Chapter 18

Minimax lower bounds

Treat Assouad, Fano, and Le Cam, as in Bin Yu paper?

1. Why minimax?

Much recent asymptotic statistical literature has been devoted to questions of rates of convergence, especially for problems involving infinite dimensional parameters. At first sight the various strange rates appear driven by the analytic details of particular smoothness assumptions and other regularity conditions; it is not always readily apparent that simple probabilistic principles can explain the rates as consequences of geometric properties of the models.

The existence of a best rate of convergence depends on a requirement that an estimator do well not just at a fixed model, but also at a sequence of models that lie nearby. The rate refers not just to pointwise convergence, but rather to convergence uniformly over models in small neighborhoods of some particular model of interest. The idea is formalized as a calculation of (local) minimax risk

Let $\mathcal{P}$ be a collection of models—probability measures on some fixed measurable space $(\Omega, \mathcal{A})$. We could suppose that $\mathcal{P}$ is indexed by a parameter, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, and that we seek estimates of θ. More generally, we could consider estimates of some function of θ, a function that might not uniquely determine the whole distribution $\mathbb{P}_\theta$. In that case, it is more elegant to abandon the parametric representation altogether and treat θ as a functional on $\mathcal{P}$, that is, treat θ as a map from $\mathcal{P}$ into some metric space (Θ, d). An estimator for $\theta(\mathbb{P})$ is then a measurable map $\widehat{\theta} : \Omega \to \Theta$.

Suppose the estimator is judged by means of an expected loss, $\mathbb{P}L(\widehat{\theta}(\omega), \theta(\mathbb{P}))$, where L is a loss function on Θ^2. I will assume L is nonnegative. You could safely think of $L(t, \theta)$ as a function that increases as t moves away from θ. For example, if $\Theta = \mathbb{R}^k$ then a common choice is $L(t, \theta) = |t - \theta|^2$, quadratic loss. The mimimax criterion seeks an estimator to minimize the maximum expected loss, the ***minimax risk***,

$$\mathcal{R}(\widehat{\theta}, \mathcal{P}) := \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}L(\widehat{\theta}(\omega), \theta(\mathbb{P})).$$

If we add a subscript n to $\widehat{\theta}$, θ, $\mathcal{P}$, and maybe even L, then it makes sense to ask how rapidly $\mathcal{R}_n(\widehat{\theta}_n, \mathcal{P}_n)$ can converge to zero.

Why should we require a rate of convergence to hold uniformly in shrinking neighborhoods of a particular parameter value? That is, why should the estimator be judged by its worst behaviour along a sequence of alternatives converging, in some sense, to a fixed model? As you saw in Chapter ChapEfficiency/, the uniformity has mathematical appeal because it excludes superefficient estimators, which exploit the weaknesses in a definition influenced only by pointwise limit behaviour.

If one intends to make inferences based on asymptotic approximations, uniformity also has statistical appeal. For example, one sometimes constructs confidence intervals using an asymptotic distribution as if it were an exact distribution. The operation of inversion of (approximate) probability statements for $\mathbb{P} \in \mathcal{P}_n$ makes little sense for a fixed n unless the approximations hold uniformly in $\mathcal{P}_n$.

2. Lower bounds for minimax risks

The derivation of a minimax rate of convergence for an estimator involves a series of minimax calculations for different sample sizes. There is no initial advantage in making the dependence on the sample size explicit. Consider then the problem of finding a lower bound for the minimax risk

$$\mathcal{R}(\widehat{\theta}, \mathcal{P}) = \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}L(\widehat{\theta}(\omega), \theta(\mathbb{P})).$$

Recall the argument from Chapter 3 where lower bounds for rates of estimation were obtained via total variation distances between models. With θ regarded as a functional on $\mathcal{P}$, the result derived in that Chapter takes the following form. Suppose

<1> $$\mathbb{P}\{d(\widehat{\theta}, \theta(\mathbb{P})) \ge \delta\} \le \epsilon \qquad \text{for all } \mathbb{P} \text{ in } \mathcal{P}$$

If $\mathbb{P}_0$ and $\mathbb{P}_1$ are members of $\mathcal{P}$ for which $d(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \ge 2\delta$ and if

$$A_0 := \{\omega : d(\widehat{\theta}(\omega), \theta(\mathbb{P}_0)) < \delta\}$$

then $\mathbb{P}_0 A_0 \ge 1 - \epsilon$ and $\mathbb{P}_1 A_0 \le \epsilon$, from which it follows that

<2> $$\|\mathbb{P}_0 - \mathbb{P}_1\|_{\text{TV}} \ge \mathbb{P}_0 A_0 - \mathbb{P}_1 A_0 \ge 1 - 2\epsilon.$$

Conversely, suppose we wish to find a lower bound for

$$\mathcal{R}_0(\widehat{\theta}, \mathcal{P}) := \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\{d(\widehat{\theta}, \theta(\mathbb{P}) \ge \delta\}$$

If we can find $\mathbb{P}_0$ and $\mathbb{P}_1$ with $d(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \ge 2\delta$ and $\|\mathbb{P}_0 - \mathbb{P}_1\|_{\text{TV}} < 1 - 2\epsilon$ then, with the same A_0 as before, we can argue

$$\begin{aligned}
2\mathcal{R}_0(\widehat{\theta}, \mathcal{P}) &\ge \mathbb{P}_0\{d(\widehat{\theta}, \theta(\mathbb{P}_0) \ge \delta\} + \mathbb{P}_1\{d(\widehat{\theta}, \theta(\mathbb{P}_1) \ge \delta\} \\
&\ge \mathbb{P}_0 A_0^c + \mathbb{P}_1 A_0 \\
&\ge 1 - \sup_B |\mathbb{P}_0 B - \mathbb{P}_1 B| \\
&> 2\epsilon,
\end{aligned}$$

from which it follows that $\mathcal{R}_0(\widehat{\theta}, \mathcal{P}) > \epsilon$.

