Abstract

Minimax Lower Bounds

Adityanand Guntuboyina

2011

This thesis deals with lower bounds for the minimax risk in general decision-theoretic problems. Such bounds are useful for assessing the quality of decision rules. After providing a unified treatment of existing techniques, we prove new lower bounds which involve f-divergences, a general class of dissimilarity measures between probability measures. The proofs of our bounds rely on elementary convexity facts and are extremely simple. Special cases and straightforward corollaries of our results include many well-known lower bounds. As applications, we study a covariance matrix estimation problem and the problem of estimation of convex bodies from noisy support function measurements.

Minimax Lower Bounds

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Adityanand Guntuboyina

Dissertation Director: David B. Pollard

December 2011

Contents

Dedicated to my parents and younger brother.

Acknowledgements

I am indebted to my advisor David Pollard from whom I have learned as much about research as about writing, teaching, grading, giving talks and even biking. In addition to him, my research interests have been greatly influenced by those of Andrew Barron, Harrison Zhou and Mokshay Madiman. I am extremely thankful to them and also to Hannes Leeb with whom I wrote my first real paper.

I have been exceedingly fortunate to have many delightful friends from the Statistics Department and Helen Hadley Hall. They are the sole reason why the past five years have been so thoroughly enjoyable in spite of the rigours of the PhD program.

Chapter 1

Introduction

In statistical decision theory, a widespread way of assessing the quality of a given decision rule is to compare its maximum possible risk to the minimax risk of the problem. One uses the maximum risk of the decision rule as opposed to working with its risk directly because the risk typically depends on the unknown parameter. It is however typically impossible (especially in nonparametric problems) to determine the minimax risk exactly. Consequently, one attempts to obtain good lower bounds on the minimax risk and the maximum risk of a given decision rule is then compared to these lower bounds. Lower bounds on the minimax risk are the subject of this thesis.

Chapter 2 provides a unified view of the techniques commonly used in the literature to establish minimax bounds. We explain why techniques due to Le Cam, Assouad and Fano are all simple consequences of a well known expression for the Bayes risk in general decision-theoretic problems.

In Chapter 3, we prove a class of lower bounds for the minimax risk (one for each convex function f) using f-divergences between the underlying probability measures. The f-divergences are a general class of measures of dissimilarity between probability measures. Kullback-Leibler divergence, chi-squared divergence, total variation distance distance and Hellinger distance are all special cases of f-divergences. The proof of our bound is extremely simple: it is based on an elementary pointwise inequality and a couple of applications of Jensen's inequality. Special cases and straightforward corollaries of our bound include well-known minimax lower bounds like Fano's inequality and Pinsker's inequality.

We also generalize a technique of Yang and Barron (1999) for obtaining minimax lower bounds using covering and packing numbers of the whole parameter space. The results in Yang and Barron (1999), which are based on Kullback-Leibler divergences, have been successfully applied to several nonparametric problems with very large (infinite-dimensional) parameter spaces. On the other hand, for finite dimensional problems, their results usually produce sub-optimal rates, which lends support to the statistical folklore that global covering and packing numbers alone are not enough to recover classical parametric rates of convergence. As Chapter 3 shows, the folklore is wrong as far as lower bounds are concerned: with a different f-divergence (chi-squared), the analogue of the results of Yang and Barron (1999) does give the correct rate for several finite dimensional problems.

Remark 1.0.1. *After the paper Guntuboyina (2011), on which Chapter 3 is based, was accepted, Professor Alexander Gushchin pointed out to me that one of the main theorems of Chapter 3 appears in his paper, Gushchin (2003). The details of the overlap with Gushchin's paper are described in Section 3.4 of Chapter 3.*

In Chapter 4, we illustrate the use of the bounds from Chapter 3 by means of an application to a covariance matrix estimation problem, which was recently studied by Cai, Zhang, and Zhou (2010).

Chapter 5 presents another illustration of our bounds. We study the problem of estimating a compact, convex set from noisy support function measurements. We improve results due to Gardner, Kiderlen, and Milanfar (2006) by identifying the correct (achievable) minimax rate for the problem.

Bibliography

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics 38*, 2118–2144.

Gardner, R. J., M. Kiderlen, and P. Milanfar (2006). Convergence of algorithms for reconstructing convex bodies and directional measures. *Annals of Statistics 34*, 1331–1374.

Guntuboyina, A. (2011). Lower bounds for the minimax risk using f divergences, and applications. *IEEE Transactions on Information Theory 57*, 2386–2399.

Gushchin, A. A. (2003). On Fano's lemma and similar inequalities for the minimax risk. *Theor. Probability and Math. Statist. 67*, 29–41.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Chapter 2

Standard Minimax lower bounds

2.1 Introduction

This chapter reviews commonly used methods for bounding the minimax risk from below in statistical problems. We work in the standard decision-theoretic setting (see Ferguson, 1967, Chapter 1). Let Θ and $\mathcal{A}$ denote the parameter space and action space respectively with the (non-negative) loss function denoted by $L(\theta, a)$. We observe X whose distribution P_θ depends on the unknown parameter value. It is assumed that P_θ is a probability measure on a space $\mathcal{X}$ having a density p_θ with respect to a common dominating sigma finite measure μ. (Nonrandomized) Decision rules are functions mapping $\mathcal{X}$ to $\mathcal{A}$. The risk of a decision rule $\mathfrak{d}$ is defined by $\mathbb{E}_\theta L(\theta, \mathfrak{d}(X))$, where $\mathbb{E}_\theta$ denotes expectation taken under the assumption that X is distributed according to P_θ. The minimax risk for this problem is defined by

$$R_{\text{minimax}} := \inf_{\mathfrak{d}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \mathfrak{d}(X))$$

We first prove a general minimax lower bound that is based on a classically known exact expression for the Bayes risk in decision-theoretic problems. We then demonstrate that standard lower bound techniques due to Le Cam, Assouad and Fano can all be viewed as simple corollaries of this general bound. Previously (see, for example, Yu, 1997 and Tsybakov, 2009, Chapter 2), these three techniques have been treated separately.

2.2 General Minimax Lower Bound

The minimax risk R_{minimax} is bounded from below by the Bayes risk with respect to every proper prior. Let w be a probability measure on Θ. The Bayes risk with respect to w is defined by

$$R_{\text{Bayes}}(w) := \inf_{\mathfrak{d}} \int_{\Theta} \mathbb{E}_{\theta} L(\theta, \mathfrak{d}(X)) w(d\theta).$$

The inequality $R_{\text{minimax}} \geq R_{\text{Bayes}}(w)$ holds for every w. The decision rule $\mathfrak{d}$ for which $R_{\text{Bayes}}(w)$ is minimized can be determined as a posterior expected loss given the data (Lehmann and Casella, 1998, page 228), which results in an exact expression for $R_{\text{Bayes}}(w)$. Indeed, for every $\mathfrak{d}$, assuming conditions for interchanging the order of integration, we have

$$\int_{\Theta} \mathbb{E}_{\theta} L(\theta, \mathfrak{d}(X)) w(d\theta) = \int_{\mathcal{X}} \int_{\Theta} L(\theta, \mathfrak{d}(x)) p_{\theta}(x) w(d\theta) \mu(dx) \geq \int_{\mathcal{X}} B_{w,L}(x) \mu(dx)$$

where $B_{w,L}(x) := \inf_{a \in \mathcal{A}} B_{w,L}^{a}(x)$ and $B_{w,L}^{a}(x) := \int_{\Theta} L(\theta, a) p_{\theta}(x) w(d\theta)$. Morever, equality is achieved for $\mathfrak{d}(x) := \operatorname{argmin}_{a \in \mathcal{A}} B_{w,L}(x, a)$. Thus $R_{\text{Bayes}}(w)$ is equal to $\int_{\mathcal{X}} B_{w,L}(x) \mu(dx)$ and we have the following minimax lower bound:

$$R_{\text{minimax}} \geq \int_{\mathcal{X}} B_{w,L}(x) \mu(dx) \qquad \text{for every } w. \tag{2.1}$$

2.3 Review of Standard Techniques

Standard lower bound techniques including those of Assouad, Le Cam and Fano are reviewed here. These bounds are well-known but we shall provide simple proofs using the general bound (2.1). Our main point is that each of these bounds is a special case of (2.1) for a particular choice of the prior w. In fact, all minimax lower bound techniques that I know are based on bounding from below the Bayes risk with respect to a prior w. Since the right hand side of (2.1) is exactly equal to the Bayes risk under w, other minimax lower bound techniques that we do not discuss in this chapter (e.g., Massart, 2007, Corollary 2.18 and Cai and Low, 2011, Corollary 1) can also be derived from (2.1).

In the sequel, the following notions are often used:

1. $d(\theta_1, \theta_2) := \inf\{L(\theta_1, a) + L(\theta_2, a) : a \in \mathcal{A}\}$ for $\theta_1, \theta_2 \in \Theta$.

2. $d(\Theta_1, \Theta_2) := \inf\{d(\theta_1, \theta_2) : \theta_1 \in \Theta_1, \theta_2 \in \Theta_2\}$ for subsets Θ_1 and Θ_2 of Θ.

3. We say that a finite subset F of Θ is η-separated if $d(\theta_1, \theta_2) \geq \eta$ for all $\theta_1, \theta_2 \in F$ with $\theta_1 \neq \theta_2$.

