Assouad’s method

Assouad’s method provides a somewhat different technique for proving lower bounds. Instead of reducing the estimation problem to a multiple hypothesis test or simpler estimation problem, as with Le Cam’s method and Fano’s method from the preceding lectures, here we transform the original estimation problem into multiple binary hypothesis testing problems, using the structure of the problem in an essential way. Assouad’s method applies only problems where the loss we care about is naturally related to identification of individual points on a hypercube.

3.1 The method

3.1.1 Well-separated problems

To describe the method, we begin by encoding a notion of separation and loss, similar to what we did in the classical reduction of estimation to testing. For some \( d \in \mathbb{N} \), let \( V = \{-1, 1\}^d \), and let us consider a family \( \{P_v\}_{v \in V} \subset \mathcal{P} \) indexed by the hypercube. We say that the family \( P_v \) induces a 2\( \delta \)-Hamming separation for the loss \( \Phi \circ \rho \) if there exists a function \( \hat{v} : \theta(P) \to \{-1, 1\}^d \) satisfying

\[
\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \{[\hat{v}(\theta)]_j \neq v_j\}.
\]

(3.1.1)

That is, we can take the parameter \( \theta \) and test the individual indices via \( \hat{v} \).

Example 3.1 (Estimation in \( \ell_1 \)-error): Suppose we have a family of multivariate Laplace distributions on \( \mathbb{R}^d \)—distributions with density proportional to \( p(x) \propto \exp(-\|x - \mu\|_1) \)—and we wish to estimate the mean in \( \ell_1 \)-distance. For \( v \in \{-1, 1\}^d \) and some fixed \( \delta > 0 \) let \( p_v \) be the density

\[
p_v(x) = \frac{1}{2} \exp(-\|x - \delta v\|_1),
\]

which has mean \( \theta(P_v) = \delta v \). Under the \( \ell_1 \)-loss, we have for any \( \theta \in \mathbb{R}^d \) that

\[
\|\theta - \theta(P_v)\|_1 = \sum_{j=1}^{d} |\theta_j - \delta v_j| \geq \delta \sum_{j=1}^{d} \{\text{sign}(\theta_j) \neq v_j\},
\]

so that this family induces a \( \delta \)-Hamming separation for the \( \ell_1 \)-loss. ♠
3.1.2 From estimation to multiple binary tests

As in the standard reduction from estimation to testing, we consider the following random process: nature chooses a vector \( V \in \{-1, 1\}^d \) uniformly at random, after which the sample \( X \) is drawn from the distribution \( P_v \) conditional on \( V = v \). Then, if we let \( P_{\pm j} \) denote the joint distribution over the random index \( V \) and \( X \) conditional on the \( j \)-th coordinate \( V_j = \pm 1 \), we obtain the following sharper version of Assouad’s lemma [2] (see also the paper [1]); we provide a proof in Section 3.1.3 to follow.

**Lemma 3.2.** Under the conditions of the previous paragraph, we have

\[
\mathcal{M}(\theta(P), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} \inf_{\Psi} [P_{+j}(\Psi(X) \neq +1) + P_{-j}(\Psi(X) \neq -1)] .
\]

While Lemma 3.2 requires conditions on the loss \( \Phi \) and metric \( \rho \) for the separation condition (3.1.1) to hold, it is sometimes easier to apply than Fano’s method. Moreover, while we will not address this in class, several researchers [1, 2] have noted that it appears to allow easier application in so-called “interactive” settings—those for which the sampling of the \( X_i \) may not be precisely i.i.d. It is closely related to Le Cam’s method, discussed previously, as we see that if we define \( P_{v, +j} = 2^{1-d} \sum_{v \in \{-1, 1\}^d} P_v \) (and similarly for \( P_{v, -j} \)), Lemma 3.2 is equivalent to

\[
\mathcal{M}(\theta(P), \Phi \circ \rho) \geq \delta \sum_{j=1}^{d} [1 - \|P_{+j} - P_{-j}\|_{TV}] .
\] (3.1.2)

There are standard weakenings of the lower bound (3.1.2) (and Lemma 3.2). We give one such weakening. First, we note that the total variation is convex, so that if we define \( P_{v, +j} \) to be the distribution \( P_v \) where coordinate \( j \) takes the value \( v_j = 1 \) (and similarly for \( P_{v, -j} \)), we have

\[
P_{+j} = \frac{1}{2^d} \sum_{v \in \{-1, 1\}^d} P_{v, +j} \quad \text{and} \quad P_{-j} = \frac{1}{2^d} \sum_{v \in \{-1, 1\}^d} P_{v, +j} .
\]