The calculation leading from <1> to <2> implicitly uses the zero-one loss function $L(t, \theta) = \{d(t, \theta) \ge \delta\}$ together with the fact that

$$L(t, \theta_0) + L(t, \theta_1) \ge 1 \qquad \text{for all } t \text{ if } d(\theta_0, \theta_1) \ge 2\delta.$$

A similar argument can be made for a general loss function if there exists a function $c_0(\cdot, \cdot)$ for which

<3> $$\inf_{t\in\Theta} \big(L(t, \theta_0) + L(t, \theta_1)\big) = c_0(\theta_0, \theta_1) > 0,$$

for each pair θ_0 and θ_1 in Θ. That is, $c(\theta_0, \theta_1)$ is the largest constant for which

<4> $$L(t, \theta_0) + L(t, \theta_1) \ge c \qquad \text{for all } t.$$

For example, if $\Theta = \mathbb{R}^k$ and $L(t, \theta) = |t - \theta|^2$ then, for all t,

$$\tfrac{1}{2}|\theta_0 - \theta_1|^2 + 2|t - (\theta_0 + \theta_1)/2|^2 \le L(t, \theta_0) + L(t, \theta_1).$$

That is, inequality <3> holds with $c_0(\theta_0, \theta_1) = \frac{1}{2}|\theta_0 - \theta_1|^2$.

In general, $c_0(\theta_0, \theta_1)$ will be an increasing function of $d(\theta_0, \theta_1)$. A lower bound for $\inf\{c_0(\theta_0, \theta_1) : \theta_0 \in \Theta_0, \theta_1 \in \Theta_1\}$ thus corresponds to a lower bound for the distance between the sets Θ_0 and Θ_1.

Inequality <4> fits with the definition (Chapter 3) of the affinity between two probability measures:

$$1 - \tfrac{1}{2}\|\mathbb{P}_0 - \mathbb{P}_1\|_1 = \alpha_1(\mathbb{P}_0, \mathbb{P}_1) := \inf_{f_0 + f_1 = 1} \mathbb{P}_0 f + \mathbb{P}_1 f_1,$$

where the infimum runs over nonnegative measurable functions for which $f_0(\omega) + f_1(\omega) = 1$ for all ω. As a consequence,

<5> $$\mathbb{P}_0 f + \mathbb{P}_1 f_1 \geq \alpha_1(\mathbb{P}_0, \mathbb{P}_1) \qquad \text{if } f_i \geq 0 \text{ and } f_1 + f_2 \geq 1.$$

In particular, if the loss function satisfies <4> and if $\theta_i = \theta(\mathbb{P}_i)$, then the choice $f_i(\omega) = L(\widehat{\theta}(\omega), \theta_i)/c(\theta_0, \theta_1)$ gives

<6> $$2\mathcal{R}(\widehat{\theta}, \mathcal{P}) \geq \mathbb{P}_0 L(\widehat{\theta}, \theta(\mathbb{P}_0)) + \mathbb{P}_1 L(\widehat{\theta}, \theta(\mathbb{P}_1)) \geq c(\theta_0, \theta_1)\alpha_1(\mathbb{P}_0, \mathbb{P}_1).$$

The result in Chapter 3 for the zero-one loss function is a special case of this inequality.

We can get a much better lower bound for the maximum risk by taking convex combinations of probability measures, then exploiting linearity of the map $\mathbb{P} \mapsto \mathbb{P}g$ for fixed g.

<7> **Definition.** *The convex hull of a set of measures $\mathcal{M}$ is the set of all finite convex combinations $\sum_i \alpha_i \mu_i$, with $\mu_i \in \mathcal{M}$ and $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. Denote the set of all such convex combinations by* $\text{co}\,(\mathcal{M})$.

REMARK. It would not be difficult to extend the calculations to more general mixtures, but the applications to minimax lower bounds seem to require nothing more than finite convex combinations.

<8> **Lemma.** *Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be subsets of $\mathcal{P}$ for which there exists a positive constant $C = C(\mathcal{P}_0, \mathcal{P}_1)$ such that*

$$L(t, \theta(\mathbb{P}_0)) + L(t, \theta(\mathbb{P}_1)) \geq C \qquad \text{for all } t \text{ if } \mathbb{P}_0 \in \mathcal{P}_0 \text{ and } \mathbb{P}_1 \in \mathcal{P}_1.$$

Then

$$2\mathcal{R}(\widehat{\theta}, \mathcal{P}) \geq C \sup\{\alpha_1(\mathbb{Q}_0, \mathbb{Q}_1) : \mathbb{Q}_i \in \text{co}\,(\mathcal{P}_i)\}$$

REMARK. If the $c_0(\theta_0, \theta_1)$ from <3> is an increasing function of $d(\theta_0, \theta_1)$, the condition on the loss function effectively sets a lower bound for $d(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1))$ when $\mathbb{P}_i \in \mathcal{P}_i$.

Proof. Define $f_i(\omega) := \inf_{\mathbb{P}_i \in \mathcal{P}_i} L(\widehat{\theta}(\omega), \theta(\mathbb{P}_i))/C$. Note that $f_i \geq 0$ and $f_0 + f_1 \geq 1$. As in the proof of <6>,

$$\begin{aligned} 2\mathcal{R}(\widehat{\theta}, \mathcal{P}) &\geq \mathbb{P}_0 L(\widehat{\theta}(\omega), \theta(\mathbb{P}_0)) + \mathbb{P}_1 L(\widehat{\theta}(\omega), \theta(\mathbb{P}_1)) \\ &\geq C\left(\mathbb{P}_0 f_0 + \mathbb{P}_1 f_1\right) \qquad \text{for all } \mathbb{P}_i \in \mathcal{P}_i. \end{aligned}$$

The right-hand side is linear in both probability measures. The inequality is preserved if we replace each $\mathbb{P}_i$ by a finite convex combination of measures from $\mathcal{P}_i$, giving probability measures $\mathbb{Q}_i$ in the convex hulls. That is,

$$2\mathcal{R}(\widehat{\theta}, \mathcal{P})/C \geq \alpha_1(\mathbb{Q}_0, \mathbb{Q}_1) \qquad \text{for all } \mathbb{Q}_i \in \text{co}\,(\mathcal{P})_i.$$

□ Take suprema over $\mathbb{Q}_i$ to complete the proof.