4. For finitely many probability measures $P_1, \dots, P_N$ on $\mathcal{X}$ and weights $\rho_i \geq 0, \sum_{i=1}^N \rho_i = 1$, we define

$$\bar{r}_\rho(P_1, \dots, P_N) := 1 - \int_{\mathcal{X}} \max_{1 \leq i \leq N} [\rho_i p_i(x)] \, \mu(dx) \qquad \text{where } p_i := dP_i/d\mu.$$

   When the probability measures $P_1, \dots, P_N$ are clear from the context, we just write $\bar{r}_\rho$. Also, when $\rho_i = 1/N$, we simply write $\bar{r}(P_1, \dots, P_N)$ or $\bar{r}$.

5. Hamming distance on the hypercube $\{0, 1\}^m$: $\Upsilon(\tau, \tau') = \sum_{i=1}^m \{\tau_i \neq \tau_i'\}$.

6. The total variation distance $||P-Q||_{TV}$ between two probability measures P and Q is defined as $\frac{1}{2}\int_{\mathcal{X}}|p-q|d\mu$ where p and q denote the densities of P and Q with respect to μ.

7. Testing affinity $||P \wedge Q||_1 := \int (p \wedge q) d\mu = 2\bar{r}(P,Q) = 1 - ||P-Q||_{TV}$.

8. Kullback-Leibler divergence, $D_1(P||Q) = \int p \log(p/q) d\mu$. We use D_1 for the Kullback-Leibler divergence because it is a member (for $\alpha = 1$) of a family of divergences D_α introduced in the next chapter.

Example 2.3.1 (Multiple Hypothesis Testing)**.** Suppose that $\Theta = \mathcal{A} = \{1, \ldots, N\}$ and $L(\theta, a) = \{\theta \neq a\}$. Then,

$$R_{\text{minimax}} \geq \bar{r}_\rho(P_1, \ldots, P_N) \qquad \text{for every } \rho_i \geq 0, \sum_{i=1}^{N} \rho_i = 1. \tag{2.2}$$

This is a direct consequence of (2.1). Indeed, for every $a \in \mathcal{A}$ and $x \in \mathcal{X}$, we can write

$$B^a_{\rho,L}(x) = \sum_{i=1}^{N} \{a \neq i\}\, p_i(x)\rho_i = \sum_{i=1}^{N} p_i(x)\rho_i - p_a(x)\rho_a$$

It follows therefore that $\inf_{a\in\mathcal{A}} B^a_{\rho,L}(x) = \sum_i p_i(x)\rho_i - \max_i[p_i(x)\rho_i]$ from which (2.2) immediately follows.

□

The bound (2.1), with a multiplicative factor, can be obtained for R_{minimax} even in general decision-theoretic problems, as explained in the following example.

Example 2.3.2. [General Testing Bound] For every η-separated finite subset F of Θ, we have

$$R_{\text{minimax}} \geq \frac{\eta}{2}\bar{r}_\rho(P_\theta, \theta \in F) \qquad \text{for all } \rho_\theta \geq 0, \theta \in F \text{ with } \sum_{\theta \in F} \rho_\theta = 1. \tag{2.3}$$

This can be proved from (2.1) by choosing w to be the discrete probability measure on F with $w\{\theta\} = \rho_\theta, \theta \in F$. Indeed, for this prior w, we use the inequality $L(\theta, a) \geq (\eta/2)\{L(\theta, a) \geq \eta/2\}$ to write

$$B^a_{w,L}(x) \geq \frac{\eta}{2}\left(\sum_{\theta \in F} \rho_\theta p_\theta(x) - \sum_{\theta \in F} \rho_\theta p_\theta(x)\{L(\theta, a) < \eta/2\}\right)$$

for every $a \in \mathcal{A}$ and $x \in \mathcal{X}$. Because F is η-separated, for every action a, the loss $L(\theta, a)$ is strictly smaller than $\eta/2$ for at most $\theta \in F$. It follows therefore that

$$B_{w,L}(x) \geq \frac{\eta}{2}\left(\sum_{\theta \in F} \rho_\theta p_\theta(x) - \max_{\theta \in F}[\rho_\theta p_\theta(x)]\right)$$

which implies (2.1).

□

Example 2.3.3 (Assouad)**.** Suppose that Θ and $\mathcal{A}$ denote the hypercube $\{0, 1\}^m$ with the loss function $L(\theta, a) = \Upsilon(\theta, a) = \sum_{i=1}^m \{\theta_i \neq a_i\}$. Then

$$R_{\text{minimax}} \geq \frac{m}{2} \min_{\Upsilon(\theta, \theta')=1} ||P_\theta \wedge P_{\theta'}||_1. \tag{2.4}$$

We shall prove this using (2.1) by taking w to be the uniform probability measure on Θ. For every $a \in \mathcal{A}$ and $x \in \mathcal{X}$,

$$B^a_{w,\Upsilon}(x) = 2^{-m} \sum_{i=1}^m \sum_{\theta \in \{0,1\}^m} \{\theta_i \neq a_i\} p_\theta(x)$$

and consequently

$$B_{w,\Upsilon}(x) = \frac{1}{2}\sum_{i=1}^m \min\left(\frac{\sum_{\theta:\theta_i=0} p_\theta(x)}{2^{m-1}}, \frac{\sum_{\theta:\theta_i=1} p_\theta(x)}{2^{m-1}}\right).$$

Thus by (2.1),

$$R_{\text{minimax}} \geq \frac{1}{2}\sum_{i=1}^{m}\left|\left|\left(2^{-(m-1)}\sum_{\theta:\theta_i=0} P_\theta\right) \wedge \left(2^{-(m-1)}\sum_{\theta:\theta_i=1} P_\theta\right)\right|\right|_1$$

Each of the terms in the above summation can be seen to be bounded from below by $\min_{\Upsilon(\theta,\theta')=1} ||P_\theta \wedge P_{\theta'}||_1$ which gives (2.4).

□

Assouad's method also applies to general problems as explained below.

Example 2.3.4 (General Assouad)**.** Consider a map $\psi : \{0,1\}^m \to \Theta$ and suppose that ζ is a positive real number such that $d(\psi(\tau), \psi(\tau')) \geq \zeta\Upsilon(\tau, \tau')$ for every pair $\tau, \tau' \in \{0,1\}^m$. Then

$$R_{\text{minimax}} \geq \frac{m\zeta}{4} \min_{\Upsilon(\tau,\tau')=1} ||P_{\psi(\tau)} \wedge P_{\psi(\tau')}||_1. \tag{2.5}$$

In order to prove this, for $a \in \mathcal{A}$, we define $\tau_a \in \{0,1\}^m$ by $\tau_a := \operatorname{argmin}_\tau L(\psi(\tau), a)$. Then

$$L(\psi(\tau), a) \geq \frac{L(\psi(\tau), a) + L(\psi(\tau_a), a)}{2} \geq \frac{\zeta}{2}\Upsilon(\tau, \tau_a).$$

Thus by choosing w to be the image of the uniform probability measure on $\{0,1\}^m$ under the map ψ, we get

$$B^a_{w,L}(x) \geq \frac{\zeta}{2}\frac{1}{2^m}\sum_{\tau\in\{0,1\}^m} \Upsilon(\tau, \tau_a) p_{\psi(\tau)}(x)$$

for every $x \in \mathcal{X}$ and $a \in \mathcal{A}$. From here, we proceed as in the previous example to obtain (2.5).

□

Example 2.3.5 (Le Cam)**.** Let w_1 and w_2 be two probability measures that are supported on subsets Θ_1 and Θ_2 of the parameter space respectively. Also let m_1 and m_2 denote the marginal densities of X with respect to w_1 and w_2 respectively i.e., $m_i(x) := \int_{\Theta_i} p_\theta(x) w_i(d\theta)$ for $i = 1, 2$. Le Cam (1973) proved the following inequality

$$R_{\text{minimax}} \geq \frac{1}{2} d\left(\Theta_1, \Theta_2\right) \left|\left|m_1 \wedge m_2\right|\right|_1 . \tag{2.6}$$

For its proof, we use (2.1) with the mixture prior $w = (w_1 + w_2)/2$. For every $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$\begin{aligned} B^a_{w,L}(x) &= \frac{1}{2} B^a_{w_1,L}(x) + \frac{1}{2} B^a_{w_2,L}(x) \\ &\geq \frac{1}{2} m_1(x) \inf_{\theta_1 \in \Theta_1} L(\theta_1, a) + \frac{1}{2} m_2(x) \inf_{\theta_2 \in \Theta_2} L(\theta_2, a) \\ &\geq \frac{1}{2} \min\left(m_1(x), m_2(x)\right) d\left(\Theta_1, \Theta_2\right), \end{aligned}$$

which, at once, implies (2.6).

□

Example 2.3.6 (Fano)**.** Fano's inequality states that for every finite η-separated subset of Θ with cardinality denoted by N, we have

$$R_{\text{minimax}} \geq \frac{\eta}{2} \left(1 - \frac{\log 2 + \frac{1}{N} \sum_{\theta \in F} D_1(P_\theta || \bar{P})}{\log N}\right), \tag{2.7}$$

where $\bar{P} := \sum_{\theta \in F} P_\theta / N$. The quantity $J_1 := \sum_{\theta \in F} D_1(P_\theta || \bar{P})/N$ is known as the Jensen-Shannon divergence. It is also Shannon's mutual information (Cover and Thomas, 2006, Page 19) between the random parameter θ distributed according to the uniform distribution on F and the observation X whose conditional distribution given θ equals P_θ.

