Problem Set 4
Estimation and Decision Theory

For the Exercise Session on Oct 29 — **Due: Tue, November 4, 10am, on Moodle**

1 Problem for Class

Problem 1: Tweedie's Formula

For the special case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, *Tweedie's formula* says that the conditional mean (that is, the MMSE estimator) can be expressed as

$$\mathbb{E}[D|\, X = x] = x + \sigma^2 \ell'(x), \tag{1}$$

where

$$\ell'(x) = \frac{d}{dx} \log f_X(x), \tag{2}$$

where $f_X(x)$ denotes the marginal PDF of X. In this exercise, we derive this formula.

(a) Assume that $f_{X|D}(x|d) = e^{\alpha d x - \psi(d)} f_0(x)$ for some functions $\psi(d)$ and $f_0(x)$ and some constant α (such that $f_{X|D}(x|d)$ is a valid PDF for every value of d). Define

$$\lambda(x) = \log \frac{f_X(x)}{f_0(x)}, \tag{3}$$

where $f_X(x)$ is the marginal PDF of X, i.e., $f_X(x) = \int f_{X|D}(x|\delta) f_D(\delta) d\delta$. With this, establish that

$$\mathbb{E}[D|\, X = x] = \frac{1}{\alpha} \frac{d}{dx} \lambda(x). \tag{4}$$

(b) Show that the case where $X = D + N$, where N is Gaussian noise of mean zero and variance σ^2, is indeed of the form required in Part *(a)* by finding the corresponding $\psi(d), f_0(x)$, and α. Show that in this case, we have

$$\frac{f_0'(x)}{f_0(x)} = -\frac{x}{\sigma^2}, \tag{5}$$

and use this fact in combination with Part *(a)* to establish Tweedie's formula.

Solution 1. This formula is due to M. C. K. Tweedie, "Functions of a statistical variate with given means, with special reference to Laplacian distributions," *Proc. Camb. Phil. Soc.,* Vol. 43 (1947), pp.41-49.

(a) Simply plugging in, we find

$$\frac{d}{dx}\lambda(x) = \frac{d}{dx}\log\left(\frac{\int f_{X|D}(x|\delta)f_D(\delta)d\delta}{f_0(x)}\right) \tag{6}$$

$$= \frac{d}{dx}\log\left(\frac{\int e^{\alpha\delta x-\psi(\delta)}f_0(x)f_D(\delta)d\delta}{f_0(x)}\right) \tag{7}$$

$$= \frac{d}{dx}\log\int e^{\alpha\delta x-\psi(\delta)}f_D(\delta)d\delta \tag{8}$$

$$= \frac{1}{\int e^{\alpha\delta x-\psi(\delta)}f_D(\delta)d\delta}\int \alpha\delta e^{\alpha\delta x-\psi(\delta)}f_D(\delta)d\delta \tag{9}$$

But since we know that

$$\int e^{\alpha\delta x-\psi(\delta)}f_0(x)f_D(\delta)d\delta = f_X(x), \tag{10}$$

we can rewrite

$$\frac{d}{dx}\lambda(x) = \frac{f_0(x)}{f_X(x)}\int \alpha\delta e^{\alpha\delta x-\psi(\delta)}f_D(\delta)d\delta \tag{11}$$

$$= \alpha\int \delta \underbrace{\frac{e^{\alpha\delta x-\psi(\delta)}f_0(x)f_D(\delta)}{f_X(x)}}_{f_{D|X}(d|x)}d\delta \tag{12}$$

$$= \alpha\mathbb{E}\left[D|\,X=x\right] \tag{13}$$

as claimed.

(b) In this case, we have

$$f_{X|D}(x|d) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}e^{-\frac{d^2}{2\sigma^2}}e^{\frac{1}{\sigma^2}xd}. \tag{14}$$

Pattern matching with the desired form

$$f_{X|D}(x|d) = e^{\alpha dx-\psi(d)}f_0(x), \tag{15}$$

it is quickly verified that $\alpha = 1/\sigma^2$, and

$$f_0(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}, \tag{16}$$

and thus,

$$f_0'(x) = -\frac{1}{\sqrt{2\pi}\sigma}\frac{2x}{2\sigma^2}e^{-\frac{x^2}{2\sigma^2}}, \tag{17}$$

giving the claimed result.

Putting things together, we have

$$\mathbb{E}\left[D|\,X=x\right] = \frac{1}{\alpha}\frac{d}{dx}\lambda(x) = \sigma^2\left(\frac{d}{dx}\log f_X(x) - \frac{d}{dx}\log f_0(x)\right) \tag{18}$$

$$= \sigma^2\left(\frac{d}{dx}\log f_X(x) - \frac{f_0'(x)}{f_0(x)}\right) \tag{19}$$

$$= \sigma^2\left(\frac{d}{dx}\log f_X(x) + \frac{x}{\sigma^2}\right) \tag{20}$$

$$= x + \sigma^2\frac{d}{dx}\log f_X(x), \tag{21}$$

which is the claimed formula.

