
Problem Set 5
Estimation and Decision Theory
For the Exercise Session of Nov 3 - Nov 10 — **Due: Tue, November 17, 10am, on Moodle**

---

1 The Homework

The following two problems are the core of this homework. You work out solutions and turn them in. Problem 1 is suitable for the exercise session November 3. Problem 2 is suitable for November 10.

Problem 1: Fano method

In this problem, we will develop a framework to find lower bounds on the estimation error of the minimax distribution estimator. We will use Fano's inequality, which we saw in class. First, recall the minimax distribution estimation problem:

$$r_{k,n}^{L} = \min_{q} \sup_{p \in \Delta_k} \mathbb{E}_{X^n \sim p^n}[L(p, q(X^n))].$$

Assume that the loss L is symmetric in its arguments, satisfies the triangle inequality, and that $L(x,x) = 0 \ \forall x$.

(i) Let $\mathcal{P} := \{P_1, \ldots, P_m\} \subseteq \Delta_k$ where $m \geq 2$, such that $L(P_i, P_j) \geq \delta > 0$ for $i \neq j$.
Show that

$$\sup_{p \in \Delta_k} \mathbb{E}_{X^n \sim p^n}[L(p, q(X^n))] \geq \frac{\delta}{2} \max_j \mathbb{P}_{X^n \sim P_j^n}\left[L(P_j, q(X^n)) \geq \frac{\delta}{2}\right].$$

Hint: Markov's inequality is your friend here.

(ii) Now let $(V, X^n) \sim P_{V,X^n}$ be jointly distributed such that V is uniformly distributed over $\{1, \ldots, m\}$ and $\mathbb{P}[X^n = x^n | V = j] = P_j^n(x^n)$. Define $Z := \arg\min_j L(q(X^n), P_j)$. Show that

$$\mathbb{P}[Z \neq V] \leq \max_j \mathbb{P}_{X^n \sim P_j^n}\left[L(q(X^n), P_j) \geq \frac{\delta}{2}\right].$$

(iii) Use Fano's inequality to show that

$$\max_j \mathbb{P}_{X^n \sim P_j^n}\left[L(q(X^n), P_j) \geq \frac{\delta}{2}\right] \geq 1 - \frac{I(X^n; V) + \log 2}{\log m}.$$

Hint: Recall $I(Y;W) = H(Y) - H(Y|W)$.

(iv) Show that

$$I(X^n; V) \leq \frac{1}{m^2} \sum_{i,j=1}^{m} D(P_i^n \| P_j^n) \leq n \max_{i,j} D(P_i \| P_j),$$

and thus

$$r_{k,n}^{L} \geq \frac{\delta}{2}\left(1 - \frac{n \max_{i,j} D(P_i \| P_j) + \log 2}{\log m}\right).$$

Hint: $I(W;Y) = D(P_{W|Y} \| P_W | P_Y)$, and the KL divergence is a convex function.

Problem 2: Parameter Estimation and Fisher Information

Throughout this problem, $\theta > 0$ and we observe a single random variable X with density p_θ. Recall that the score and Fisher information are defined by

$$\ell_\theta(X) = \frac{\partial}{\partial\theta} \log p_\theta(X), \qquad I_\theta = \mathbb{E}_\theta[\ell_\theta(X)^2].$$

(a) Find the Fisher information for the family

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}, \qquad x \in \mathbb{R},$$

corresponding to $X \sim N(0, \theta)$.

(b) Find the Fisher information for the family

$$p_\theta(x) = \theta e^{-\theta x}, \qquad x \geq 0,$$

with $p_\theta(x) = 0$ for $x < 0$.

(c) What is the Cramér–Rao lower bound on

$$\mathbb{E}_\theta[(T(X) - \theta)^2]$$

for an unbiased estimator $T(X)$ of θ in parts (a) and (b)?

(d) Now suppose that $X \sim \text{Unif}([0, \theta])$, with density

$$p_\theta(x) = \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(x).$$

(i) Compute the score $\ell_\theta(X)$ and I_θ using the definitions above.

(ii) Show that $T(X) = 2X$ is unbiased and compute its mean-squared error.

(iii) Compare this mean-squared error with $1/I_\theta$. Does this contradict the Cramér–Rao theorem? Justify your answer.

2 Additional Problems

For the next three problems, you do not need to turn in solutions. For the exam, we do expect you know how to solve these problems. So, you may keep these problems around and tackle them in preparation for the Midterm and/or Final exam. Or you may solve them now.

Problem 3: 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 $f_{X|D}(x|d) = e^{\alpha dx - \psi(d)} f_0(x)$ 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.

Problem 4: Wiener Filter and Irrelevant Data

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

$$\mathbf{w} \quad = \quad R_x^{-1}\mathbf{r}_{dx}, \tag{6}$$

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.

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