The general testing bound: $R_{\text{minimax}} \geq (\eta/2)\bar{r}(P_\theta, \theta \in F)$ is the first step in the proof of (2.7). The next step is to prove that

$$\bar{r}(P_\theta, \theta \in F) \geq 1 - \frac{\log 2 + \frac{1}{N}\sum_{\theta \in F} D_1(P_\theta || \bar{P})}{\log N}. \tag{2.8}$$

Kemperman (1969, Page 135) provided a simple proof of (2.8) using the following elementary inequality: For nonnegative numbers $a_1, \dots, a_N$,

$$(\log N) \max_{1 \leq i \leq N} a_i \leq \sum_{i=1}^{N} a_i \log\left(\frac{2a_i}{\bar{a}}\right) \qquad \text{where } \bar{a} := (a_1 + \cdots + a_N)/N. \tag{2.9}$$

For a proof of (2.9), assume, without loss of generality, that $\sum_i a_i = 1$ and $a_1 = \max_{1 \leq i \leq N} a_i$. Then (2.9) is equivalent to the inequality $\sum_i a_i \log(b_i/a_i) \leq 0$ where $b_1 = 1/2$ and $b_i = 1/(2N), i = 2, \dots, N$ and this latter inequality is just a consequence of Jensen's inequality (using the convexity of $x \mapsto \log x$ and $\sum_i a_i = 1 \geq \sum_i b_i$).

Kemperman proved (2.8) by applying (2.9) to the nonnegative numbers $p_\theta(x), \theta \in F$ for a fixed $x \in \mathcal{X}$ and integrating both sides of the resulting inequality with respect to μ.

The inequality (2.7) has been extensively used in the nonparametric statistics literature for obtaining minimax lower bounds, important works being Ibragimov and Has'minskii (1977, 1980, 1981); Has'minskii (1978); Birgé (1983, 1986); Yang and Barron (1999).

□

Bibliography

Birgé, L. (1983). Approximation dans les espaces metriques et theorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65*, 181–237.

Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields 71*, 271–291.

Cai, T. T. and M. G. Low (2011). Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional. *Annals of Statistics 39*, 1012–1041.

Cover, T. and J. Thomas (2006). *Elements of Information Theory* (2 ed.). Wiley.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Boston: Academic Press.

Has'minskii, R. Z. (1978). A lower bound on the risk of nonparametric estimates of densities in the uniform metric. *Theory Probability and Its Applications 23*, 794–798.

Ibragimov, I. and R. Z. Has'minskii (1977). A problem of statistical estimation in Gaussian white noise. *Dokl. Akad. Nauk SSSR 236*, 1053–1055.

Ibragimov, I. and R. Z. Has'minskii (1980). On estimate of the density function. *Zap. Nauchn. Semin. LOMI 98*, 61–85.

Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. In *Probability and Information Theory*. Springer-Verlag. Lecture Notes in Mathematics, 89, pages 126–169.

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

Lehmann, E. L. and G. Casella (1998). *Theory of Point Estimation* (2nd ed.). New York: Springer.

Massart, P. (2007). *Concentration inequalities and model selection. Lecture notes in Mathematics*, Volume 1896. Berlin: Springer.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer-Verlag.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Yu, B. (1997). Assouad, Fano, and Le Cam. In D. Pollard, E. Torgersen, and G. L. Yang (Eds.), *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 423–435. New York: Springer-Verlag.

Chapter 3

Bounds via f-divergences

3.1 f-divergences: What are they?

In this chapter, we shall prove minimax lower bounds using f-divergences.

Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. The limits $f(0) := \lim_{x \downarrow 0} f(x)$ and $f'(\infty) := \lim_{x \uparrow \infty} f(x)/x$ exist by convexity although they can be $+\infty$.

For two probability measures P and Q having densities p and q with respect to μ, Ali and Silvey (1966) defined the f-divergence $D_f(P||Q)$ between P and Q by

$$D_f(P||Q) := Qf(p/q) + f'(\infty)P\{q = 0\}. \tag{3.1}$$

This notion was also independently introduced by Csisźar (1963).

$D_f(P||Q)$ can be viewed as a measure of distance between P and Q. It is usually not a metric however with the exception of the total variation distance $||P - Q||_{TV}$, which corresponds to $f(x) = |x-1|/2$ (an interesting fact, whose proof can be found in Vajda, 2009, is that D_f is a metric if and only if it equals, up to a constant, the total variation distance).

When P is absolutely continuous with respect to Q, the second term in the right hand side of (3.1) equals zero (note the convention $\infty \times 0 = 0$) and thus the definition reduces to $Qf(p/q)$. When $f(x) = |x-1|/2$, the second term is necessary in order to ensure that (3.1) agrees with the usual definition for total variation distance in the case when Q does not dominate P. For convex functions f with $f'(\infty) = \infty$ (such as $x \log x$ or $x^2 - 1$), $D_f(P||Q)$ equals $+\infty$ when P is not absolutely continuous with respect to Q.

It is easily checked that the right hand side above is unchanged if $f(x)$ is replaced by $f(x) + c(x-1)$ for any constant c. With an appropriate choice of c, we can always arrange for f to be minimized at $x = 1$ which, because $f(1) = 0$, ensures that f is nonnegative.

$D_f(P||Q)$ is convex in each argument; convexity in P is obvious while convexity in Q follows from $D_f(P||Q) = D_{f^*}(Q||P)$ for $f^*(x) = xf(1/x)$.

The power divergences constitute an important subfamily of the f-divergences. They correspond to the convex functions $f_\alpha, \alpha \in \mathbb{R}$ defined by

$$f_\alpha(x) = \begin{cases} x^\alpha - 1 & \text{for } \alpha \notin [0,1] \\ 1 - x^\alpha & \text{for } \alpha \in (0,1) \\ x \log x & \text{for } \alpha = 1 \\ -\log x & \text{for } \alpha = 0 \end{cases}$$

For simplicity, we shall denote the divergence D_{f_α} by D_α. One has the identity $D_\alpha(P||Q) = D_{1-\alpha}(Q||P)$. Some examples of power divergences are:

1. Kullback-Leibler divergence: $\alpha = 1$; $D_1(P||Q) = \int p \log(p/q) d\mu$.

2. Chi-squared divergence: $\alpha = 2$; $D_2(P||Q) = \int (p^2/q) d\mu$.

3. Square of the Hellinger distance: $\alpha = 1/2$; $D_{1/2}(P||Q) = 1 - \int \sqrt{pq} d\mu$.

The total variation distance $||P - Q||_{TV}$ is an f-divergence (with $f(x) = |x-1|/2$) but not a power divergence.

The power divergences are particularly handy in applications where the underlying probabilities are product measures, for which the calculation of power divergences reduces to calculations on the marginal distributions. Indeed, it can be readily checked that

$$D_\alpha\left(P_1 \times \cdots \times P_n || Q_1 \times \cdots \times Q_n\right) = \begin{cases} \prod_{i=1}^n \left(D_\alpha(P_i||Q_i) + 1\right) - 1 & \text{for } \alpha \notin [0,1] \\ 1 - \prod_{i=1}^n \left(1 - D_\alpha(P_i||Q_i)\right) & \text{for } \alpha \in (0,1) \\ \sum_{i=1}^n D_\alpha(P_i||Q_i) & \text{for } \alpha \in \{0,1\} \end{cases}$$

3.2 Main Result

Consider the quantity $\bar{r} = \bar{r}(P_1, \dots, P_N) = 1 - \frac{1}{N}\int \max_i p_i d\mu$ for probability measures $P_1, \dots, P_N$ having densities $p_1, \dots, p_N$ with respect to μ. As explained in Chapter 2, the quantity $\bar{r}(P_1, \dots, P_N)$ appears in almost all the standard minimax lower bound techniques. For example, the general testing bound uses $\bar{r}(P_1, \dots, P_N)$ directly and the methods of Assouad and Le Cam use the affinity term $||P_1 \wedge P_2|| = 2\bar{r}(P_1, P_2)$.

The following theorem provides a lower bound for $\bar{r}$ in terms of f-divergences. As we shall demonstrate in the rest of this chapter, it implies a number of very useful lower bounds for the minimax risk in general decision-theoretic problems.

Theorem 3.2.1. *Consider probability measures $P_1, \dots, P_N$ on a space $\mathcal{X}$ and a convex function f on $(0, \infty)$ with $f(1) = 0$. For every probability measure Q on $\mathcal{X}$, we have $\sum_{i=1}^N D_f(P_i||Q) \geq g(\bar{r})$ where $\bar{r} = \bar{r}(P_1, \dots, P_N)$ and*

$$g(a) := f\left(N(1-a)\right) + (N-1) f\left(\frac{Na}{N-1}\right). \tag{3.2}$$

Proof. To make the basic idea clearer, we assume that $P_1, \dots, P_N$ are all dominated by Q and write p_i for the density of P_i with respect to Q. For the undominated case, see the proof of Theorem 3.3.1. We start with the following simple inequality for nonnegative numbers $a_1, \dots, a_N$

$$\sum_{i=1}^{N} f(a_i) \geq f(\max_i a_i) + (N-1) f\left(\frac{\sum_{i=1}^{N} a_i - \max_i a_i}{N-1}\right). \tag{3.3}$$

To see this, assume without loss of generality that $a_1 = \max_i a_i$, rewrite the sum $\sum_i f(a_i)$ as $f(a_1) + (N-1)\sum_{i\geq 2}(f(a_i)/(N-1))$ and use convexity on the final sum.