Thus, by the triangle inequality, we have

\[
\|P_{+j} - P_{-j}\|_{TV} = \left\| \frac{1}{2^d} \sum_{v \in \{-1, 1\}^d} P_{v, +j} - P_{v, -j} \right\|_{TV} \leq \frac{1}{2^d} \sum_{v \in \{-1, 1\}^d} \|P_{v, +j} - P_{v, -j}\|_{TV} \leq \max_{v, j} \|P_{v, +j} - P_{v, -j}\|_{TV} .
\]

Then as long as the loss satisfies the per-coordinate separation (3.1.1), we obtain the following:

\[
\mathcal{M}(\theta(P), \Phi \circ \rho) \geq d \delta \left( 1 - \max_{v, j} \|P_{v, +j} - P_{v, -j}\|_{TV} \right) .
\] (3.1.3)

This is the version of Assouad’s lemma most frequently presented.

We also note that by the Cauchy-Schwarz inequality and convexity of the variation-distance, we have

\[
\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{TV} \leq \sqrt{d} \left( \sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{TV}^2 \right)^{1/2} \leq \sqrt{d} \left( \sum_{j=1}^{d} \frac{1}{2^d} \sum_{v} \|P_{v, +j} - P_{v, -j}\|_{TV}^2 \right)^{1/2} ,
\]
and consequently we have a not quite so terribly weak version of inequality (3.1.2):

$$
\mathcal{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta d \left[ 1 - \left( \frac{1}{d} \sum_{j=1}^{d} \sum_{v \in \{-1,1\}^d} \|P_{v,+j} - P_{v,-j}\|_{TV}^2 \right)^{1/2} \right].
$$

(3.1.4)

Regardless of whether we use the sharper version (3.1.2) or weakened versions (3.1.3) or (3.1.4), the technique is essentially the same. We simply seek a setting of the distributions $P_v$ so that the probability of making a mistake in the hypothesis test of Lemma 3.2 is high enough—say $1/2$—or the variation distance is small enough—such as $\|P_{+j} - P_{-j}\|_{TV} \leq 1/2$ for all $j$. Once this is satisfied, we obtain a minimax lower bound of the form

$$
\mathcal{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \delta \left[ 1 - \left( \frac{1}{2} \right)^{1/2} \right] = \frac{d\delta}{2}.
$$

### 3.1.3 Proof of Lemma 3.2

Fix an (arbitrary) estimator $\hat{\theta}$. By assumption (3.1.1), we have

$$
\Phi(\rho(\theta, \theta(P_v))) \geq 2\delta \sum_{j=1}^{d} \mathbb{1}\{[\hat{v}(\theta)]_j \neq v_j\}.
$$

Taking expectations, we see that

$$
\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \Phi(\rho(\hat{\theta}(X), \theta(P))) \right] \geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathbb{E}_{P_v} \left[ \Phi(\rho(\hat{\theta}(X), \theta_v)) \right]
$$

$$
\geq \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} 2\delta \sum_{j=1}^{d} \mathbb{E}_{P_v} \left[ \mathbb{1}\{[\hat{v}(\theta)]_j \neq v_j\} \right]
$$

as the average is smaller than the maximum of a set and using the separation assumption (3.1.1). Recalling the definition of the mixtures $P_{\pm j}$ as the joint distribution of $V$ and $X$ conditional on $V_j = \pm 1$, we swap the summation orders to see that

$$
\frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} P_v \left( [\hat{v}(\theta)]_j \neq v_j \right) = \frac{1}{|\mathcal{V}|} \sum_{v:v_j=1} P_v \left( [\hat{v}(\theta)]_j \neq v_j \right) + \frac{1}{|\mathcal{V}|} \sum_{v:v_j=-1} P_v \left( [\hat{v}(\theta)]_j \neq v_j \right)
$$

$$
= \frac{1}{2} P_{+j} \left( [\hat{v}(\theta)]_j \neq v_j \right) + \frac{1}{2} P_{-j} \left( [\hat{v}(\theta)]_j \neq v_j \right).
$$

This gives the statement claimed in the lemma, while taking an infimum over all testing procedures $\Psi : \mathcal{X} \to \{-1,+1\}$ gives the claim (3.1.2).