REMARK. If you are troubled about possible nonmeasurability of the f_i, work with finite subsets of the $\mathcal{P}_i$, then take suprema at the end of the argument.

If the loss function satisfies <3>, the Lemma suggests that we search for pairs of subsets $\mathcal{P}_i$ of $\mathcal{P}$ for which the separation in total variation,

$$\begin{aligned} d_{\text{TV}}(\mathcal{P}_0, \mathcal{P}_1) &:= \tfrac{1}{2}\inf\{\|\mathbb{Q}_0 - \mathbb{Q}_1\|_1 : \mathbb{Q}_i \in \text{co}\,(\mathcal{P}_i)\} \\ &= 1 - \sup\{\alpha(\mathbb{Q}_0, \mathbb{Q}_1) : \mathbb{Q}_i \in \text{co}\,(\mathcal{P}_i)\} \end{aligned}$$

is small, but which are well separated by the functional $\theta(\cdot)$ in the sense that

$$\inf\{c(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) : \mathbb{P}_i \in \mathcal{P}_i\}$$

is nonzero.

It may not be obvious that we gain much by introducing convex combinations into the lower bound. In fact, the gains can be substantial. A simple example with the normal distribution gives a hint of what is to come.

<9> **Example.** As shown in Chapter 3, the total variation distance between the $N(\theta, 1)$ and the $N(0, 1)$ distributions decreases like $\sqrt{2/\pi}|\theta|$ as $\theta \to 0$. More precisely,

$$\phi(x - \theta) = \phi(x) + \theta x \phi(x) + \tfrac{1}{2}\theta^2(x^2 - 1)\phi(x) + \dots$$

so that

$$\int |\phi(x-\theta) - \phi(x)|\,dx = |\theta| \int |x|\phi(x)\,dx + O(\theta^2)$$

Less precisely,

$$\|N(\theta, 1) - N(0, 1)\|_1^2 \le \int \phi(x)|\phi(x-\theta)/\phi(x) - 1|^2\,dx = \exp(\theta^2) - 1.$$

A similar argument suggests that the mixture $P_\theta = \frac{1}{2}N(\theta, 1) + \frac{1}{2}N(-\theta, 1)$ converges to the $N(0, 1)$ at an even faster rate:

$$\tfrac{1}{2}\left(\phi(x-\theta) + \phi(x+\theta)\right) = \phi(x) + \tfrac{1}{2}\theta^2(x^2 - 1)\phi(x) + \dots$$

so that

$$\int \left|\tfrac{1}{2}\phi(x-\theta) + \tfrac{1}{2}\phi(x+\theta) - \phi(x)\right|\,dx = \tfrac{1}{2}\theta^2 \int |x^2 - 1|\phi(x)\,dx + O(\theta^4)$$

Integration by parts gives $\frac{1}{2}\int |x^2 - 1|\phi(x)\,dx = 2\phi(1) \approx 0.48$.

It is not too difficult to make these calculations rigorous. The second moment bound gives the same rate of convergence even more easily.

$$\begin{aligned} \|P_\theta - P_0\|_1^2 \le P_0\left|\frac{dP_\theta}{dP_0} - 1\right|^2 &= P_0\left|\frac{dP_\theta}{dP_0}\right|^2 - 1 \\ &= \tfrac{1}{4}P_0\left|\exp\left(\theta x - \frac{\theta^2}{2}\right) + \exp\left(-\theta x - \frac{\theta^2}{2}\right)\right|^2 - 1 \\ &= \tfrac{1}{2}\left(\exp(\theta^2) + \exp(-\theta^2)\right) - 1 \\ &= \frac{\theta^4}{2!} + \frac{\theta^8}{4!} + \dots \le \exp(\tfrac{1}{2}\theta^4) - 1 \end{aligned}$$

The bound on the distance $\|P_\theta - P_0\|_1$ decreases like $\theta^2/\sqrt{2}$, an overestimate by a constant factor of approximately 1.5. □

<10> **Example.** Suppose $\mathbb{P}_\theta = \otimes_{i \le n} N(\theta_i, \sigma^2)$, where $\theta = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$. For a fixed $\xi \in \mathbb{R}_+^n$ let $\mathcal{P}_1 = \{\mathbb{P}_\theta : |\theta_i| = \xi_i \text{ for each } i\ \}$ and $\mathcal{P}_0 = \{\mathbb{P}_0\}$.

For each $\mathbb{P}_\theta$ in $\mathcal{P}_1$, the distance $\|\mathbb{P}_\theta - \mathbb{P}_0\|_1$ decreases like $|\xi|/\sigma$.

The uniform mixture over the 2^n measures in $\mathcal{P}_1$ is, in fact, also a product measure,

$$\mathbb{Q} = \otimes_{i \le n}\left(\tfrac{1}{2}N(\xi_i, \sigma^2) + \tfrac{1}{2}N(-\xi_i, \sigma^2)\right),$$

for which

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \prod\nolimits_{i\le n} \left(\tfrac12 \exp\left(\frac{\alpha_i \xi_i x_i}{\sigma^2} - \tfrac12 \frac{\xi_i^2}{\sigma^2} \right) + \tfrac12 \exp\left(\frac{-\alpha_i \xi_i x_i}{\sigma^2} - \tfrac12 \frac{\xi_i^2}{\sigma^2} \right) \right)$$

Thus

$$\begin{aligned} \mathbb{P}\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)^2 &= \mathbb{P}\prod\nolimits_{i\le n} \left(\cosh^2(\xi_i x_i/\sigma^2)\exp(-\xi_i^2/\sigma^2)\right) \\ &= \prod\nolimits_{i\le n} \tfrac12 \left(1 + \mathbb{P}\cosh(2\xi_i x_i)\right)\exp(-\xi_i^2) \\ &= \prod\nolimits_{i\le n} \cosh(\xi_i^2/\sigma^2) \\ &\le \exp\left(\sum\nolimits_{i\le n} \xi_i^4/\sigma^4 \right) \end{aligned}$$

The quadratic bound now gives

$$\|\mathbb{Q}-\mathbb{P}\|_1^2 \le \exp\left(\tfrac12 \sum\nolimits_{i\le n} \xi_i^4/\sigma^4 \right) - 1.$$

Compare with the corresponding quadratic bound for a single θ,

$$\|\mathbb{P}_\theta-\mathbb{P}\|_1^2 \le \exp\left(\tfrac12 \sum\nolimits_{i\le n} \xi_i^2/\sigma^2 \right) - 1.$$

Effectively, the averaging over the prior has replaced $\sum_{i\le n}(\xi_i/\sigma)^2$ in the exponent by $\sum_{i\le n}(\xi_i/\sigma)^4$, which can be much smaller if each $|\xi_i|/\sigma$ is small. □

3. Achievability

Section needs to be checked against Le Cam (1973). $\mathcal{L}^1$ or TV?