We now fix $x \in \mathcal{X}$ and apply (3.3) with $a_i := p_i(x)$ to obtain

$$\sum_{i=1}^{N} f\left(p_i(x)\right) \geq f(\max_i p_i(x)) + (N-1) f\left(\frac{\sum_{i=1}^{N} p_i(x) - \max_i p_i(x)}{(N-1)}\right).$$

The required inequality $\sum_i D_f(P_i||Q) \geq g(\bar{r})$ is now deduced by integrating both sides of the above pointwise inequality with respect to Q and using Jensen's inequality on the right hand side. □

Remark 3.2.1. *As already mentioned in Remark 1.0.1, after the journal acceptance of the paper Guntuboyina (2011), on which the present chapter is based, Professor Alexander Gushchin brought to my notice the fact that the above theorem appears in Gushchin (2003). The extent of the overlap with Gushchin's paper and the differences between our's and Gushchin's proof of the theorem are described in Section 3.4 of Chapter 3.*

Remark 3.2.2. *The special case of Theorem 3.2.1 for the Kullback-Leibler divergence ($f(x) = x \log x$) has appeared in the literature previously: implicitly in Han and Verdú (1994, Proof of Theorem 1) and explicitly, without proof, in Birgé (2005, Theorem 3). The proof in Han and Verdú (1994) is based on information-theoretic arguments.*

The following argument shows that Theorem 3.2.1 provides a lower bound for $\bar{r}$. Note that $\bar{r}$ is at most $1-1/N$ (which directly follows from the definition of $\bar{r}$) and g is non-increasing on $[0, 1-1/N]$. To see this, observe that for every $a \in (0, 1-1/N]$, we have

$$\frac{g'_L(a)}{N} = f'_L\left(\frac{Na}{N-1}\right) - f'_R(N(1-a)),$$

where g'_L and f'_L represent left derivatives and f'_R represents right derivative (note that f'_L and f'_R exist because of the convexity of f). Because $Na/(N-1) \le N(1-a)$ for every $a \in [0, 1-1/N]$ and f is convex, we see that $g'_L(a) \le 0$ for every $a \in (0, 1-1/N]$ which implies that g is non-increasing on $[0, 1-1/N]$.

We also note that the convexity of f implies that g is convex as well. The following techniques are useful for converting the inequality given in Theorem 3.2.1 into an explicit lower bound for $\bar{r}$:

1. **Explicit inversion of** g: For certain functions f, the function g given by (3.2) can be explicitly inverted. Examples are given below.

   (a) (Chi-squared divergence) For $f(x) = f_2(x) = x^2 - 1$,

   $$g(\bar{r}) = \frac{N^3}{N-1}\left(1 - \frac{1}{N} - \bar{r}\right)^2 \ge N^2\left(1 - \frac{1}{N} - \bar{r}\right)^2.$$

   Because $\bar{r} \le 1 - 1/N$, the inequality $\sum_i D_2(P_i||Q) \ge g(\bar{r})$ can be inverted to yield

   $$\bar{r}(P_1, \ldots, P_N) \ge 1 - \frac{1}{N} - \frac{1}{\sqrt{N}}\sqrt{\frac{\sum_{i=1}^N D_2(P_i||Q)}{N}} \qquad \text{for every } Q. \quad (3.4)$$

   (b) (Total variation distance) For $f(x) = |x-1|/2$, because $\bar{r} \le 1 - 1/N$, it can be checked that $g(\bar{r}) = N - 1 - N\bar{r}$. We, thus, have the explicit inequality:

$$\bar{r} \geq 1 - \frac{1}{N} - \frac{\sum_{i=1}^{N} ||P_i - Q||_{TV}}{N} \qquad \text{for every } Q.$$

2. **Lower bounds for** g: Lower bounds for g can often lead to useful inequalities. For example, if $f(x) = f_\alpha(x) = x^\alpha - 1$ with $\alpha > 1$, then the function g has the simple lower bound:

   $$g(\bar{r}) = N^\alpha(1-\bar{r})^\alpha - N + (N-1)\left(\frac{N\bar{r}}{N-1}\right)^\alpha \geq N^\alpha(1-\bar{r})^\alpha - N.$$

   This results in the following explicit bound for $\bar{r}$:

   $$\bar{r} \geq 1 - \left(\frac{1}{N^{\alpha-1}} + \frac{\sum_{i=1}^{N} D_\alpha(P_1||Q)}{N^\alpha}\right)^{1/\alpha} \qquad \text{for every } Q \text{ and } \alpha > 1. \tag{3.5}$$

   When $\alpha = 2$, the above inequality is weaker than (3.4) but for large N, the two bounds are almost the same.

3. **Linear approximation for** g: We have a seemingly crude method that works for every f. Because the function g is convex and non-increasing, for every a in $(0, 1-1/N]$, the left derivative $g'_L(a)$ is less than or equal to 0 and $g(\bar{r})$ is at least $g(a) + g'_L(a)(\bar{r} - a)$. Theorem 3.2.1 implies therefore that $\sum_i D_f(P_i||Q)$ is at least $g(a) + g'_L(a)(\bar{r} - a)$ which, when rearranged, results in

   $$\bar{r} \geq a + \frac{\sum_{i=1}^{N} D_f(P_i||Q) - g(a)}{g'_L(a)} \tag{3.6}$$

   for every Q and $a \in (0, 1-1/N]$ with $g'_L(a) < 0$. As explained in Section 3.6, this crude result is strong enough for Theorem 3.2.1 to yield Fano's inequality.

3.3 A more general result

In this section, we show that the method of proof used for Theorem 3.2.1 also gives an inequality for $\bar{r}_w = \bar{r}_w(P_1, \dots, P_N) = 1 - \int \max_i(w_i p_i) d\mu$ for general weights $w_i \geq 0$ with $\sum_i w_i = 1$. This result has been included here just for completeness and will not be used in the sequel. Theorem 3.2.1 is a special case of the following theorem obtained by taking $w_i = 1/N$. Moreover, in the proof of the following theorem, we do not necessarily assume that $P_1, \dots, P_N$ are dominated by Q.

Theorem 3.3.1. *For every $f : (0, \infty) \to \mathbb{R}$ and every probability measure Q,*

$$\sum_{i=1}^{N} w_i D_f(P_i || Q) \geq W f\left(\frac{1 - \bar{r}_w}{W}\right) + (1 - W) f\left(\frac{\bar{r}_w}{1 - W}\right), \tag{3.7}$$

where $W := \int_{\mathcal{X}} w_{T(x)} Q(dx)$ with $T(x) := \operatorname{argmax}_{1 \leq i \leq N}(w_i p_i(x))$.

Proof. We assume, without loss of generality, that all the weights $w_1, \dots, w_N$ are strictly positive. Suppose $dP_i/d\mu = p_i, i = 1, \dots, N$ and $dQ/d\mu = q$. Consider the following pointwise inequality: For nonnegative numbers $a_1, \dots, a_N$ and every $1 \leq \tau \leq N$,

$$\sum_{i=1}^{N} w_i f(a_i) \geq w_\tau f(a_\tau) + (1 - w_\tau) f\left(\frac{\sum_{i=1}^{N} w_i a_i - w_\tau a_\tau}{1 - w_\tau}\right).$$

Applying this inequality to $a_i = p_i(x)/q(x)$ and $\tau := T(x) = \operatorname{argmax}_i(w_i p_i(x))$ for a fixed x with $q(x) > 0$, we obtain

$$\sum_{i=1}^{N} w_i f\left(\frac{p_i(x)}{q(x)}\right) \geq w_{T(x)} f\left(\frac{p_{T(x)}(x)}{q(x)}\right) + (1 - w_{T(x)}) f\left(\frac{\sum_i w_i p_i(x) - w_{T(x)} p_{T(x)}(x)}{(1 - w_{T(x)}) q(x)}\right).$$

Integrating both sides with respect to Q, we obtain that $\sum_i w_i Q f(p_i/q)$ is greater than or equal to

$$W \int f\left(\frac{p_{T(x)}(x)}{q(x)}\right) Q'(dx) + (1-W)\int f\left(\frac{\sum_i w_i p_i(x) - w_{T(x)} p_{T(x)}}{(1-w_{T(x)})q(x)}\right) Q''(dx), \quad (3.8)$$

where $W = \int w_{T(x)} Q(dx)$ and

$$Q'(dx) := \frac{w_{T(x)}}{W} Q(dx) \qquad \text{and} \qquad Q''(dx) := \frac{1-w_{T(x)}}{1-W} Q(dx).$$

By the application of Jensen's inequality to each of the terms in (3.8), we deduce that $\sum_i w_i Q f(p_i/q)$ is greater than or equal to

$$Wf\left(\int_{\{q>0\}} \frac{\max_i(w_i p_i)}{W} d\mu\right) + (1-W) f\left(\int_{\{q>0\}} \frac{\sum_i w_i p_i - \max_i(w_i p_i)}{1-W} d\mu\right).$$

Also note that $\sum_i w_i P_i\{q = 0\}$ equals

$$W \int_{\{q=0\}} \frac{\max_i(w_i p_i)}{W} d\mu + (1-W) \int_{\{q=0\}} \frac{\sum_i w_i p_i - \max_i(w_i p_i)}{1-W} d\mu.$$