### 3.2 Example applications of Assouad’s method

We now provide two example applications of Assouad’s method. The first is a standard finite-dimensional lower bound, where we provide a lower bound in a normal mean estimation problem. For the second, we consider estimation in a logistic regression problem, showing a similar lower bound. In Chapter [4] to follow, we show how to use Assouad’s method to prove strong lower bounds in a standard nonparametric problem.
Example 3.3 (Normal mean estimation): For some $\sigma^2 > 0$ and $d \in \mathbb{N}$, we consider estimation of mean parameter for the normal location family

$$\mathcal{N} := \left\{ \mathcal{N}(\theta, \sigma^2 I_{d \times d}) : \theta \in \mathbb{R}^d \right\}$$

in squared Euclidean distance. We now show how for this family, the sharp Assouad's method implies the lower bound

$$\mathcal{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{d\sigma^2}{8n}. \tag{3.2.1}$$

Up to constant factors, this bound is sharp; the sample mean has mean squared error $d\sigma^2/n$. We proceed in (essentially) the usual way we have set up. Fix some $\delta > 0$ and define $\theta_v = \delta v$, taking $P_v = \mathcal{N}(\theta_v, \sigma^2 I_{d \times d})$ to be the normal distribution with mean $\theta_v$. In this case, we see that the hypercube structure is natural, as our loss function decomposes on coordinates: we have $\|\theta - \theta_v\|^2 \geq \delta^2 \sum_{j=1}^d \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\}$. The family $P_v$ thus induces a $\delta^2$-Hamming separation for the loss $\|\cdot\|_2^2$, and by Assouad’s method (3.1.2), we have

$$\mathcal{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^d \left[1 - \frac{1}{\mathcal{D}_{kl}(P_n^v, P_n^{v'})} \right],$$

where $P_{\pm j} = 2^{1-d} \sum_{v : v_j = \pm 1} P_n^v$. It remains to provide upper bounds on $\mathcal{D}_{kl}(P_n^v, P_n^{v'})$. By the convexity of $\|\cdot\|_2^2_{TV}$ and Pinsker’s inequality, we have

$$\mathcal{D}_{kl}(P_n^v, P_n^{v'}) \leq \frac{1}{2} \mathcal{D}_{ham}(v, v') \leq \frac{1}{2} \mathcal{D}_{ham}(v, v') \mathcal{D}_{kl}(P_n^v, P_n^{v'}).$$

But of course, for any $v$ and $v'$ differing in only 1 coordinate,

$$\mathcal{D}_{kl}(P_n^v, P_n^{v'}) = \frac{n}{2\sigma^2} \|\theta_v - \theta_{v'}\|^2 = \frac{2n}{2\sigma^2} \delta^2,$$

giving the minimax lower bound

$$\mathcal{M}_n(\theta(\mathcal{N}), \|\cdot\|_2^2) \geq \frac{2\delta^2}{2} \sum_{j=1}^d \left[1 - \sqrt{\frac{2n\delta^2}{\sigma^2}} \right].$$

Choosing $\delta^2 = \sigma^2/8n$ gives the claimed lower bound (3.2.1). ♦

Example 3.4 (Logistic regression): In this example, consider the logistic regression model, where we have known (fixed) regressors $X_i \in \mathbb{R}^d$ and an unknown parameter $\theta \in \mathbb{R}^d$; the goal is to infer $\theta$ after observing a sequence of $Y_i \in \{-1, 1\}$, where for $y \in \{-1, 1\}$ we have

$$P(Y_i = y \mid X_i, \theta) = \frac{1}{1 + \exp(-y X_i^\top \theta)}.$$

Denote this family by $\mathcal{P}_\log$, and for $P \in \mathcal{P}_\log$, let $\theta(P)$ be the predictor vector $\theta$. We would like to estimate the vector $\theta$ in squared $\ell_2$ error. As in Example 3.3, if we choose some $\delta > 0$ and for each $v \in \{-1, 1\}^d$, we set $\theta_v = \delta v$, then we have the $\delta^2$-separation in Hamming metric $\|\theta - \theta_v\|^2 \geq \delta^2 \sum_{j=1}^d \mathbb{1}\{\text{sign}(\theta_j) \neq v_j\}$. Let $P_n^v$ denote the distribution of the $n$ independent
observations \( Y_i \) when \( \theta = \theta_v \). Then we have by Assouad’s lemma (and the weakening of \ref{eq:3.2.4}) that
\[
\mathcal{M}_n(\theta(P_{\log}), \|\cdot\|_2^2) \geq \frac{\delta^2}{2} \sum_{j=1}^{d} \left[ 1 - \| P_{v,j}^n - P_{v,-j}^n \|_{TV} \right]
\geq \frac{d \delta^2}{2} \left[ 1 - \left( \frac{1}{d} \sum_{j=1}^{d} \frac{1}{2^d} \sum_{v \in \{-1,1\}^d} \| P_{v,j}^n - P_{v,-j}^n \|_{TV}^2 \right)^{1/2} \right]. \tag{3.2.2}
\]

