $\mathbb{E}[XY]$; there is always a choice of σ_W^2 such that we have conditional independence.

(e) Following the hint, we will use $a = \mathbb{E}[XZ]$ and $b = \mathbb{E}[YZ]$. Observe that

$$a = \mathbb{E}[XZ] = \mathbb{E}[X(\alpha X + \beta Y + W)] = \alpha + \beta\rho, \tag{17}$$

$$b = \mathbb{E}[YZ] = \mathbb{E}[Y(\alpha X + \beta Y + W)] = \alpha\rho + \beta, \tag{18}$$

so there is a simple bijection between the pairs (a, b) and (α, β). We choose to express our results in terms of a and b.

In class, we have seen that for jointly Gaussians, the conditional mean is equal to the linear MMSE, which is the key to this exercise. Specifically,

$$\min_f \mathbb{E}[(Z - f(X,Y))^2] = \mathbb{E}[(Z - \mathbb{E}[Z|X,Y])^2] \tag{19}$$

$$= \mathbb{E}[(Z - (\mathbb{E}[XZ], \mathbb{E}[YZ])K^{-1}_{(X,Y)}\binom{X}{Y})^2] \tag{20}$$

$$= \mathbb{E}[Z^2] - (\mathbb{E}[XZ], \mathbb{E}[YZ])K^{-1}_{(X,Y)}(\mathbb{E}[XZ], \mathbb{E}[YZ])^T. \tag{21}$$

Plugging in the special case, and noting from Part *(d)* that we need $ab = \rho$,

$$\min_f \mathbb{E}[(Z - f(X,Y))^2] = 1 - (a, \frac{\rho}{a})\frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} a \\ \frac{\rho}{a} \end{pmatrix} \tag{22}$$

$$= 1 - \frac{1}{1-\rho^2}(a - \frac{\rho^2}{a}, -\rho a + \frac{\rho}{a})\begin{pmatrix} a \\ \frac{\rho}{a} \end{pmatrix} \tag{23}$$

$$= 1 - \frac{1}{1-\rho^2}\left(a^2 - \rho^2 - \rho^2 + \frac{\rho^2}{a^2}\right) \tag{24}$$

$$= 1 - \frac{1}{1-\rho^2}\left(a^2 - 2\rho^2 + \frac{\rho^2}{a^2}\right) \tag{25}$$

Taking the derivative of the expression in parentheses with respect to the variable a^2 directly gives $a^2 = \rho$ as the extremum. Its second derivative is always non-negative, so this is a minimum (for the expression in parentheses). Hence, this choice maximizes the resulting mean-squared error.

More explicitly now, this says that the maximizing choice satisfies

$$\mathbb{E}[XZ] = \pm\sqrt{|\rho|}. \tag{26}$$

where the sign can be selected either way without affecting the resulting mean-squared error, and

$$\mathbb{E}[YZ] = \frac{\rho}{\mathbb{E}[XZ]}. \tag{27}$$

For example, assuming that $\rho \geq 0$, a maximizing choice is

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] = \sqrt{\rho}. \tag{28}$$

For interpretation, one could say that in a (geometric mean) sense, this Z sits "exactly halfway" between X and Z.

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 $\langle A, B\rangle = \text{trace}(B^H A)$ 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 $\langle A, B\rangle = \text{trace}(B^H W A)$, 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.

Solution: Note: In the following, we solve assuming the more general case of complex valued matrices. It should be easier to solve for the real-valued case.

(a) We need to check some properties. Because the space is that of matrices,

1. Commutativity holds.
2. Associativity holds.
3. Distributivity holds.
4. The 0 element is the all 0's matrix $\mathbf{0} \in \mathbb{C}^{m\times n}$.
5. For all $A \in \mathbb{C}^{m\times n}$, we have that the element $-A \in \mathbb{C}^{m\times n}$ is such that $A + (-A) = \mathbf{0}$.
6. For all $A \in \mathbb{C}^{m\times n}$, we have that $I_{m\times m}A = A$.

So this is indeed a vector space.

(b) Here, we check the properties of an inner product space. Letting $A, B, C \in \mathbb{C}^{m\times n}, \alpha \in \mathbb{C}$, we have that

1. $\langle A + C, B\rangle = \text{Tr}(B^H(A + C)) = \text{Tr}(B^H A + B^H C) = \text{Tr}(B^H A) + \text{Tr}(B^H C) = \langle A, B\rangle + \langle C, B\rangle$, where we used the linearity of the trace operator.
2. $\langle \alpha A, B\rangle = \text{Tr}(B^H \alpha A) = \alpha \text{Tr}(B^H A) = \alpha\langle A, B\rangle$, where we used the linearity of the trace operator.
3. $\langle A, B\rangle = \text{Tr}(B^H A) = \text{Tr}((A^H B)^H) = \text{Tr}(A^H B)^* = \langle B, A\rangle^*$, where we used the linearity of the trace operator and that conjugation is also linear.

4. We want $\langle A, A\rangle = \mathrm{Tr}(A^H A) \geq 0$. Since $A^H A$ is normal, it is also positive semi-definite, and so all its eigenvalues are positive. One of the property of the trace is that it is equal to the sum of eigenvalues of the matrix considered. In our case, this means $\mathrm{Tr}(A^H A) \geq 0$ since the eigenvalues are all non-negative.

(c) We have that the norm is $\sqrt{\langle A, A\rangle} = \sqrt{\mathrm{Tr}(A^H A)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |A_{ij}|^2}$, which we recognize to be the Frobenius norm. So $\sqrt{\langle A, A\rangle} = \|A\|_F$.