Le Cam (1986, p476) (also in 1973 paper? Kraft 1955) has established a direct connection between variation distances between convex hulls and the existence of randomized tests. For arbitrary sets of probability measure $\mathcal{P}_0$ and $\mathcal{P}_1$ on $(\Omega, \mathcal{A})$,

<11> $$d_{\text{TV}}(\text{co}\,(\mathcal{P}_0)\,, \text{co}\,(\mathcal{P}_1)) = 1 - \inf_{0\le\psi\le1} \sup\{\mathbb{P}_0\psi + \mathbb{P}_1(1-\psi) : \mathbb{P}_i \in \mathcal{P}_i\}.$$

The infimum on the right-hand side runs over measurable functions ψ on Ω with $0 \le \psi \le 1$. If we interpret $\psi(\omega)$ as a probability for rejecting $\mathcal{P}_0$, then $\mathbb{P}_0\psi + \mathbb{Q}(1-\psi)$ becomes a sum of the two probabilities of error with ψ as a randomized test between $\mathcal{P}_0$ and $\mathcal{P}_1$. If $d_{\text{TV}}(\text{co}\,(\mathcal{P}_0)\,, \text{co}\,(\mathcal{P}_1)) > 1 - 2\epsilon$ then there exists a randomized test ψ_0 for which

$$\mathbb{P}_0\psi_0 + \mathbb{P}_1(1-\psi_0) < 2\epsilon \qquad \text{for all } \mathbb{P}_0 \text{ in } \mathcal{P}_0 \text{ and } \mathbb{P}_1 \text{ in } \mathcal{P}_1.$$

Donohu and Liu (1991a) used a sequence of such tests to construct an estimator that comes close to achieving an accuracy of estimation, in the sense of the uniform bound <REGULARITY.AGAIN>, associated with a total variation distance between convex hulls.

4. Bounds on total variation distance

In Chapter 3 we used a simple $\mathcal{L}^2$ bound for the $\mathcal{L}^1$ distance between measures $\mathbb{P}$ and $\mathbb{Q}$ with densities p and q with respect to a probability measure λ,

$$\|\mathbb{Q}-\mathbb{P}\|_1^2 = \left(\lambda|q-p|\right)^2 \le \lambda|q-p|^2$$

In particular, for $\lambda = \mathbb{P}$, the bound reduces to $\mathbb{P}(q-1)^2 = \mathbb{P}(q^2) - 1$. This inequality is easy to handle when both $\mathbb{P}$ and $\mathbb{Q}$ are product measures, $\mathbb{P} = \otimes_{i\le n} P_i$ and $\mathbb{Q} = \otimes_{i\le n} Q_i$ with $dQ_i/dP_i = q_i$, in which case

<12> $$\|\mathbb{Q}-\mathbb{P}\|_1^2 \le -1 + \prod\nolimits_{i\le n} P_i(q_i^2)$$

There is an analogous simplification for convex combinations of product measures. The measure $\mathbb{Q}$ will not be a product measure itself, except in trivial cases. The useful factorization properties of Hellinger affinity do not apply; the Hellinger calculations are not as useful as in Chapter 3 for bounding the $\mathcal{L}^1$ distance between $\mathbb{Q}$ and $\mathbb{P}$. The $\mathcal{L}^2$ method still delivers a useful bound.

<13> **Lemma.** *Suppose $\mathbb{P} = \otimes_{i\le n} P_i$ and $\mathbb{Q} = \sum_{\alpha\in\mathbb{A}} w_\alpha \mathbb{Q}_\alpha$, a convex combination of product measures $\mathbb{Q}_\alpha = \otimes_{i\le n} Q_{i,\alpha}$ with $dQ_{i,\alpha}/dP_i = q_{i,\alpha} = q_i(\omega,\alpha)$. Then*

$$\|\mathbb{Q}-\mathbb{P}\|_1^2 \le -1 + \sum\nolimits_{\alpha,\beta\in\mathbb{A}} w_\alpha w_\beta \prod\nolimits_{i\le n} P_i\left(q_{i,\alpha}q_{i,\beta}\right)$$

Proof. Note that

$$\left(d\mathbb{Q}/d\mathbb{P}\right)^2 = \left(\sum\nolimits_{\alpha\in\mathbb{A}} w_\alpha q_{i,\alpha}\right)\left(\sum\nolimits_{\beta\in\mathbb{A}} w_\beta q_{i,\beta}\right)$$

□ Take $\mathbb{P}$ expectations, invoking the usual factorizations.

REMARK. We could use the bound

$$\|\mathbb{Q}-\mathbb{P}\|_1 \le \sum\nolimits_{\alpha\in\mathbb{A}} \|\mathbb{Q}_\alpha - \mathbb{P}\|_1 \le \max_{\alpha\in\mathbb{A}} \|\mathbb{Q}_\alpha - \mathbb{P}\|_1,$$

but that would wipe out the effect of any cancellations due to the averaging over $\mathbb{A}$.

