

## Problem

Throughout the problem, we let  $\mathcal{B}$  be the Borel subsets of  $\mathbb{R}^d$ .

**A. Preliminaries.** Let  $f$  and  $g$  be two probability densities on  $\mathbb{R}^d$ , that is, nonnegative functions such that

$$\int f = \int g = 1.$$

(All integrals are evaluated with respect to the Lebesgue measure.)

1. Show that

$$\int |f - g| = 2 \int_{A_{fg}} (f - g),$$

where  $A_{fg}$  is the set  $\{f > g\}$ , i.e.,

$$A_{fg} = \{x \in \mathbb{R}^d : f(x) > g(x)\}.$$

2. Deduce that

$$\int |f - g| = 2 \sup_{B \in \mathcal{B}} \left| \int_B f - \int_B g \right|.$$

This result is known as *Scheff 's theorem*.

**B. A selection problem.** Assume we are given a sample of independent random variables  $X_1, \dots, X_n$  with common **unknown** density  $f$ . We denote by  $\mathcal{F}$  a collection of densities parameterized by  $\theta$ :

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}.$$

Our goal is to select in  $\mathcal{F}$  the “best” possible density, using only  $X_1, \dots, X_n$ .

1. Let  $\mu_n$  be the empirical measure associated with  $X_1, \dots, X_n$ . Explain why the strategy that chooses  $\theta$  in  $\Theta$  by minimizing the quantity

$$\sup_{B \in \mathcal{B}} \left| \int_B f_\theta - \mu_n(B) \right|$$

is not a good idea.

2. Introduce the collection of sets

$$\mathcal{A} = \{\{f_\theta > f_{\theta'}\} : (\theta, \theta') \in \Theta^2\}.$$

In order to choose the “best” density in  $\mathcal{F}$ , a possible route is to minimize in  $\theta$  the following criterion:

$$\Delta(\theta) = \sup_{A \in \mathcal{A}} \left| \int_A f_\theta - \mu_n(A) \right|.$$

We denote by  $\theta^*$  an element of  $\Theta$  such that  $\Delta(\theta^*) = \inf_{\theta \in \Theta} \Delta(\theta)$ .

2.a Let  $\bar{\theta}$  be an element of  $\Theta$  such that

$$\int |f_{\bar{\theta}} - f| = \inf_{\theta \in \Theta} \int |f_\theta - f|.$$

Prove that

$$\int |f_{\theta^*} - f_{\bar{\theta}}| \leq 4 \sup_{A \in \mathcal{A}} \left| \int_A f_{\bar{\theta}} - \mu_n(A) \right|.$$

2.b Next, show that

$$\int |f_{\theta^*} - f| \leq 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + 4\Delta_n,$$

where  $\Delta_n$  is some explicit random quantity.

3. 3.a Recall the definition of  $\mathbf{S}_{\mathcal{A}}(n)$ , the shatter coefficient of  $n$  points by the class  $\mathcal{A}$ .

3.b Show that

$$\mathbb{E} \left( \int |f_{\theta^*} - f| \right) \leq 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + O \left( \sqrt{\frac{\log(\mathbf{S}_{\mathcal{A}}(n))}{n}} \right).$$

3.c Provide a statistical interpretation of this inequality.

**C. Application.** On the real line  $\mathbb{R}$ , we let  $\mathcal{F}$  be the set of Gaussian densities, parameterized by their mean and variance, i.e.,

$$\mathcal{F} = \left\{ f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/(2\sigma^2)} : \theta = (m, \sigma^2) \in \mathbb{R} \times (0, \infty) \right\}.$$

1. Prove that  $\mathcal{A}$  is contained in a class of sets  $\mathcal{B}_2$  that can be easily described.
2. Determine the Vapnik-Chervonenkis dimension  $V$  of  $\mathcal{B}_2$ .
3. Conclude that

$$\mathbb{E}\left(\int |f_{\theta^*} - f|\right) \leq 3 \inf_{\theta \in \Theta} \int |f_{\theta} - f| + O\left(\sqrt{\frac{V \log n}{n}}\right).$$