# Problem Set 4 **(Graded)** —*Due Tuesday, November 5, before class starts*

For the Exercise Sessions on Oct 15 and Oct 29

| Last name | First name | SCIPER Nr | Points |
|---|---|---|---|
| | | | |

**Problem 1: Lipschitz Bandits**

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) Bound the expected regret as a function of $K$, $n$, $L$ and the placement of points.

b) 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 1.**

a) 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

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

b) 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 $c\sqrt{Kn\log(n)} + \frac{1}{2}nL/K$ wrt to $K$ and then set the result to 0 and solve for $K$. We get $-((nL - cK\sqrt{Kn\log(n)})/(2K^2)) = 0$ which gives us (ignoring integer constraints) $K = L^{(2/3)}n^{(1/3)}/(c^{(2/3)}\log(n)^{(1/3)})$. If we plug this back into the expression we arrive at $\frac{3}{2}c^{(2/3)}L^{(1/3)}n^{(2/3)}\log(n)^{(1/3)}$.

**Problem 2: MMSE Estimation**

Consider the scenario where $p(x|d) = de^{-dx}$, for $x \geq 0$ (and zero otherwise), that is, the observed data $x$ is distributed according to an exponential with mean $1/d$. Moreover, the desired variable $d$ itself is also exponentially distributed, with parameter $\lambda$, that is, $p(d) = \lambda e^{-\lambda d}$.

*(a)* Find the MMSE estimator of $d$ given $x$, and calculate the corresponding mean-squared error incurred by this estimator.

*(b)* Find the MAP estimator of $d$ given $x$.

**Solution 2.** *(a)* This problem is about the exponential distribution. Let us start by recalling the basics. This distribution is supported on the non-negative real line. Its probability density function is parameterized as $\lambda e^{-\lambda d}$. The mean is $1/\lambda$ and the variance is $1/\lambda^2$. Thus, the second moment is $2/\lambda^2$.

Then, the joint pdf of $D$ and $X$ is

$$p(d, x) = p(d)p(x|d) = \lambda e^{-\lambda d} de^{-dx} = \lambda d e^{-(\lambda+x)d}. \tag{2}$$

Moreover, the marginal distribution of $X$ is

$$p(x) = \int_d p(d, x) = \int_d \lambda d e^{-(\lambda+x)d}. \tag{3}$$

This integral is elementary and can be solved directly. Alternatively, we may rewrite it as

$$p(x) = \int_d \lambda d e^{-(\lambda+x)d} = \frac{\lambda}{\lambda + x}\int_d d(\lambda + x)e^{-(\lambda+x)d}, \tag{4}$$

at which point we recognize the integral to be the mean of an exponential distribution with parameter $(\lambda + x)$. We know that this mean value is $1/(\lambda + x)$. Plugging in, we thus find

$$p(x) = \frac{\lambda}{(\lambda + x)^2}. \tag{5}$$

Finally, plugging in, we find the conditional distribution

$$p(d|x) = \frac{p(d, x)}{p(x)} = (\lambda + x)^2 d e^{-(\lambda+x)d} \tag{6}$$

which, by the way, is known as the Gamma distribution $\Gamma(2, \lambda + x)$, but we will not use this observation in our solutions.

Now, on to the MMSE estimator. As we have seen in class, this is simply the conditional expectation,

$$\hat{D}_{MMSE}(X = x) = \mathbb{E}[D|X = x] = \int_d dp(d|x) = \int_d d(\lambda + x)^2 d e^{-(\lambda+x)d} \tag{7}$$

$$= (\lambda + x)^2 \int_d d^2 e^{-(\lambda+x)d}. \tag{8}$$

The last integral can be solved with elementary techniques. But we will again choose to liken it to an exponential distribution integral. Pattern matching, we would thus write

$$\hat{D}_{MMSE}(x) = (\lambda + x) \int_d d^2 (\lambda + x) e^{-(\lambda+x)d}, \tag{9}$$

and observe that this integral is simply the second moment of the exponential distribution with parameter $(\lambda + x)$, which we know very well, thus,

$$\hat{D}_{MMSE}(x) = (\lambda + x)\frac{2}{(\lambda + x)^2} \tag{10}$$

$$= \frac{2}{\lambda + x} \tag{11}$$

To calculate the corresponding mean-squared error, we may proceed as follows:

$$\mathcal{E} \quad = \quad \int_{d,x} \left( d - \frac{2}{\lambda + x} \right)^2 p(x, d), \tag{12}$$

and we already have the formula for $p(x, d) = \lambda d e^{-(\lambda+x)d}$. So we could just brute-force tackle this integral (which is not undoable). Alternatively, we can simplify our task by leveraging the orthogonality principle that we have seen in class:

$$\mathcal{E} = \mathbb{E}_{D,X}[(D - \hat{D}_{MMSE})^2] \tag{13}$$

$$= \mathbb{E}[D^2] - \mathbb{E}_X[\hat{D}^2_{MMSE}]. \tag{14}$$

The first term is merely $\mathbb{E}[D^2] = 2/\lambda^2$. Moreover, for the second term,

$$\mathbb{E}_X[\hat{D}^2_{MMSE}] = \int_x \left( \frac{2}{\lambda + x} \right)^2 \frac{\lambda}{(\lambda + x)^2} = \int_{x=0}^{\infty} \frac{4\lambda}{(\lambda + x)^4} = \frac{4}{3\lambda^2}. \tag{15}$$

Thus, we find

$$\mathcal{E} = \mathbb{E}[D^2] - \mathbb{E}_X[\hat{D}^2_{MMSE}] = \frac{2}{\lambda^2} - \frac{4}{3\lambda^2} \tag{16}$$

$$= \frac{2}{3\lambda^2} \tag{17}$$

*(b)* MAP estimator is

$$\begin{aligned}\hat{d}_{MAP}(x) &= \arg\max_d p(d|x) &(18)\\ &= \arg\max_d (\lambda+x)^2 d e^{-(\lambda+x)d} &(19)\\ &= \arg\max_d d e^{-(\lambda+x)d}. &(20)\end{aligned}$$

Now, to extremize (max or min) any expression, we of course simply start by finding stationary points, that is, by setting the derivative equal to zero:

$$\frac{d}{dd}\left(de^{-(\lambda+x)d}\right) = e^{-(\lambda+x)d} - d(\lambda+x)e^{-(\lambda+x)d} = e^{-(\lambda+x)d}\left(1-d(\lambda+x)\right). \quad (21)$$

Setting this to zero, we find $d = 1/(\lambda+x)$ as the single stationary point. Thus, the extremum of $p(d|x)$ can only be either at the domain boundaries or at the stationary point we have found. It's easy to see that at both domain boundaries, $d=0$ and $d=\infty$, the function $p(d|x)$ is zero, and that at the stationary point, the function $p(d|x)$ is positive. This implies that we have indeed found the maximum, and thus,

$$\hat{d}_{MAP}(x) = \frac{1}{\lambda+x}. \quad (22)$$

**Problem 3: Conditional Independence and MMSE**

For simplicity, throughout this problem, **all random variables are assumed to be zero-mean.**

*(a)* Show that if $X$ and $Y$ are conditionally independent given $Z$, then

$$\mathbb{E}[(X-\mathbb{E}[X|Z])(Y-\mathbb{E}[Y|Z])] = 0. \quad (23)$$

*(b)* Now let $X$ and $Y$ be jointly Gaussian (zero-mean). It is well known that if $\mathbb{E}[XY]=0$, then $X$ and $Y$ are independent. Establish this fact starting from the observation that for (zero-mean) Gaussian random variables $X$ and $Y$, we may always write $Y = \alpha X + W$, for some constant $\alpha$, where $W$ is zero-mean Gaussian *independent of* $X$. *Note: This prepares you for Part (c).*

*(c)* Let $X, Y, Z$ be jointly Gaussian (and zero-mean, as throughout this problem). Prove that if

$$\mathbb{E}[(X-\mathbb{E}[X|Z])(Y-\mathbb{E}[Y|Z])] = 0, \quad (24)$$

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 $\gamma$ and $\delta$, where $V$ is Gaussian and independent of $X$ and $Z$.