If the prior is degenerate, the upper bound from the Corollary reduces to the bound from inequality <12>

<14> **Example.** A key step in a beautiful calculation by Mammen (1986) was the bounding of the total variation distance between a product measure $\mathbb{P} = P^n$ and a mixture

$$\mathbb{Q} = \frac{1}{n}\sum_{\alpha=1}^{n} \mathbb{Q}_\alpha \qquad \text{where } \mathbb{Q}_\alpha := P^{\alpha-1}\otimes Q \otimes P^{n-\alpha}$$

Mammen used the second moment method with dominating measure $(P+Q)/2$ to obtain an upper bound in terms of $H^2(P,Q)$. (See Problem [2] for his bound.) A similar bound is even easier to derive when Q has density $1+\Delta$ with respect to P, with $P\Delta^2 < \infty$.

In the notation of Lemma <13>, we have

$$q_{i,\alpha} = \begin{cases} 1+\Delta(x_i) & \text{if } i=\alpha \\ 1 & \text{otherwise}\end{cases}$$

Thus

$$\mathbb{P}q_{i,\alpha}q_{i,\beta} = \begin{cases} 1+P\Delta^2 & \text{if } i=\alpha=\beta \\ 1 & \text{otherwise}\end{cases}$$

The assertion of the Lemma simplifies to

$$\|\mathbb{Q}-\mathbb{P}\|_1^2 \le \frac{1}{n^2}\sum_{\alpha=1}^{n} P\Delta^2 = \frac{P\Delta^2}{n}$$

In the typical case where Δ is bounded, so that $P\Delta^2$ is of the same order of magnitude as $H^2(P,Q)$, and $H(P,Q)$ is of order $O(1/\sqrt{n})$, the bound $\sqrt{P\Delta^2/n}$ for $\|\mathbb{Q}-\mathbb{P}\|_1$ converges to zero at a $1/n$ rate. By a direct calculation of $\mathcal{L}^1$ norms,

$$\|\mathbb{Q}_a - \mathbb{P}\|_1 = \|Q-P\|_1$$

which in typical parametric situations converges to zero at only a $O(1/\sqrt{n})$ rate. The mixing greater improves the rate of convergence. □

5. Quadratic functional of Gaussian means

For linear functionals of models, the task of distinguishing between pairs of models is often difficult enough to determine the best minimax rate of convergence, as in Chapter 3. For only slightly more complicated functionals, more difficult tasks are needed. The simplest, and most explored, case involves estimation of a quadratic function of the sequence $\eta = (\eta_1, \eta_2, \ldots)$ of means of a set $x_1, x_2 \ldots$ of independent $N(\eta_i, \sigma^2)$ random variables. The asymptotic theory involves a limit as σ tends to zero.

Several surprisingly rich problems have been studied in detail. A typical example is estimation of the functional

$$\theta(\eta) := \sum\nolimits_i \beta_i \eta_i^2,$$

subject to the inequality constraints

$$|\eta_i| \le A_i \qquad \text{for each } i,$$

where $\{\beta_i\}$ and $\{A_i\}$ are given decreasing sequences. Write $\mathcal{E}$ for the constraint set. With squared error loss, $L(t,\theta) = (t-\theta)^2$, the minimax risk is

$$\mathcal{R}(\widehat{\theta}, \mathcal{E}) = \sup_{\eta \in \mathcal{E}} \mathbb{P}_{\eta,\sigma}(\widehat{\theta} - \theta(\eta))^2.$$

The interest lies in determining the rate at which the risk decreases as a function of σ.

For the particular case $\beta_i = i^{2k}$ and $A_i = i^{-\alpha}$, for various positive α, Fan (1991) explained a surprising cutoff phenomenon. The minimax rate is of order

Find the correct values.

<15> $$\begin{cases} \sigma^4 & \text{if } \alpha > 2k + 1/4 \\ \sigma^{8(\alpha-k)/(4\alpha+1)} & \text{if } k < \alpha < 2k + 1/4. \end{cases}$$

The bound is due to the combined effect of two mechanisms for contolling the rate. Each contributes a lower bound, obtained as the solution to a convex optimization problem. The maxima of the two bounds gives the achievable rate.

It is most instructive to compare these two bounds by means of their consequences for a minimax rate as $\sigma \to 0$.

At $\mathbb{Q}_0$ the functional θ takes the value 0. At each $\mathbb{Q}_\lambda$ it takes the value $\theta(\xi) = \sum_i \beta_i \xi_i^2$. For quadratic loss, $c(\theta(\mathbb{Q}_0), \theta(\mathbb{Q}_\lambda)) = 1/2\,\theta(\xi)^2$. If we take $\mathbb{P}_0 = \mathbb{Q}_0$ and $\mathbb{P}_1 = \mathbb{Q}_\lambda$ in Lemma <13:5> we get

$$\mathcal{R}(\widehat{\theta}, \mathcal{E}) := \sup_{\xi \in \mathcal{E}} \mathbb{P}_{\eta,\sigma}|\widehat{\theta} - \theta(\xi)|^2 \ge 1/4\,\theta(\xi)^2 \alpha(\mathbb{Q}_0, \mathbb{Q}_\lambda),$$

where

$$\alpha(\mathbb{Q}_0, \mathbb{Q}_\lambda) = 1 - 1/2\|\mathbb{Q}_\lambda - \mathbb{Q}_0\|_1.$$

If we take $\mathcal{P}_0 = \{\mathbb{Q}_0\}$ and $\mathcal{P}_1 = \{\mathbb{Q}_\lambda : \lambda \in \Lambda\}$ in Lemma <13:7> we get an analogous lower bound with the affinity replaced by

$$\alpha(\mathcal{P}_0, \mathcal{P}_1) \ge 1 - 1/2\|\mathbb{Q} - \mathbb{Q}_0\|_1.$$

The following strategy now suggests itself. Use inequalities <SINGLE.L1> and <CONVEX.L1> to keep the affinity above a certain constant level, such as $1 - 1/2\sqrt{e-1}$. (Where does that number come from?) This lower bound places a constraint on ξ, in addition to the constraint that defines $\mathcal{E}$. Then maximize $\theta(\xi)$ subject to the constraints. More specifically, the two-point (degenerate prior) case corresponds to the problem:

$$\begin{array}{rll} \text{maximize} & \sum_i \beta_i \xi_i^2 & \\ \text{subject to} & \sum_i \xi_i^2/\sigma^2 \le 1 & \\ \text{and} & |\xi_i| \le A_i & \text{all } i \end{array}$$

and the convex-hull case corresponds to the problem

$$\begin{array}{rll} \text{maximize} & \sum_i \beta_i \xi_i^2 & \\ \text{subject to} & {}^1\!/\!_2 \sum_i \xi_i^4/\sigma^4 \le 1 & \\ \text{and} & |\xi_i| \le A_i & \text{all } i \end{array}$$

Both problems have explicit, closed form solutions.