It remains to bound \( \| P_{v,j}^n - P_{v,-j}^n \|_{TV}^2 \) to find our desired lower bound. To that end, use the shorthands \( p_v(x) = 1/(1 + \exp(\delta x^\top v)) \) and let \( D_{kl}(p\|q) \) be the binary KL-divergence between Bernoulli(\( p \)) and Bernoulli(\( q \)) distributions. Then we have by Pinsker’s inequality (recall Proposition \ref{prop:2.2}) that for any \( v, v' \),
\[
\| P_{v,j}^n - P_{v',j}^n \|_{TV} \leq \frac{1}{4} \left[ D_{kl}(P_v^n \| P_{v'}^n) + D_{kl}(P_{v'}^n \| P_{v}^n) \right] = \frac{1}{4} \sum_{i=1}^{n} \left[ D_{kl}(p_v(X_i) \| p_{v'}(X_i)) + D_{kl}(p_{v'}(X_i) \| p_v(X_i)) \right].
\]

Let us upper bound the final KL-divergence. Let \( p_a = 1/(1 + e^a) \) and \( p_b = 1/(1 + e^b) \). We claim that
\[
D_{kl}(p_a \| p_b) + D_{kl}(p_b \| p_a) \leq (a - b)^2. \tag{3.2.3}
\]

Deferring the proof of claim \ref{eq:3.2.3}, we immediately see that
\[
\| P_{v,j}^n - P_{v',j}^n \|_{TV} \leq \frac{\delta^2}{4} \sum_{i=1}^{n} \left( X_i^\top (v - v') \right)^2.
\]

Now we recall inequality \ref{eq:3.2.2} for motivation, and we see that the preceding display implies
\[
\frac{1}{2^d} \sum_{j=1}^{d} \sum_{v \in \{-1,1\}^d} \| P_{v,j}^n - P_{v,-j}^n \|_{TV}^2 \leq \frac{\delta^2}{4d} \sum_{v \in \{-1,1\}^d} \sum_{j=1}^{d} \sum_{i=1}^{n} (2X_{ij})^2 = \frac{\delta^2}{d} \sum_{i=1}^{n} \sum_{j=1}^{d} X_{ij}^2.
\]

Replacing the final double sum with \( \| X \|_{Fr}^2 \), where \( X \) is the matrix of the \( X_i \), we have
\[
\mathcal{M}_n(\theta(P_{\log}), \|\cdot\|_2^2) \geq \frac{d \delta^2}{2} \left[ 1 - \left( \frac{\delta^2}{d} \| X \|_{Fr}^2 \right)^{1/2} \right].
\]

Setting \( \delta^2 = d/4 \| X \|_{Fr}^2 \), we obtain
\[
\mathcal{M}_n(\theta(P_{\log}), \|\cdot\|_2^2) \geq \frac{d \delta^2}{4} = \frac{d^2}{16 \| X \|_{Fr}^2} = \frac{d}{n} \cdot \frac{1}{16 \sum_{i=1}^{n} \| X_i \|_{Fr}^2}.
\]

That is, we have a minimax lower bound scaling roughly as \( d/n \) for logistic regression, where \( \text{“large” } X_i \) (in \( \ell_2 \)-norm) suggest that we may obtain better performance in estimation. This is intuitive, as a larger \( X_i \) gives a better signal to noise ratio.

We now return to prove the claim \ref{eq:3.2.3}. Indeed, by a straightforward expansion, we have
\[
D_{kl}(p_a \| p_b) + D_{kl}(p_b \| p_a) = p_a \log \frac{p_a}{p_b} + (1 - p_a) \log \frac{1 - p_a}{1 - p_b} + p_b \log \frac{p_b}{p_a} + (1 - p_b) \log \frac{1 - p_b}{1 - p_a}
\]
\[
= (p_a - p_b) \log \frac{p_a}{p_b} + (p_b - p_a) \log \frac{1 - p_a}{1 - p_b} = (p_a - p_b) \log \left( \frac{p_a}{1 - p_a} \frac{1 - p_b}{p_b} \right).
\]

38
Now note that \( p_a/(1 - p_a) = e^{-a} \) and \( (1 - p_b)/p_b = e^b \). Thus we obtain

\[
D_{\text{kl}}(p_a || p_b) + D_{\text{kl}}(p_b || p_a) = \left( \frac{1}{1 + e^a} - \frac{1}{1 + e^b} \right) \log \left( \frac{e^{b-a}}{1 + e^a} \right) = (b-a) \left( \frac{1}{1 + e^a} - \frac{1}{1 + e^b} \right)
\]

Now assume without loss of generality that \( b \geq a \). Noting that \( e^x \geq 1 + x \) by convexity, we have

\[
\frac{1}{1 + e^a} - \frac{1}{1 + e^b} = \frac{e^b - e^a}{(1 + e^a)(1 + e^b)} \leq \frac{e^b - e^a}{e^b} = 1 - e^{a-b} \leq 1 - (1 + (a - b)) = b - a,
\]

yielding claim (3.2.3). ♦
Bibliography