By the definition of $D_f(P_i||Q)$, we deduce that $\sum_i w_i D_f(P_i||Q)$ is bounded from below by $WT_1 + (1-W)T_2$ where T_1 and T_2 equal

$$f\left(\int_{\{q>0\}} \frac{\max_i(w_i p_i)}{W} d\mu\right) + f'(\infty) \int_{\{q=0\}} \frac{\max_i(w_i p_i)}{W} d\mu$$

and

$$f\left(\int_{\{q>0\}} \frac{\sum_i w_i p_i - \max_i(w_i p_i)}{1-W} d\mu\right) + f'(\infty) \int_{\{q=0\}} \frac{\sum_i w_i p_i - \max_i(w_i p_i)}{1-W} d\mu$$

respectively. Now by the convexity of f, the inequality $f(y_0) + (y - y_0) f'(\infty) \geq f(y)$ holds for every $0 \leq y_0 \leq y$. Using this with $y_0 := \int_{\{q>0\}} \max_i(w_i p_i) d\mu / W$ and

$y = \int \max_i (w_i p_i) d\mu / W$, we obtain that $T_1 \geq f((1 - \bar{r}_w)/W)$. It is similarly shown that $T_2 \geq f(\bar{r}_w/(1-W)$ which implies that $WT_1 + (1-W)T_2$ is larger than or equal to the right hand side of (3.7). □

3.4 Overlap with Gushchin (2003)

As mentioned in Remark 3.2.1, Professor Alexander Gushchin pointed out to me (after the acceptance of Guntuboyina, 2011) that Theorem 3.2.1 and its non-uniform prior version, Theorem 3.3.1, appear in his paper Gushchin (2003). Specifically, in a different notation, Theorem 3.2.1 appears as Theorem 1 and inequality (3.7) appears in Section 4.3 in Gushchin (2003). Except for these two theorems and the observation that Fano's inequality is a special case of Theorem 3.2.1 (which we make in Section 3.6), there is no other overlap between this thesis and Gushchin (2003).

Also, the proof of Theorem 3.2.1 (and Theorem 3.3.1) given in Gushchin (2003) is different from our proof. In order to make this transparent, we shall sketch Gushchin's proof of Theorem 3.2.1 here:

1. The proof starts with the observation that $\sum_i D_f(P_i||Q)/N$ equals $D_f(\tilde{P}||\tilde{Q})$ where $\tilde{P}$ and $\tilde{Q}$ denote probability measures on $\mathcal{X} \times \{1, \dots, N\}$ defined by $\tilde{P}(B \times \{i\}) = P_i(B)/N$ and $\tilde{Q}(B \times \{i\}) = Q(B)/N$ for $B \subseteq \mathcal{X}$.

2. Let $A_1, \dots, A_N$ denote a partition of $\mathcal{X}$ such that $p_i(x)$ equals $\max_i p_i(x)$ for $x \in A_i$. Consider the test function ϕ on $\mathcal{X} \times \{1, \dots, N\}$ defined by $\phi(x, i) = \{x \notin A_i\}$. It can be checked that $\tilde{P}\phi = \bar{r}$ and $\tilde{Q}(1 - \phi) = 1/N$.

3. Gushchin (2003) then invokes a general result (Liese and Vajda, 1987, Theorem 1.24) relating f-divergences to the type I and type II errors of tests to deduce Theorem (3.2.1).

Our proof, which is based on the elementary pointwise inequality (3.3) and two applications of Jensen's inequality, is clearly simpler.

3.5 Special Case: $N = 2$

For $N = 2$, Theorem 3.2.1 gives

$$D_f(P_1||Q) + D_f(P_2||Q) \geq f(2(1 - \bar{r})) + f(2\bar{r}).$$

The quantity $\bar{r}(P_1, P_2)$ is related to the total variation distance V between P_1 and P_2 via $V = 1 - 2\bar{r}(P_1, P_2)$. Thus the above inequality can be rewritten in terms of total variation distance as follows:

$$D_f(P_1||Q) + D_f(P_2||Q) \geq f(1 + V) + f(1 - V) \qquad \text{for every } Q. \tag{3.9}$$

We have singled out this special case of Theorem 3.2.1 because

1. It adds to the many inequalities that exist in the literature which relate the f-divergence between two probability measures to their total variation distance.
2. As may be recalled from the previous chapter, lower bounds for $\bar{r}(P_1, P_2)$, which also equals one-half the affinity $||P_1 \wedge P_2||_1$, for two probability measures P_1 and P_2 are necessary for the application of the bounds of Assouad and Le Cam.

Inequality (3.9) is new although its special case for $f(x) = x \log x$ has been obtained by Topsøe (2000, Equation (24)). Topsøe (2000) also explained how to use this inequality to deduce Pinsker's inequality with sharp constant: $D_1(P_1||P_2) \geq 2V^2$.

3.6 Fano's inequality

Fano's inequality, which is commonly used in nonparametric statistics, bounds $\bar{r}$ from below using the Kullback-Leibler divergence between the P_i's and their average, $\bar{P} := (P_1 + \cdots + P_N)/N$:

$$\bar{r} \geq 1 - \frac{\log 2 + \frac{1}{N}\sum_{i=1}^{N} D_1(P_i||\bar{P})}{\log N}. \tag{3.10}$$

It is a consequence of (3.6) for $f(x) = x \log x$, $Q = \bar{P}$ and $a = (N-1)/(2N-1)$. Indeed, with these choices, (3.6) gives

$$\bar{r} \geq 1 - \frac{\log((2N-1)/N) + \frac{1}{N}\sum_{i=1}^{N} D_1(P_i||\bar{P})}{\log N}.$$

which clearly implies (3.10) because $\log((2N-1)/N) \leq \log 2$. It may be helpful to note here that for the Kullback-Leibler divergence D_1, the probability measure Q which minimizes $\sum_i D_1(P_i||Q)$ equals $\bar{P}$ and this follows from the following well-known identity (sometimes referred to as the *compensation identity*, see for example Topsøe, 2000, Page 1603):

$$\sum_{i=1}^{N} D_1(P_i||Q) = \sum_{i=1}^{N} D_1(P_i||\bar{P}) + N D_1(\bar{P}||Q) \qquad \text{for every } Q.$$

Remark 3.6.1. Our proof of Theorem 3.2.1 is similar in spirit to Kemperman's proof of Fano's inequality described in the last chapter (see Example 2.3.6). The starting point in both proofs is a pointwise inequality involving the maximum of a finite number of nonnegative numbers. Kemperman's proof starts with the pointwise inequality:

$$m \log N \leq \sum_{i=1}^{N} a_i \log\left(\frac{2a_i}{\bar{a}}\right) \qquad \text{for } a_i \geq 0 \text{ with } m := \max_{1\leq i \leq N} a_i. \tag{3.11}$$

By homogeneity, we may assume that $\sum_i a_i = 1$. The inequality is then equivalent to

$$\sum_i a_i \log a_i \geq -\log 2 - (1-m)\log N. \tag{3.12}$$

Our proof of Theorem 3.2.1 starts with (3.3) which, for $f(x) = x\log x$ and $\sum_i a_i = 1$ becomes

$$\sum_{i=1}^{N} a_i \log a_i \geq m \log m + (1-m)\log(1-m) - (1-m)\log(N-1). \tag{3.13}$$

This inequality is stronger than Kemperman's inequality (3.12) because of the elementary inequality: $m\log m + (1-m)\log(1-m) \geq -\log 2$ for all $m \in [0,1]$.

3.7 Upper bounds for $\inf_Q \sum_i D_f(P_i||Q)$

For successful application of Theorem 3.2.1, one needs useful upper bounds for the quantity $J_f := \inf_Q \sum_{i=1}^{N} D_f(P_i||Q)/N$. When $f = f_\alpha$, we write J_α for J_f. Such bounds are provided in this section.

For $f(x) = x\log x$, the following inequality has been frequently used in the literature (see, for example, Birgé, 1983 and Nemirovski, 2000):

$$J_1 \leq \frac{1}{N}\sum_{i=1}^{N} D_1(P_i||\bar{P}) \leq \frac{1}{N^2}\sum_{i,j} D_1(P_i||P_j) \leq \max_{i,j} D_1(P_i||P_j).$$

This is just a consequence of the convexity of $D_1(P||Q)$ in Q and, for the same reason, holds for all f-divergences. The inequality is analogous to using $\max(a_i - a_j)^2$ as an upper bound for $\inf_c \sum_{i=1}^N (a_i - c)^2/N$ and, quite often, $\max_{i,j} D_f(P_i||P_j)$ is not a good upper bound for J_f.

Yang and Barron (1999, Page 1571) improved the upper bound in the case of the Kullback-Leibler divergence. Specifically, they showed that for every set of probability measures $Q_1, \dots, Q_M$,

$$\inf_Q \frac{1}{N} \sum_{i=1}^N D_1(P_i||Q) \le \log M + \max_{1\le i\le N} \min_{1\le j\le M} D_1(P_i||Q_j). \tag{3.14}$$

The M probability measures $Q_1, \dots, Q_M$ can be viewed as an approximation of the N probability measures $P_1, \dots, P_N$. The term $\max_i \min_j D_1(P_i||Q_j)$ then denotes the approximation error, measured via the Kullback-Leibler divergence. The right hand side of inequality (3.14) can therefore be made small if it is possible to choose not too many probability measures $Q_1, \dots, Q_M$ which well approximate the given set of probability measures $P_1, \dots, P_N$.