Two-point problem

This case is a trivial linear programming problem in $y_i = \xi_i^2$. The solution is $\xi_i^2 = A_i$ for all i less some i_0, and equal to zero for i larger than i_0, with ξ_{i_0} chosen to achieve equality in the first constraint. When $\sigma^2 \le A_1$, the solution degenerates to $\xi_1 = \sigma$ with all other ξ_i equal to zero. When σ is small, the task set for the estimator is decide whether the single observation x_1 comes from a $N(0, \sigma^2)$ or a $N(\sigma, \sigma^2)$ distribution. According to Lemma <13:5>, the minimax risk is then greater than $\beta_1^2\sigma^4$??. As shown in Problem [EXACT.MINIMAX], an exact calculation is possible for this trivial problem; the exact lower bound is σ^4??. The two-point case provides the σ^4 branch of the lower bound <15>.

Calculate exact mmax risk.

Convex-hull problem

This case reduces to a neat quadratic programming problem after substitution $\xi_i = \sqrt{y_i\beta_i}$:
Maximize $\sum_i \beta_i^2 y_i$ subject to the constraints:
(i) $\sum_i \beta_i^2 y_i^2 \le 2\sigma^4$;
(ii) $0 \le y_i \le A_i^2/\beta_i$ for all i.

<16> **Lemma.** *Suppose $\{\gamma_i\}$ and $\{B_i\}$ are sequences of strictly positive numbers for which $\sum_i \gamma_i B_i < \infty$ and $\sum_i \gamma_i B_i^2 < \infty$. Let $y_i = B_i \wedge t$, where t is the largest value (possibly $+\infty$) for which*

$$\sum_i \gamma_i y_i^2 = \sum_i \gamma_i (B_i \wedge t)^2 \le C.$$

Then $\{y_i\}$ maximizes $\theta(\mathbf{y}) = \sum_i \gamma_i y_i$ subject to the constraints
(i) $\sum_i \gamma_i y_i^2 \le C$, for a given positive C;
(ii) $0 \le y_i \le B_i$ for all i.

Proof. If $y_{i'} = B_{i'}$ for all i' then clearly $\theta(\mathbf{x})$ is the maximum. We may therefore assume that $x_{i_0} = t < B_{i_0}$ for at least one i_0. In that case, equality must be achieved in constraint (i), for otherwise an increase in t would increase $\theta(\mathbf{y})$. We must also have each $y_{i'}$ equal to either t or $B_{i'}$, whichever is smaller.

The feasible set is compact for pointwise convergence. The function θ must achieve its finite supremum at some feasible $\mathbf{x}$. Suppose $\theta(\mathbf{x}) > \theta(\mathbf{x})$. For some i we must have $B_i \ge x_i > y_i = B_i \wedge t$, which forces $y_i = t \ge y_{i'}$ for all i'. If $x_{i'} \geq y_{i'}$ for all i', constraint (i) would be violated. Thus $x_j < y_j$ for some j. Note that $x_i > t \geq x_j$.

Consider small perturbation of these two coordinates. Choose a constant $\tau > 1$ with $x_i > \tau x_j$. For small $\epsilon > 0$ consider the effect of replacing x_i by $x_i - \epsilon/\gamma_i$ and x_j by $x_j + \tau\epsilon/\gamma_j$. If ϵ is small enough, the new vector is feasible, because

$$\gamma_i x_i^2 + \gamma_j x_j^2 - \gamma_i(x_i - \epsilon/\gamma_i)^2 - \gamma_j(x_j + \tau\epsilon/\gamma_j)^2 = 2\epsilon(x_i - \tau x_j) + O(\epsilon^2).$$

The coefficient of ϵ is strictly positive; for small enough $\epsilon > 0$ constraint (i) still holds. The modified $\mathbf{x}$ is still feasible, but the change increases $\theta(\mathbf{x})$ by $(\tau - 1)\epsilon > 0$: a contradiction. □

Question: What happens if constraint (ii) is replaced by $\sum_i \delta_i x_i^p \leq C'$, for some constants $p > 0$ and C'?

When σ is small enough, the solution to the convex-hull problem is $y_i = (A_i^2/\beta_i) \wedge t$, where

$$\sum_i A_i^4 \wedge (\beta_i^2 t^2) = 2\sigma^4.$$

The maximizing ξ is given by $\xi_i = A_i \wedge \sqrt{\beta_i t}$.

For the special case $\beta_i = i^{2k}$ and $A_i = i^{-\alpha}$, the ratio A_i^2/β_i is decreasing. Thus $y_i = t$ for $i \leq m$ and $y_i = A_i^2/\beta_i$ for $i > m$, where m is the solution to

$$What?$$

6. Quadratic functionals of densities

Editing needed

Suppose independent observations are taken from a distribution P having density f concentrated on [0, 1]. Consider the functional (slight abuse of notation here)

$$\theta(f) = \int_0^1 f'(x)^2\,dx \qquad \text{OR} \int_0^1 f(x)^2\,dx?$$

Suppose f is constrained to lie in the smoothness class LIP of functions on [0, 1] for which

$$\sup_{0\leq x\leq 1} |f^{(i)}(x)| \leq C \qquad \text{for } i = 1, \ldots, k,$$

and

$$|f^{(k)}(x) - f^{(k)}(y)| \leq C|x - y|^\nu \qquad \text{for } 0 \leq x, y \leq 1.$$

Here C is a fixed constant, k is a positive integer, and ν is such that $0 < \nu \leq 1$ and $s = k + \nu > 1$.