2 The Homework

Problem 2: Bernoulli Data with Beta Prior

Let $S \in [0,1]$ be distributed with a Beta distribution with parameters $(1/2, 1/2)$, which, as we have seen in class, is $p(s) = \frac{1}{\pi} s^{-\frac{1}{2}}(1-s)^{-\frac{1}{2}}$. We make n observations $X_1, X_2, \ldots, X_n$ that are (conditionally) independent Bernoulli (S) random variables.

a) Calculate the conditional distribution $p(s|x_1, x_2, \ldots, x_n)$. Express it in terms of the integer t, which is the number of '1's in the sample $(x_1, x_2, \ldots, x_n)$.

Hint: For $a, b \in \mathbb{R}^+$, we have $\int_0^1 y^{a-1}(1-y)^{b-1}dy = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, where $\Gamma(\cdot)$ denotes the Gamma function.

b) We would like to estimate S from $X_1, X_2, \ldots, X_n$ such as to minimize the mean-squared error $\mathbb{E}[(S - \hat{S}(X_1, X_2, \ldots, X_n)^2]$. Find the optimum estimate $\hat{S}(X_1, X_2, \ldots, X_n)$. Simplify your result as much as possible.

Hint: The Gamma function satisfies the property, for $c \in \mathbb{R}^+$, that $\Gamma(c+1) = c\Gamma(c)$.

Solution 2. The MMSE estimator is the conditional expectation. Let t denote the number of ones in the sample $(x_1, x_2, \ldots, x_n)$.

Let us first find the conditional distribution $p(s|x_1, x_2, \ldots, x_n)$.

$$p(s, x_1, x_2, \ldots, x_n) = p(s)p(x_1, x_2, \ldots, x_n|s) \tag{22}$$

$$= \frac{s^{-\frac{1}{2}}(1-s)^{-\frac{1}{2}}}{\pi} s^t (1-s)^{n-t} \tag{23}$$

$$= \frac{1}{\pi} s^{t-\frac{1}{2}}(1-s)^{n-t-\frac{1}{2}} \tag{24}$$

and thus,

$$p(x_1, x_2, \ldots, x_n) = \frac{1}{\pi}\int_0^1 s^{t-\frac{1}{2}}(1-s)^{n-t-\frac{1}{2}} ds \tag{25}$$

$$= \frac{1}{\pi}\frac{\Gamma(t+\frac{1}{2})\Gamma(n-t+\frac{1}{2})}{\Gamma(n+1)} \tag{26}$$

Thus,

$$p(s|x_1, x_2, \ldots, x_n) = \frac{p(s)p(x_1, x_2, \ldots, x_n|s)}{p(x_1, x_2, \ldots, x_n)} \tag{27}$$

$$= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} s^{t-1/2}(1-s)^{n-t-1/2} \tag{28}$$

To calculate the conditional mean, we now proceed as follows:

$$\begin{aligned}
&\mathbb{E}[S|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] \\
&= \int_0^1 sp(s|x_1, x_2, \ldots, x_n)ds && (29)\\
&= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} \int_0^1 s \cdot s^{t-1/2}(1-s)^{n-t-1/2}ds && (30)\\
&= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} \int_0^1 s^{t+1/2}(1-s)^{n-t-1/2}ds && (31)\\
&= \frac{\Gamma(n+1)}{\Gamma(t+1/2)\Gamma(n-t+1/2)} \cdot \frac{\Gamma(t+3/2)\Gamma(n-t+1/2)}{\Gamma(n+2)} && (32)\\
&= \frac{\Gamma(n+1)}{\Gamma(n+2)} \cdot \frac{\Gamma(t+3/2)}{\Gamma(t+1/2)} && (33)\\
&= \frac{t+1/2}{n+1}, && (34)
\end{aligned}$$

which, intriguingly, is exactly the "add-1/2" estimator that we have studied (from a different perspective) in the chapter on Distribution Estimation...

3 Additional Problems

Problem 3: Parameter Estimation and Fisher Information

Find the Fisher information for the following families:

(a) $f_\theta(x) = N(0, \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$

(b) $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$

(c) What is the Cramèr Rao lower bound on $\mathbb{E}_\theta(\hat{\theta}(X) - \theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

Solution 3. (a) Since $f_\theta(x) = N(0, \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$, we have

$$f'_\theta = -\frac{1}{2}\frac{1}{\sqrt{2\pi\theta^3}} e^{-\frac{x^2}{2\theta}} + \frac{x^2}{2\theta^2}\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}} \tag{35}$$

and

$$\frac{f'_\theta}{f_\theta} = \left(-\frac{1}{2\theta} + \frac{x^2}{2\theta^2}\right). \tag{36}$$

Therefore the Fisher information,

$$\begin{aligned}
J(\theta) &= \mathbb{E}_\theta \left(\frac{f'_\theta}{f_\theta}\right)^2 && (37)\\
&= \mathbb{E}_\theta \left(\frac{1}{4\theta^2} - 2\frac{1}{2\theta}\frac{x^2}{2\theta^2} + \frac{x^4}{4\theta^4}\right) && (38)\\
&= \frac{1}{4\theta^2} - \frac{1}{\theta}\frac{\theta}{2\theta^2} + \frac{3\theta^2}{4\theta^4} && (39)\\
&= \frac{1}{2\theta^2}, && (40)
\end{aligned}$$

where for $X \sim N(0,\theta)$, $\mathbb{E}[X^2] = \theta$ and $\mathbb{E}[X^4] = 3\theta^2$.