Inequality (3.14) can be rewritten using covering numbers. For $\epsilon > 0$, let $M_1(\epsilon)$ denote the smallest number M for which there exist probability measures $Q_1, \dots, Q_M$ that form an ϵ^2-*cover* for $P_1, \dots, P_N$ in the Kullback-Leibler divergence i.e.,

$$\min_{1\le j\le M} D_1(P_i||Q_j) \le \epsilon^2 \qquad \text{for every } 1 \le i \le N.$$

Then (3.14) is equivalent to

$$\inf_Q \frac{1}{N} \sum_{i=1}^N D_1(P_i||Q) \le \inf_{\epsilon>0} \left(\log M_1(\epsilon) + \epsilon^2\right). \tag{3.15}$$

Note that $\log M_1(\epsilon)$ is a decreasing function of ϵ. The right hand side of the above inequality involves the usual increasing versus decreasing trade-off. The next Theorem generalizes the bound (3.14) to arbitrary f-divergences.

Theorem 3.7.1. *Let $Q_1, \dots, Q_M$ be probability measures having densities $q_1, \dots, q_M$ respectively with respect to μ. Let us denote their average by $\bar{Q} = (Q_1 + \dots + Q_M)/M$ with $\bar{q} = (q_1 + \dots + q_M)/M$. Then for every convex function f on $(0, \infty)$ with $f(1) = 0$, we have*

$$J_f \leq \frac{1}{N} \sum_{i=1}^{N} \min_{1 \leq j \leq M} \int_{\mathcal{X}} \frac{q_j}{M} f\left(\frac{Mp_i}{q_j}\right) d\mu + \left(1 - \frac{1}{M}\right) f(0) + f'(\infty) \bar{P} \left\{\bar{q} = 0\right\}. \quad (3.16)$$

Proof. We assume, without loss of generality, that $f(0) < \infty$. Clearly for each $i \in \{1, \dots, N\}$,

$$D_f(P_i || \bar{Q}) = \int_{\mathcal{X}} \bar{q} \left[f\left(\frac{p_\theta}{\bar{q}}\right) - f(0)\right] + f(0) + f'(\infty) P_i \left\{\bar{q} = 0\right\}.$$

The convexity of f implies that the map $y \mapsto y[f(a/y) - f(0)]$ is non-increasing for every nonnegative a. Using this and the fact that $\bar{q} \geq q_j/M$ for every j, we get that for every $i \in \{1, \dots, N\}$,

$$D_f(P_i || \bar{Q}) \leq \min_{1 \leq j \leq M} \int_{\mathcal{X}} \frac{q_j}{M} \left[f\left(\frac{Mp_i}{q_j}\right) - f(0)\right] d\mu + f(0) + f'(\infty) P_i \left\{\bar{q} = 0\right\}.$$

Inequality (3.16) is deduced by averaging these inequalities over $1 \leq i \leq N$. □

For $f(x) = x \log x$, the inequality (3.16) gives

$$J_1 \leq \log M + \frac{1}{N} \sum_{i=1}^{N} \min_j D_1(P_i || Q_j) + \infty \cdot \bar{P} \left\{\bar{q} = 0\right\}.$$

This clearly implies (3.14) (note that the $\infty \cdot \bar{P}\{\bar{q} = 0\}$ term is redundant because if $\bar{P}$ is not absolutely continuous with respect to $\bar{Q}$, then $\min_j D_1(P_i || Q_j)$ would be $+\infty$ for some i).

Power divergences ($f(x) = f_\alpha(x), \alpha > 0$) are considered in the examples below.

Theorem 3.7.1 gives meaningful conclusions for power divergences only when $\alpha > 0$ because $f_\alpha(0)$ equals $+\infty$ when $\alpha \leq 0$.

Analogous to $M_1(\epsilon)$, let us define $M_\alpha(\epsilon)$ as the smallest number of probability measures needed to form an ϵ^2-cover of $P_1, \dots, P_N$ in the D_α divergence.

Example 3.7.2. Let $f(x) = x^\alpha - 1$ with $\alpha > 1$. Applying inequality (3.16), we get that

$$J_\alpha \leq M^{\alpha-1} \left(\frac{1}{N} \sum_{i=1}^{N} \min_{1 \leq j \leq M} D_\alpha(P_i || Q_j) + 1 \right) - 1 + \infty \cdot \bar{P} \left\{ \bar{q} = 0 \right\}.$$

As a consequence, we obtain (note that the $\infty \cdot \bar{P}\{\bar{q} = 0\}$ term is again redundant)

$$J_\alpha \leq M^{\alpha-1} \left(\max_{1 \leq i \leq N} \min_{1 \leq j \leq M} D_\alpha(P_i || Q_j) + 1 \right) - 1. \tag{3.17}$$

Rewriting in terms of the cover numbers $M_\alpha(\epsilon)$, we get

$$J_\alpha \leq \inf_{\epsilon > 0} \left(1 + \epsilon^2 \right) M_\alpha(\epsilon)^{\alpha-1} - 1. \tag{3.18}$$

Note that $M_\alpha(\epsilon)$ is a decreasing function of ϵ.

□

Example 3.7.3. Let $f(x) = 1 - x^\alpha$ for $0 < \alpha < 1$. The inequality (3.16) gives (note that $f'_\alpha(\infty) = 0$)

$$J_\alpha \leq 1 - \frac{1}{M^{1-\alpha}} \left(1 - \frac{1}{N} \sum_{i=1}^{N} \min_{1 \leq j \leq M} D_\alpha(P_i || Q_j) \right).$$

and thus

$$J_\alpha \leq 1 - \frac{1}{M^{1-\alpha}} \left(1 - \max_{1 \leq i \leq N} \min_{1 \leq j \leq M} D_\alpha(P_i, Q_j) \right).$$

In terms of $M_\alpha(\epsilon)$, we have

$$J_\alpha \leq 1 - \sup_{\epsilon>0}(1-\epsilon^2)M_\alpha(\epsilon)^{\alpha-1}.$$

Once again, the usual increasing versus decreasing trade-off is involved.

□

3.8 General Bounds

By combining Theorem 3.2.1: $J_f = \inf_Q \sum_{i=1}^N D_f(P_i||Q)/N \geq g(\bar{r})/N$ with the upper bound for J_f given in Theorem 3.7.1, we get lower bounds for $\bar{r}$ in terms of covering numbers of $\{P_1,\ldots,P_N\}$ measured in terms of the divergence D_f. For example, in the case of the convex function $f_\alpha(x) = x^\alpha - 1, \alpha > 1$ for which the inequality given by Theorem 3.2.1 can be approximately inverted to yield (3.5), combining (3.18) with (3.5) results in

$$\bar{r}(P_1,\ldots,P_N) \geq 1 - \left(\frac{1}{N^{\alpha-1}} + \frac{(1+\epsilon^2)M_\alpha(\epsilon)^{\alpha-1}}{N^{\alpha-1}}\right)^{1/\alpha} \qquad \text{for every } \epsilon > 0 \text{ and } \alpha > 1.$$

When $\alpha = 2$, we can use (3.4) instead of (3.5) to get

$$\bar{r}(P_1,\ldots,P_N) \geq 1 - \frac{1}{N} - \sqrt{\frac{(1+\epsilon^2)M_2(\epsilon)}{N}} \qquad \text{for every } \epsilon > 0.$$

One more special case is when $\alpha = 1$ (Kullback-Leibler divergence). Here we combine (3.10) with (3.15) to deduce

$$\bar{r}(P_1,\ldots,P_N) \geq 1 - \frac{\log 2 + \log M_1(\epsilon) + \epsilon^2}{\log N} \qquad \text{for every } \epsilon > 0.$$

If we employ the general testing bound (Chapter 2), then the above inequalities can be converted to produce inequalities for the minimax risk in general decision-theoretic problems. The general testing bound asserts that

$$R_{\text{minimax}} \geq (\eta/2)\bar{r}(P_\theta, \theta \in F) \qquad \text{for every } \eta\text{-separated finite subset } F \text{ of } \Theta.$$

Let us recall that a finite subset F of Θ is η-separated if $L(\theta_1, a) + L(\theta_2, a) \geq \eta$ for every $a \in \mathcal{A}$ and $\theta_1, \theta_2 \in F$ with $\theta_1 \neq \theta_2$.

The testing lower bound, therefore, implies that for every $\eta > 0$ and every finite η-separated subset F of Θ, the right hand side of each of the above three inequalities multiplied by $\eta/2$ would be a lower bound for R_{minimax}. This leads to the following three inequalities (the first inequality holds for every $\alpha > 1$)

$$R_{\text{minimax}} \geq \frac{\eta}{2}\left(1 - \left(\frac{1}{N^{\alpha-1}} + \frac{(1+\epsilon^2)M_\alpha(\epsilon; F)^{\alpha-1}}{N^{\alpha-1}}\right)^{1/\alpha}\right) \tag{3.19}$$

$$R_{\text{minimax}} \geq \frac{\eta}{2}\left(1 - \frac{1}{N} - \sqrt{\frac{(1+\epsilon^2)M_2(\epsilon; F)}{N}}\right), \tag{3.20}$$

$$R_{\text{minimax}} \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + \log M_1(\epsilon; F) + \epsilon^2}{\log N}\right), \tag{3.21}$$

where N is the cardinality of F and we have written $M_\alpha(\epsilon; F)$ in place of $M_\alpha(\epsilon)$ to stress that the covering number corresponds to $P_\theta, \theta \in F$. Inequality (3.21) is essentially due to Yang and Barron (1999) although they state their result for the estimation problem from n independent and identically distributed observations.

The first step in the application of these inequalities to a specific problem is the choice of η and the η-separated finite subset $F \subseteq \Theta$. This is usually quite involved and problem-specific. For example, refer to Chapter 4, where an application of these inequalities to a covariance matrix estimation problem is provided.