<17> **Definition.** *Let s and δ be strictly positive. Let k be the largest integer strictly less than s and let $s = k + \alpha$, with $0 < \alpha \leq 1$. For a function f defined and k times differentiable at least on the interval $(-\delta, \delta)$, define the norm $\|f\|_{s,\delta}$ as the smallest constant C (possibly infinite) for which*

$$\sup_{|t|<\delta} |f^{(i)}(t)| \leq C \qquad \text{for } i = 0, 1, \ldots, k$$
$$|f^{(k)}(t_1) - f^{(k)}(t_2)| \leq C|t_1 - t_2|^{\alpha} \qquad \text{for } |t_1|, |t_2| < \delta.$$

Call f locally s-smooth if $\|f\|_{s,\delta}$ is finite for some positive δ. Write $\mathfrak{S}_s(0,\delta,C)$ for the class of such functions with $\|f\|_{s,\delta} \leq C$.

Notice that finiteness of $\|f\|_{1,\delta}$ requires only a Lipschitz condition on f near the origin.

Let f_0 denote the uniform density on $[-1/2, 1/2]$. For a fixed small $\delta > 0$ and a constant $C > 1$, let us work with a smoothness class $\mathfrak{S}_s = \mathfrak{S}_s(0,\delta,C)$ for the remainder of the Section. Of course $f_0 \in \mathfrak{S}_s$. Let P be the Uniform[0, 1] distribution, for which (more abuse of notation) $\theta(P) = 0$.

The value of m will be chosen later. For each i let $g_i = (g_{i1}, \ldots, g_{im})$ be an m-vectors of orthogonal functions on [0, 1] with $\mathbb{Q}_0 g_i = 0$ and finite second moment matrix $\mathbb{Q}_0 g_i g_i' = \Gamma_i$. Define

$$G_i(x_i, \lambda) = \lambda' g_i(x_i) = \sum_{j=1}^{m} \lambda_j g_{ij}(x_i).$$

Then $\tau_i(\lambda, \mu) = \lambda' \Gamma_i \mu$ and

$$\mathbb{P}\prod_{i\leq n}(1 + \tau_i(\lambda,\mu)) = \mathbb{P}\prod_{i\leq n}\left(1 + \lambda'\Gamma_i\mu\right) \leq \mathbb{P}\exp(\lambda'\sum_i \Gamma_j \mu).$$

The random variable $\lambda' \sum_i \Gamma_j \mu$ has a symmetric distribution. If you expanded the product and took expectations term by term you would observe that odd powers are wiped out. The exponential bound achieves the same effect (Problem [1]). The effect is particularly easy to see if the components of g_i are orthogonal in $L^2(\mathbb{Q}_0)$, with $\Gamma_i = \sigma_i^2 I_m$, for then the exponent becomes cW, with $c = \sum_i \sigma_i^2$ and $W = \lambda'\mu$, a sum of m independent Rademacher variables.

$$\mathbb{P}\exp(cW) = \left(1/2 e^{c} + 1/2 e^{-c}\right) m \leq \exp\left(1/2 m c^2\right).$$

In summary, with the orthogonal components, inequality <CONVEX.BOUND> gives

$$\|\mathbb{Q} - \mathbb{Q}_0\|_1^2 \leq \exp\left(1/2 m (\sum_i \sigma_i^2)^2\right) - 1.$$

Notice that the bound is of order $O(m(\sum_i \sigma_i^2)^2)$ near zero.

Theorem <QUADRATIC.BOUND> will show that the best rate of uniform convergence, in the sense of inequality <UNIFORM.RATE>, cannot be faster than $O_p(r_n)$ where

$$r_n = n^{-(4s-4)/(4s+1)}.$$

Construct alternative models starting from a fixed, $k+1$-times differentiable function g with support in [0, 1], and such that $Pg = 0$ and $Pg^2 = \tau > 0$ and $P(g')^2 = \beta > 0$. Partition [0, 1] into disjoint intervals of length $1/m$ with centers at $b_1, \ldots, b_m$. Define

$$g_\alpha(x) = cm^{-s} g(mx - b_\alpha),$$

for a small, positive constant c, to be determined. If c is small enough, the perturbed densities $1+\sum_\alpha \lambda_\alpha g_\alpha$, corresponding to the corners of the hypercube, lie in LIP(s, C). Check that

$$Pg_\alpha^2 = c^2\tau m^{-(2s+1)},$$

from which

$$n^2\sum_\alpha (Pg_\alpha^2)^2 = c^4\tau^2 n^2 m^{-(4s+1)}.$$

Let m increase like $n^{2/(4s+1)}$, to keep the last expression bounded. Given $\epsilon > 0$, choose a c small enough to ensure that

$$\limsup_n \exp\left(\tfrac{1}{2}n^2\sum_\alpha (Pg_\alpha^2)^2\right) - 1 < \big(2(1-2\epsilon)\big)^2,$$

as required by <RATE.DETERMINING>. Define $D_{n1} = \{0\}$ and $D_{n2} = \{\gamma_n\}$, where

$$\gamma_n = \theta(1+\sum_\alpha \lambda_\alpha g_\alpha) = \sum_\alpha P(g_\alpha')^2 = c^2\beta m^{-(2s-2)}, \qquad \text{for all } \boldsymbol{\lambda}.$$

Choose K small enough to ensure that $2Kr_n < \gamma_n$, which makes the sets D_{n1} and D_{n2} at least $2Kr_n$-separated. Then the asserted lower bound for the rate follows.

7. Notes

Hasminskii (1979)

The basic idea was explained by Le Cam—see his 1973 paper in particular—with recent embellishments and extensions due to

Donoho & Liu (1987) and Donoho & Liu (1991).

Hall and Marron.

Donoho, Liu and MacG.

Donoho and Nussbaum.

Ingster.

Bickel & Ritov (1988), Ritov & Bickel (1990), Birgé & Massart (1992) Donoho & Liu (1991) Bretagnolle & Huber (1979) Le Cam (1973) Ritov & Bickel (1990) Fan (1991)

Ingster (1986) for quadratic bound.

Huber (1997), Yu (1997) for Assouad et al?

Bretagnolle & Huber (1979)

8. Problems

[1] Let λ and μ be independent m-vectors of Rademachers, and let Γ be an $m\times m$ matrix.