*(d)* Let $X, Y, Z$ be jointly Gaussian (and zero-mean, as throughout this problem). Prove that $X$ and $Y$ are conditionally independent given $Z$ if and only if

$$\mathbb{E}[XY]\mathbb{E}[Z^2] = \mathbb{E}[XZ]\mathbb{E}[YZ]. \quad (25)$$

*(e)* Continuing from Part *(d)*, let us simplify: $\mathbb{E}[X^2] = \mathbb{E}[Y^2] = \mathbb{E}[Z^2] = 1$, and use the notation $\rho = \mathbb{E}[XY]$. Define $a = \mathbb{E}[XZ]$ and $b = \mathbb{E}[YZ]$. Find

$$\arg\max_{a,b} \min_f \mathbb{E}[(Z - f(X,Y))^2], \quad (26)$$

where the inner minimum is over all measurable functions $f(x,y)$, and the maximum is over all choices $a, b$ such that $X$ and $Y$ are conditionally independent given $Z$.

**Solution 3.** *(a)* Plug in, using the factorization of the pdf:

$$\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] \tag{27}$$

$$= \int_{x,y,z} (x - \mathbb{E}[X|Z = z])(y - \mathbb{E}[Y|Z = z])p(z)p(x|z)p(y|z)dxdydz \tag{28}$$

$$= \int_z p(z) \left( \int_x (x - \mathbb{E}[X|Z = z])p(x|z)dx \right) \left( \int_y (y - \mathbb{E}[Y|Z = z])p(y|z)dy \right) dz \tag{29}$$

$$= \int_z p(z) \underbrace{\left( \int_x xp(x|z)dx - \mathbb{E}[X|Z = z] \right)}_{=0} \underbrace{\left( \int_y yp(y|z)dy - \mathbb{E}[Y|Z = z] \right)}_{=0} dz \tag{30}$$

$$= 0, \tag{31}$$

where we have the equality in the second line by using the conditional independence of $X$ and $Y$ given $Z$. The third line follows by reordering the integrals for $x$, $y$, and $z$ separately. The fourth line is by using the fact that $E(X|Z = z)$ does not depend on $x$ and that $\int_x p(x|z) = 1$. And, the whole expression being zero follows by the definition of $E(X|Z = z)$ and $E(Y|Z = z)$.

*(b)* This is a well-known fact and can be proved directly by writing the joint (Gaussian) PDF of $X$ and $Y$. Specifically, if $\mathbb{E}[XY] = 0$, the covariance matrix is a diagonal matrix. Hence, its inverse is also a diagonal matrix. Plugging this into the general formula for the Gaussian PDF, we can see that it breaks into two factors, one involving only $x$ and the other involving only $y$, meaning that indeed, $X$ and $Y$ are independent random variables.

However, in this sub-problem, in preparation for the following parts, you are asked to establish the same fact from the observation that for jointly Gaussian random variables, we may write $Y = \alpha X + W$. To accomplish this, write

$$\mathbb{E}[XY] = \mathbb{E}[X(\alpha X + W)] \tag{32}$$

$$= \alpha\mathbb{E}[X^2] + \mathbb{E}[XW], \tag{33}$$

where the last term is zero since $X$ and $W$ are independent of each other and $W$ has mean zero. Therefore, the only way in which we can have $\mathbb{E}[XY] = 0$ is if we set $\alpha = 0$. This, in turn, means that $Y = \alpha X + W = W$, thus implying that $X$ and $Y$ are independent.

*(c)* As in the hint, express $Y = \gamma X + \delta Z + V$, where $V$ is independent of $X$ and $Z$. The key is to observe that with this,

$$\mathbb{E}[Y|Z] = \mathbb{E}[\gamma X + \delta Z + V|Z] = \gamma\mathbb{E}[X|Z] + \delta Z, \tag{34}$$

by the properties of conditional expectation (and the fact that $V$ is independent of $Z$ and zero-mean). Then, plugging in as in *(b)*,

$$0 = \mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] \tag{35}$$

$$= \mathbb{E}[(X - \mathbb{E}[X|Z])(\gamma X + \delta Z + V - \gamma\mathbb{E}[X|Z] - \delta Z)] \tag{36}$$

$$= \mathbb{E}[(X - \mathbb{E}[X|Z])(\gamma X + V - \gamma\mathbb{E}[X|Z])] \tag{37}$$

$$= \mathbb{E}[(X - \mathbb{E}[X|Z])(\gamma(X - \mathbb{E}[X|Z]) + V)] \tag{38}$$

$$= \gamma\mathbb{E}[(X - \mathbb{E}[X|Z])^2] + \mathbb{E}[(X - \mathbb{E}[X|Z])V]. \tag{39}$$