Yang and Barron (1999) suggested a clever way of applying (3.21) which does not require explicit construction of an η-separated subset F. Their first suggestion is to take F to be a *maximal* (as opposed to arbitrary) η-separated subset of Θ. Here maximal means that F is η-separated and no $F' \supseteq F$ is η-separated. For this F, they recommend the trivial bound $M_1(\epsilon; F) \leq M_1(\epsilon; \Theta)$. Here, the quantity $M_1(\epsilon; \Theta)$, or more generally, $M_\alpha(\epsilon; \Theta)$ is the covering number: smallest M for which there exist probability measures $Q_1, \dots, Q_M$ such that

$$\min_{1 \leq j \leq M} D_\alpha(P_\theta || Q_j) \leq \epsilon^2 \qquad \text{for every } \theta \in \Theta.$$

These ideas lead to the following lower bound:

$$R_{\text{minimax}} \geq \sup_{\eta > 0, \epsilon > 0} \frac{\eta}{2} \left(1 - \frac{\log 2 + \log M_1(\epsilon; \Theta) + \epsilon^2}{\log N(\eta)} \right) \tag{3.22}$$

where $N(\eta)$ denotes the size of a maximal η-separated subset of Θ.

Exactly parallel treatment of (3.19) and (3.21) leads to the following two bounds:

$$R_{\text{minimax}} \geq \sup_{\eta > 0, \epsilon > 0, \alpha > 1} \frac{\eta}{2} \left(1 - \left(\frac{1}{N(\eta)^{\alpha - 1}} + \frac{(1 + \epsilon^2) M_\alpha(\epsilon; \Theta)^{\alpha - 1}}{N(\eta)^{\alpha - 1}} \right)^{1/\alpha} \right) \tag{3.23}$$

and

$$R_{\text{minimax}} \geq \sup_{\eta > 0, \epsilon > 0} \frac{\eta}{2} \left(1 - \frac{1}{N(\eta)} - \sqrt{\frac{(1 + \epsilon^2) M_2(\epsilon; \Theta)}{N(\eta)}} \right). \tag{3.24}$$

The application of these inequalities just requires a lower bound on $N(\eta)$ and an upper bound on $M_\alpha(\epsilon; \Theta)$. Unlike the previous inequalities, these bounds do not involve an explicit η-separated subset of the parameter space.

The quantity $N(\eta)$ only depends on the structure of the parameter space Θ with respect to the loss function. It has no relation to the observational distributions

$P_\theta, \theta \in \Theta$. On the other hand, the quantity $M_\alpha(\epsilon; \Theta)$ depends only on these probability distributions and has no connection to the loss function. Both these quantities capture the global structure of the problem and thus, each of the above three inequalities can be termed as a *global minimax lower bound*.

Yang and Barron (1999) successfully applied inequality (3.22) to obtain optimal rate minimax lower bounds for standard nonparametric density estimation and regression problems where $N(\eta)$ and $M_1(\epsilon; \Theta)$ can be deduced from available results in approximation theory (for the performance of (3.22) on parametric estimation problems, see Section 3.9). In Chapter 5, we shall present a new application of these global bounds. Specifically, we shall employ the inequality (3.24) to prove a minimax lower bound having the optimal rate for the problem of estimating a convex set from noisy measurements of its support function.

We would like to remark, however, that these global bounds are not useful in applications where the quantities $N(\eta)$ and $M_\alpha(\epsilon; \Theta)$ are infinite or difficult to bound. This is the case, for example, in the covariance matrix estimation problem considered in Chapter 4, where it is problematic to apply the global bounds. In such situations, as we show for the covariance matrix estimation problem in Chapter 4, the inequalities (3.19), (3.20), (3.21) can still be effectively employed to result in optimal lower bounds.

3.9 Differences between the Global Bounds

In this section, we shall present examples of estimation problems where the global lower bound (3.22) yields results that are quite different in character from those given by inequalities (3.23) and (3.24). Specifically, we shall consider standard parametric estimation problems. In these problems, it has been observed by Yang and Barron

(1999, Page 1574) that (3.22) only results in sub-optimal lower bounds for the minimax risk. We show, on the other hand, that (3.24) (and (3.23)) produce rate-optimal lower bounds.

According to statistical folklore, one needs more than global covering number bounds (also known as global metric entropy bounds) to capture the usual minimax rate (under squared error loss) for classical parametric estimation problems. Indeed, Yang and Barron (1999, Page 1574-1575) were quite explicit on this point:

> For smooth finite-dimensional models, the minimax risk can be solved using some traditional statistical methods (such as Bayes procedures, Cramér-Rao inequality, Van Tree's inequality, etc.), but these techniques require more than the entropy condition. If local entropy conditions are used instead of those on global entropy, results can be obtained suitable for both parametric and nonparametric families of densities.

Nevertheless, as shown by the following examples, inequalities (3.24) and (3.23), that are based on divergences D_α with respect to $\alpha > 1$ as opposed to $\alpha = 1$, can derive lower bounds with optimal rates of convergence from global bounds.

We would like to stress here that these examples are presented merely as toy examples to note a difference between the two global bounds (3.22) and (3.24) (which provides a justification for using divergences other than the Kullback-Leibler divergence for minimax lower bounds) and also to emphasize the fact that global characteristics are enough to obtain minimax lower bounds even in finite dimensional problems. In each of the following examples, obtaining the optimal minimax lower bound is actually quite simple using other techniques.

In the first three examples, we take the parameter space Θ to be a bounded interval of the real line and we consider the problem of estimating a parameter

$\theta \in \Theta$ from n independent observations distrbuted according to m_θ, where m_θ is a probability measure on the real line. The probability measure P_θ accordingly equals the n-fold product of m_θ.

We work with the usual squared error loss $L(\theta, a) = (\theta - a)^2$. Because $d(\theta_1, \theta_2) = \inf_{a \in \mathbb{R}}(L(\theta_1, a) + L(\theta_2, a)) \geq (\theta_1 - \theta_2)^2/2$, the quantity $N(\eta)$ appearing in (3.22), (3.24) and (3.23), which is the size of a maximal η-separated subset of Θ, is larger than $c_1 \eta^{-1/2}$ for $\eta \leq \eta_0$ where c_1 and η_0 are positive constants depending on the bounded parameter space alone. We encounter more positive constants $c, c_2, c_3, c_4, c_5, \epsilon_0$ and ϵ_1 in the examples all of which depend possibly on the parameter space alone and thus, independent of n.

In the following, we focus on the performance of inequality (3.24). The behavior of (3.23) for $l > 1$ is similar to the $l = 2$ case.

Example 3.9.1. Suppose that m_θ equals the normal distribution with mean θ and variance 1. The chi-squared divergence $D_2(P_\theta || P_{\theta'})$ equals $\exp\left(n|\theta - \theta'|^2\right) - 1$ which implies that $D_2(P_\theta || P_{\theta'}) \leq \epsilon^2$ if and only if $|\theta - \theta'| \leq \sqrt{\log(1+\epsilon^2)}/\sqrt{n}$. Thus $M_2(\epsilon; \Theta) \leq c_2 \sqrt{n}/\sqrt{\log(1+\epsilon^2)}$ for $\epsilon \leq \epsilon_0$ and consequently, from (3.24),

$$R_n \geq \sup_{\eta \leq \eta_0, \epsilon \leq \epsilon_0} \frac{\eta}{2} \left(1 - \frac{\sqrt{\eta}}{c_1} - (\eta n)^{1/4} \sqrt{\frac{c_2(1+\epsilon^2)}{c_1 \sqrt{\log(1+\epsilon^2)}}}\right).$$

Taking $\epsilon = \epsilon_0$ and $\eta = c_3/n$, we get

$$R_n \geq \frac{c_3}{2n} \left(1 - \frac{\sqrt{c_3}}{c_1 \sqrt{n}} - c_3^{1/4} c_4\right), \tag{3.25}$$

where c_4 depends only on c_1, c_2 and ϵ_0. Hence by choosing c_3 small, we get that $R_n \geq c/n$ for all large n.

□

The next two examples consider standard irregular parametric estimation problems.

Example 3.9.2. Suppose that Θ is a compact interval of the positive real line that is bounded away from zero and suppose that m_θ denotes the uniform distribution on $[0, \theta]$. The chi-squared divergence, $D_2(P_\theta || P_{\theta'})$, equals $(\theta'/\theta)^n - 1$ if $\theta \leq \theta'$ and ∞ otherwise. It follows accordingly that $D_2(P_\theta || P_{\theta'}) \leq \epsilon^2$ provided $0 \leq n(\theta' - \theta) \leq \theta \log(1 + \epsilon^2)$. Because Θ is compact and bounded away from zero, $M_2(\epsilon; \Theta) \leq c_2 n / \log(1 + \epsilon^2)$ for $\epsilon \leq \epsilon_0$. Applying (3.24), we obtain

$$R_n \geq \sup_{\eta \leq \eta_0, \epsilon \leq \epsilon_0} \frac{\eta}{2} \left(1 - \frac{\sqrt{\eta}}{c_1} - \sqrt{n \sqrt{\eta}} \sqrt{\frac{c_2 (1 + \epsilon^2)}{c_1 \log(1 + \epsilon^2)}} \right).$$

Taking $\epsilon = \epsilon_0$ and $\eta = c_3/n^2$, we get that

$$R_n \geq \frac{c_3}{2n^2} \left(1 - \frac{\sqrt{c_3}}{n c_1} - c_3^{1/4} c_4 \right),$$

where c_4 depends only on c_1, c_2 and ϵ_0. Hence by choosing c_3 sufficiently small, we get that $R_n \geq c/n^2$ for all large n. This is the optimal minimax rate for this problem as can be seen by estimating θ by the maximum of the observations.