(i) Show that $\mathbb{P}\exp(\lambda'\Gamma\mu) \le \mathbb{P}\exp(Y'\Gamma Z/2\pi)$, with Y and Z independent $N(0, I_m)$ random vectors.

(ii) By means of a conditioning argument, followed by a diagonalization of a matrix, show that the last expectation equals

$$\mathbb{P}\exp\left(\sum_i \theta_i Z_i^2\right) = \prod_i (1-2\theta_i)^{-1/2} \qquad \text{if } \max\theta_i < 1/2,$$

where the θ_i are the eigenvalues of the symmetric matrix $\Gamma'\Gamma/2\pi$.

(iii) Show the last bound is less than $\exp\big(-\text{trace}(\Gamma'\Gamma)/\pi\big)$, if $\max\theta_i < 1/4$.

calculations need checking

[2] For $\mathbb{P}$ and $\mathbb{Q}$ as in Example <14>, bound $\|\mathbb{Q}-\mathbb{P}\|_1^2$ using the second moment method for densities with respect to $\lambda = (P+Q)/2$, by following these steps. Write $q = dQ/d\lambda = 1+\Delta$, so that $p = dP/d\lambda = 1-\Delta$. Write κ for $\lambda\Delta^2$. Note that, by <QUADRATIC.AVERAGE>, $\kappa \le H^2(P,Q)$.

(i) Show that $\Delta_{0,i} = -\Delta$ for all i, and that $\Delta_{\alpha,i} = \Delta$ if $i=\alpha$, and $-\Delta$ otherwise.

(ii) Deduce that $\tau_{0,0}(i) = \kappa$ for all i; that $\tau_{\alpha,\beta}(i) = -\kappa$ if $\alpha \ne \beta = i$ or $\beta \ne \alpha = i$, and κ otherwise; and that $\tau_{\alpha,0}(i) = -\kappa$ if $\alpha = i$, and κ otherwise.

(iii) Deduce that $1+\Psi(\tau_{0,0}) = (1+\kappa)^n$; that $1+\Psi(\tau_{\alpha,0}) = (1+\kappa)^{n-1}(1-\kappa)$; and that

$$1+\Psi(\tau_{\alpha,\beta}) = \begin{cases} (1+\kappa)^{n-2}(1-\kappa)^2 & \text{if } \alpha \ne \beta \\ (1+\kappa)^{n-1}(1-\kappa) & \text{if } \alpha = \beta \end{cases}$$

(iv) Deduce that

$$\lambda(q-p)^2 \le 4\kappa\left(\frac{1}{n}+\kappa\right)(1+\kappa)^{n-2}$$

Compare with Mammen (1986, inequality 3.7).

[3] Bound distance between $\sum_\alpha w_\alpha \mathbb{Q}_\alpha$ and $\mathbb{Q}_0$, where all $\mathbb{Q}_\alpha$ are product measures dominated by a product measure $\mathbb{P}$. Suppose

$$\frac{d\mathbb{Q}_\alpha}{d\mathbb{P}} = \prod_{i\le n}(1+\Delta_{\alpha,i}(x_i)),$$

and

$$\frac{d\mathbb{Q}_0}{d\mathbb{P}} = \prod_{i\le n}(1+\Delta_{0,i}(x_i)).$$

Show

$$\begin{aligned}\|\textstyle\sum_{\alpha\in\mathbb{A}} w_\alpha \mathbb{Q}_a - \mathbb{Q}_0\|_1^2 &\le \sum_{\alpha\in\mathbb{A}}\sum_{\beta\in\mathbb{A}} w_\alpha w_\beta\left[\Psi(\tau_{\alpha,\beta}) - 2\Psi(\tau_{\alpha,0}) + \Psi(\tau_{0,0})\right] \\ &= \sum_{\alpha\in\mathbb{A}}\sum_{\beta\in\mathbb{A}} w_\alpha w_\beta \Psi(\tau_{\alpha,\beta}) - 2\sum_{\alpha\in\mathbb{A}} w_\alpha \Psi(\tau_{\alpha,0}) + \Psi(\tau_{0,0})\end{aligned}$$

References

Bickel, P. J. & Ritov, Y. (1988), 'Estimating integrated squared density derivatives: sharp best order of convergence estimates', *Sankhyā: The Indian Journal of Statistics, Series A* **50**, 381–393.

Birgé, L. & Massart, P. (1992), Estimation of integral functionals of a density, Technical Report 024-92, Mathematical Sciences Research Institute, Berkeley.

Bretagnolle, J. & Huber, C. (1979), 'Estimation des densites: risque minimax', *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **47**, 119–137.

Donoho, D. & Liu, R. C. (1987), Geometrizing rates of convergence, I, Technical report, University of California, Berkeley. Department of Statistics, Technical Report 137.

Donoho, D. L. & Liu, R. C. (1991), 'Geometrizing rates of convergence, II', *Annals of Statistics* **19**, 633–667.

Fan, J. (1991), 'On the estimation of quadratic functionals', *Annals of Statistics* **19**, 1273–1294.

Hasminskii, R. Z. (1979), 'Lower bound for the risks of nonparametric estimates of the mode', pp. 91–97.

Huber, C. (1997), Lower bounds for function estimation, *in* D. Pollard, E. Torgersen & G. L. Yang, eds, 'A Festschrift for Lucien Le Cam', Springer-Verlag, New York, pp. 245–258.

Ingster, Y. I. (1986), 'Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics', *Theory Probability and Its Applications* **31**, 333–337.

Le Cam, L. (1973), 'Convergence of estimates under dimensionality restrictions', *Annals of Statistics* **1**, 38–53.

Mammen, E. (1986), 'The statistical information contained in additional observations', *Annals of Statistics* **14**, 665–678.

Ritov, Y. & Bickel, P. J. (1990), 'Achieving information bounds in non and semiparametric models', *Annals of Statistics* **18**, 925–938.

Yu, B. (1997), Assouad, Fano, and Le Cam, *in* D. Pollard, E. Torgersen & G. L. Yang, eds, 'A Festschrift for Lucien Le Cam', Springer-Verlag, New York, pp. 423–435.