(b) Since $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$, we have $\ln f_\theta = \ln\theta - \theta x$, and

$$\frac{f'_\theta}{f_\theta} = \frac{\partial \ln f_\theta}{\partial \theta} = \frac{1}{\theta} - x, \tag{41}$$

and therefore

$$\begin{aligned} J(\theta) &= \mathbb{E}_\theta\left(\frac{d\ln f_\theta}{d\theta}\right)^2 && (42)\\ &= \mathbb{E}_\theta\left(\frac{1}{\theta^2} - 2\frac{1}{\theta}x + x^2\right) && (43)\\ &= \frac{1}{\theta^2} - 2\frac{1}{\theta}\frac{1}{\theta} + \frac{2}{\theta^2} && (44)\\ &= \frac{1}{\theta^2} && (45)\end{aligned}$$

where for $X \sim \text{Exp}(\theta)$, $\mathbb{E}[X] = \frac{1}{\theta}$ and $\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2 = \frac{2}{\theta^2}$.

(c) The Cramèr-Rao lower bound is the reciprocal of the Fisher information, and is therefore $2\theta^2$ and θ^2 for parts (a) and (b) respectively.

Problem 4: Wiener Filter and Irrelevant Data

As we have seen in class, the (FIR) Wiener filter is given by

$$\mathbf{w} = R_x^{-1}\mathbf{r}_{dx}, \tag{46}$$

where R_x is the autocorrelation matrix of the data that's being used, and $\mathbf{r}_{dx}$ is the cross-correlation between the data and the desired output. For this to be well defined, R_x should be full rank. In this problem, we study this question in more detail.

(a) In many applications, the signal acquisition process is noisy. That is, the data $x[n] = s[n] + w[n]$, where $s[n]$ is an *arbitrary* signal, and $w[n]$ is white noise. Prove that in this case, the $p-$dimensional autocorrelation matrix R_x is full rank (i.e., invertible) for any p. (Note: Be careful not to make *any* assumptions about the signal $s[n]$.)

(b) In some other cases, R_x could be rank-deficient. To study this, prove first that if the (FIR) Wiener filter based on the data $\mathbf{x} = \{x[n]\}_{n=0}^{p-1}$ is $\mathbf{w}$, then the (FIR) Wiener filter based on the modified data $A\mathbf{x}$ (where A is an invertible matrix) is $A^{-H}\mathbf{w}$, (where we use the relatively common notation $A^{-H} = (A^{-1})^H$).

(c) Explain how to find the (FIR) Wiener filter when R_x is rank-deficient. Discuss existence and uniqueness. *Hint:* Use Part *(b)* to transform your data to a more convenient basis.

Solution 4. *(a)* The covariance matrix of the data can be easily seen to be

$$R_x = R_s + R_v = R_s + \sigma_v^2 I, \tag{47}$$

where I is the identity matrix. However, as we have seen earlier in this class, the eigenvalues of the matrix R_x satisfy

$$\lambda_i(R_x) = \lambda_i(R_s) + \sigma_v^2. \tag{48}$$

Because R_s is a covariance matrix, it satisfies $\lambda_i(R_s) \geq 0$, and thus, we must have that $\lambda_i(R_x) \geq \sigma_v^2 > 0$, hence, R_x is invertible.

(b) By assumption, the estimate based on $\mathbf{y}$ is given by

$$\hat{\mathbf{x}} \quad = \quad \mathbf{w}^H \mathbf{y}. \tag{49}$$

But then,

$$\hat{\mathbf{x}} \quad = \quad (A^{-H}\mathbf{w})^H A\mathbf{y} = \mathbf{w}^H A^{-1} A\mathbf{y} = \mathbf{w}^H \mathbf{y}, \tag{50}$$

which is the exact same operation on the data, and thus, must be optimal.

(c) The idea is simply to transfer to the *eigenspace* of the matrix R_x, an idea similar to the whitening filter discussed in class. Thus, without loss of generality, the case of a rank-deficient matrix R_x can be thought of as the case of a diagonal R_x where some diagonal elements are zero. Thus, in these eigencoordinates, we can immediately find the Wiener filter by selecting all the coefficients corresponding to the non-zero entries in R_x accordingly, and the remaining filter coefficients arbitrarily.

Problem 5: 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 θ. (Equivalently, 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 derivatives are with respect to the parameter θ, 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 θ_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 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 $\dfrac{\partial^2}{\partial\theta^2} K(\theta,\theta_0) = \displaystyle\sum_x p''(x;\theta) \log \frac{p(x;\theta)}{p(x;\theta_0)} + J(X;\theta)$ with

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

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

Solution 5. (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 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

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