(d) We check that the properties of an inner product space hold, and add conditions on W when necessary.

1. $\langle A + C, B\rangle = \mathrm{Tr}(B^H W(A + C)) = \mathrm{Tr}(B^H WA + B^H WC) = \mathrm{Tr}(B^H A) + \mathrm{Tr}(B^H C) = \langle A, B\rangle + \langle C, B\rangle$, so no restriction on W necessary here.

2. $\langle \alpha A, B\rangle = \mathrm{Tr}(B^H W\alpha A) = \alpha \mathrm{Tr}(B^H WA) = \alpha\langle A, B\rangle$, so no restriction on W necessary here.

3. On one side we have $\langle A, B\rangle = \mathrm{Tr}(B^H WA)$. On the other side we have $\langle B, A\rangle^* = \mathrm{Tr}(A^H WB)^* = \mathrm{Tr}((A^H WB)^H) = \mathrm{Tr}(B^H W^H A)$. To have both sides equal, we need $W = W^H$, i.e., W should be Hermitian.

4. We want $\langle A, A\rangle = \mathrm{Tr}(A^H WA) \geq 0$. That is we would like $A^H WA$ to be positive semi-definite. By definition, this would mean that for any $\mathbf{z} \in \mathbb{C}^n$, we want $\mathbf{z}^H A^H WA\mathbf{z} \geq 0$. Now note that $A\mathbf{z}$ is just another vector, so we can write that we want $\mathbf{z}^H A^H WA\mathbf{z} = (A\mathbf{z})^H W(A\mathbf{z}) \geq 0$ for all $\mathbf{z}$. So we conclude that this is the same as asking that W is positive semi-definite.

Hence, for the inner product $\langle A, B\rangle = \mathrm{Tr}(B^H WA)$ to be valid, we need W to be Hermitian and positive semi-definite.

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

Solution:

(a) We have

$$\begin{aligned}
\frac{\partial}{\partial\theta}K(\theta,\theta_0) &= \frac{\partial}{\partial\theta}\left(\sum_x p(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)}\right)\\
&= \sum_x \frac{\partial}{\partial\theta}\left(p(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)}\right)\\
&= \sum_x p'(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)} + p(x;\theta)\frac{p(x;\theta_0)}{p(x;\theta)}\frac{p'(x;\theta)}{p(x;\theta_0)}\\
&= \sum_x p'(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)} + \sum_x p'(x;\theta)\\
&= \sum_x p'(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)} + \frac{\partial}{\partial\theta}\underbrace{\sum_x p(x;\theta)}_{=1}\\
&= \sum_x p'(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)}.
\end{aligned}$$

(b) Using part (a), we have

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}K(\theta,\theta_0) &= \frac{\partial}{\partial\theta}\left(\sum_x p'(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)}\right)\\
&= \sum_x \frac{\partial}{\partial\theta}\left(p'(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)}\right)\\
&= \sum_x \left(p''(x;\theta)\log\frac{p(x;\theta)}{p(x;\theta_0)} + p'(x;\theta)\frac{p(x;\theta_0)}{p(x;\theta)}\frac{p'(x;\theta)}{p(x;\theta_0)}\right)\\
&= \sum_x p''(x;\theta_0)\log\frac{p(x;\theta)}{p(x;\theta_0)} + \sum_x p(x;\theta_0)\frac{p'(x;\theta)^2}{p(x;\theta)^2}\\
&= \sum_x p''(x;\theta_0)\log\frac{p(x;\theta)}{p(x;\theta_0)} + \mathbb{E}_\theta\left[\frac{p'(X;\theta)^2}{p(X;\theta)^2}\right]\\
&= \sum_x p''(x;\theta_0)\log\frac{p(x;\theta)}{p(x;\theta_0)} + J(X;\theta).
\end{aligned}$$

(c) Using the Taylor expansion of $K(\theta,\theta_0)$ around θ_0, together with the previous answers we get

$$\begin{aligned}
K(\theta,\theta_0) &= K(\theta_0,\theta_0) + \frac{\partial}{\partial\theta}K(\theta_0,\theta_0)(\theta-\theta_0) + \frac{1}{2}\frac{\partial^2}{\partial\theta^2}K(\theta_0,\theta_0)(\theta-\theta_0)^2 + o((\theta-\theta_0)^2)\\
&= \frac{1}{2}J(X,\theta_0)(\theta-\theta_0)^2 + o((\theta-\theta_0)^2).
\end{aligned}$$

(d) We have

$$\begin{aligned}
-\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right] &= -\sum_x p(x;\theta)\frac{\partial^2}{\partial\theta^2}\log p(x;\theta)\\
&= -\sum_x p(x;\theta)\frac{\partial}{\partial\theta}\frac{p'(x;\theta)}{p(x;\theta)}\\
&= -\sum_x p(x;\theta)\frac{p''(x;\theta)p(x;\theta)-p'(x;\theta)^2}{p(x;\theta)^2}\\
&= -\sum_x p''(x;\theta) - p(x;\theta)\frac{p'(x;\theta)^2}{p(x;\theta)^2}\\
&= -\frac{\partial^2}{\partial\theta^2}\underbrace{\sum_x p(x;\theta)}_{=1} + \sum_x p(x;\theta)\frac{p'(x;\theta)^2}{p(x;\theta)^2}\\
&= \mathbb{E}_\theta\left[\frac{p'(X;\theta)^2}{p(X;\theta)^2}\right]\\
&= J(X;\theta).
\end{aligned}$$

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 $\epsilon > 0$,

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