The last expectation is zero since $V$ is independent of $X$ and $Z$. Hence, we can satisfy the condition only if we select $\gamma = 0$. That is, we must have $Y = \delta Z + V$, where $V$ 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:

$$\mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])] = \mathbb{E}[(X - \frac{\mathbb{E}[XZ]}{\mathbb{E}[Z^2]}Z)(Y - \frac{\mathbb{E}[YZ]}{\mathbb{E}[Z^2]}Z)] \tag{40}$$

$$= \mathbb{E}[XY] - \frac{\mathbb{E}[XZ]\mathbb{E}[YZ]}{\mathbb{E}[Z^2]}. \tag{41}$$

Setting this to zero, the conditional independence condition is

$$\mathbb{E}[XY]\mathbb{E}[Z^2] = \mathbb{E}[XZ]\mathbb{E}[YZ]. \tag{42}$$

*(e)* Using $a = \mathbb{E}[XZ]$ and $b = \mathbb{E}[YZ]$, we can rewrite the conditional independence condition

$$\mathbb{E}[XY]\mathbb{E}[Z^2] = \mathbb{E}[XZ]\mathbb{E}[YZ] \tag{43}$$

simply as $\rho = ab$ (plugging in the normalizing assumptions in the problem statement).

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, we have

$$\min_f \mathbb{E}[(Z - f(X,Y))^2] = \mathbb{E}[(Z - \mathbb{E}[Z|X,Y])^2], \tag{44}$$

and because $(X, Y, Z)$ are jointly Gaussian, we have

$$\mathbb{E}[Z|X,Y] = (\mathbb{E}[XZ], \mathbb{E}[YZ])K_{(X,Y)}^{-1}\binom{X}{Y}, \tag{45}$$

which is referred to as the LMMSE estimator in the Lecture Notes. In class, we proved that the mean-squared error of the LMMSE estimator can be expressed as (see also the Lecture Notes)

$$\mathbb{E}[(Z - (\mathbb{E}[XZ], \mathbb{E}[YZ])K_{(X,Y)}^{-1}\binom{X}{Y})^2] = \mathbb{E}[Z^2] - (\mathbb{E}[XZ], \mathbb{E}[YZ])K_{(X,Y)}^{-1}(\mathbb{E}[XZ], \mathbb{E}[YZ])^T. \tag{46}$$

Now, plugging in the special case, and recalling that we need $\rho = ab$,

$$\min_f \mathbb{E}[(Z - f(X,Y))^2] = 1 - (a, \frac{\rho}{a})\frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\begin{pmatrix} a \\ \frac{\rho}{a} \end{pmatrix} \tag{47}$$

$$= 1 - \frac{1}{1-\rho^2}(a - \frac{\rho^2}{a}, -\rho a + \frac{\rho}{a})\begin{pmatrix} a \\ \frac{\rho}{a} \end{pmatrix} \tag{48}$$

$$= 1 - \frac{1}{1-\rho^2}\left(a^2 - \rho^2 - \rho^2 + \frac{\rho^2}{a^2}\right) \tag{49}$$

$$= 1 - \frac{1}{1-\rho^2}\left(a^2 - 2\rho^2 + \frac{\rho^2}{a^2}\right) \tag{50}$$

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

$$\mathbb{E}[XZ] = \pm\sqrt{|\rho|}. \tag{51}$$

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

$$\mathbb{E}[YZ] = \frac{\rho}{\mathbb{E}[XZ]}. \tag{52}$$

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

$$\mathbb{E}[XZ] = \mathbb{E}[YZ] = \sqrt{\rho}. \tag{53}$$

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

**Problem 4: Fisher Information and Divergence**

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 $\theta$. (Equivalently, a random variable $X$ whose distribution depends on $\theta$.) 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 $\theta$, not with respect to $x$.) We will use the notation $\mathbb{E}_{\theta_0}[\cdot]$ to denote expectations when the parameter is equal to a particular value $\theta_0$, i.e., $\mathbb{E}_\theta[g(X)] = \sum_x p(x;\theta) g(x)$.

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

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

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

$$J(X;\theta) := \mathbb{E}_\theta\big[\big(p'(X;\theta)/p(X;\theta)\big)^2\big].$$

(c) Show that when $\theta$ is close to $\theta_0$

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

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

**Solution 4.** (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

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

(c) Using the Taylor expansion of $K(\theta,\theta_0)$ around $\theta_0$, 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

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