□

Example 3.9.3. Suppose that m_θ denotes the uniform distribution on the interval $[\theta, \theta + 1]$. We argue that $M_2(\epsilon; \Theta) \leq c_2 / \left((1 + \epsilon^2)^{1/n} - 1 \right)$ for $\epsilon \leq \epsilon_0$. To see this, let us define ϵ' so that $2\epsilon' := (1 + \epsilon^2)^{1/n} - 1$ and let G denote an ϵ'-grid of points in the interval Θ; G would contain at most c_2/ϵ' points when $\epsilon \leq \epsilon_0$. For a point α in the grid, let Q_α denote the n-fold product of the uniform distribution on the interval $[\alpha, \alpha + 1 + 2\epsilon']$. Now, for a fixed $\theta \in \Theta$, let α denote the point in the grid such that $\alpha \leq \theta \leq \alpha + \epsilon'$. It can then be checked that the chi-squared divergence between P_θ

and Q_α is equal to $(1+2\epsilon')^n - 1 = \epsilon^2$. Hence $M_2(\epsilon, \Theta)$ can be taken to be the number of probability measures Q_α, which is the same as the number of points in G. This proves the claimed upper bound on $M_2(\epsilon; \Theta)$.

It can be checked by elementary calculus (Taylor expansion, for example) that the inequality

$$(1+\epsilon^2)^{1/n} - 1 \geq \frac{\epsilon^2}{n} - \frac{1}{2n}\left(1 - \frac{1}{n}\right)\epsilon^4$$

holds for $\epsilon \leq \sqrt{2}$ (in fact for all ϵ, but for $\epsilon > \sqrt{2}$, the right hand side above may be negative). Therefore for $\epsilon \leq \min(\epsilon_0, \sqrt{2})$, we get that

$$M_2(\epsilon; \Theta) \leq \frac{2nc_2}{2\epsilon^2 - (1 - 1/n)\epsilon^4}.$$

From inequality (3.24), we get that for every $\eta \leq \eta_0$ and $\epsilon \leq \min(\epsilon_0, \sqrt{2})$,

$$R_n \geq \frac{\eta}{2}\left(1 - \frac{\sqrt{\eta}}{c_1} - \sqrt{n\sqrt{\eta}}\sqrt{\frac{2(1+\epsilon^2)c_2}{c_1\left(2\epsilon^2 - (1-1/n)\epsilon^4\right)}}\right).$$

If we now take $\epsilon = \min(\epsilon_0, 1)$ and $\eta = c_3/n^2$, we see that the quantity inside the parantheses converges (as $n \to \infty$) to $1 - c_3^{1/4}c_4$ where c_4 depends only on c_1, c_2 and ϵ_0. Therefore by choosing c_3 sufficiently small, we get that $R_n \geq c/n^2$. This is the optimal minimax rate for this problem as can be seen by estimating θ by the minimum of the observations.

□

Next, we consider a d-dimensional normal mean estimation problem and show that the bound given by (3.24) has the correct dependence on the dimension d.

Example 3.9.4. Let Θ denote the ball in $\mathbb{R}^d$ of radius Γ centered at the origin. Let us consider the problem of estimating $\theta \in \Theta$ from an observation X distributed according to the normal distribution with mean θ and variance covariance matrix

$\sigma^2 I_d$, where I_d denotes the identity matrix of order d. Thus P_θ denotes the $N(\theta, \sigma^2 I_d)$ distribution. We assume squared error loss: $L(\theta, a) = ||\theta - a||^2$.

We use inequality (3.24) to show that the minimax risk R for this problem is larger than or equal to a constant multiple of $d\sigma^2$ when $\Gamma \geq \sigma\sqrt{d}$.

The first step is to note that by standard volumetric arguments, we can take

$$N(\eta) = \left(\frac{\Gamma}{\sqrt{2\eta}}\right)^d, M_2(\epsilon, \Theta) = \left(\frac{3\Gamma}{\sigma\sqrt{\log(1+\epsilon^2)}}\right)^d \tag{3.26}$$

whenever $\sigma\sqrt{\log(1+\epsilon^2)} \leq \Gamma$.

Applying inequality (3.24) with (3.26), we get that, for every $\eta > 0$ and $\epsilon > 0$ such that $\sigma\sqrt{\log(1+\epsilon^2)} \leq \Gamma$, we have

$$R \geq \frac{\eta}{2}\left(1 - \left(\frac{\sqrt{2\eta}}{\Gamma}\right)^d - \left(\frac{3\sqrt{2\eta}}{\sigma}\right)^{d/2} \frac{\sqrt{1+\epsilon^2}}{(\log(1+\epsilon^2))^{d/4}}\right).$$

Now by elementary calculus, it can be checked that the function $\epsilon \mapsto \sqrt{1+\epsilon^2}/(\log(1+\epsilon^2))^{d/4}$ is minimized (subject to $\sigma\sqrt{\log(1+\epsilon^2)} \leq \Gamma$) when $1+\epsilon^2 = e^{d/2}$. We then get that

$$R \geq \sup_{\eta>0} \frac{\eta}{2}\left(1 - \left(\frac{\sqrt{2\eta}}{\Gamma}\right)^d - \left(\frac{36e\eta}{\sigma^2 d}\right)^{d/4}\right).$$

We now take $\eta = c_1 d\sigma^2$ and since $\Gamma \geq \sigma\sqrt{d}$, we obtain

$$R \geq \frac{c_1\sigma^2 d}{2}\left(1 - (2c_1)^{d/2} - (36ec_1)^{d/4}\right).$$

We can therefore choose c_1 small enough to obtain that $R \geq cd\sigma^2$ for a constant c that is independent of d. Up to constants independent of d, this lower bound is optimal for the minimax risk R because $\mathbb{E}_\theta L(X, \theta) = d\sigma^2$.

□

Bibliography

Ali, S. M. and S. D. Silvey (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B 28*, 131–142.

Birgé, L. (1983). Approximation dans les espaces metriques et theorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65*, 181–237.

Birgé, L. (2005). A new bound for multiple hypothesis testing. *IEEE Transactions on Information Theory 51*, 1611–1615.

Csisźar, I. (1963). Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität on markoffschen ketten. *Publ. Math. Inst. Hungar. Acad. Sci., Series A 8*, 84–108.

Guntuboyina, A. (2011). Lower bounds for the minimax risk using f divergences, and applications. *IEEE Transactions on Information Theory 57*, 2386–2399.

Gushchin, A. A. (2003). On Fano's lemma and similar inequalities for the minimax risk. *Theor. Probability and Math. Statist. 67*, 29–41.

Han, T. S. and S. Verdú (1994). Generalizing the Fano inequality. *IEEE Transactions on Information Theory 40*, 1247–1251.

Liese, F. and I. Vajda (1987). *Convex Statistical Distances.* Leipzig: Teubner.

Nemirovski, A. S. (2000). Topics in nonparametric statistics. In *Lecture on Probability Theory and Statistics, École d'Été de Probabilitiés de Saint-flour XXVIII-1998*, Volume 1738. Berlin, Germany: Springer-Verlag. Lecture Notes in Mathematics.

Topsøe, F. (2000). Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Theory 46*, 1602–1609.

Vajda, I. (2009). On metric divergences of probability measures. *Kybernetika 45*, 885–900.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Chapter 4

Covariance Matrix Estimation

4.1 Introduction

In this chapter we illustrate the use of the methods from the previous two chapters to reprove a recent minimax lower bound due to Cai, Zhang, and Zhou (2010), henceforth referred to as CZZ, for the following covariance matrix estimation problem. Let $X_1, \ldots, X_n$ be independent $p \times 1$ random vectors each distributed according to $N_p(0, \Sigma)$, the p-variate normal distribution with mean zero and covariance matrix Σ, where $\Sigma \in \mathcal{M}(\alpha)$ for some $\alpha > 0$. The set $\mathcal{M}(\alpha)$ is defined to be the set of all $p \times p$ covariance matrices (σ_{ij}) for which $|\sigma_{ij}| \leq |i-j|^{-\alpha-1}$ for $i \neq j$ and whose eigenvalues all lie in $[0, 2]$. The goal is to estimate $\Sigma \in \mathcal{M}(\alpha)$ under the loss function $L(\Sigma_1, \Sigma_2) := ||\Sigma_1 - \Sigma_2||^2$, where $|| \cdot ||$ denotes spectral (or operator) norm: $||A|| := \max\{||Ax|| : ||x|| \leq 1\}$.

Amongst other results, CZZ showed that

$$R_n(\alpha) := \inf_{\hat{\Sigma}} \sup_{\Sigma \in \mathcal{M}(\alpha)} \mathbb{E}_\Sigma L(\Sigma, \hat{\Sigma}) \geq c_1 \, n^{-\alpha/(2\alpha+1)} \qquad \text{if } p \geq c_2 n^{1/(2\alpha+1)} \tag{4.1}$$

where c_1 and c_2 are positive constants depending on α alone.

CZZ proved this inequality by constructing a map $\psi : \{0,1\}^m \to \mathcal{M}(\alpha)$ and applying Assouad's inequality,

$$R_n(\alpha) \geq \frac{m\zeta}{4} \min_{\Upsilon(\tau,\tau')=1} ||P_{\psi(\tau)} \wedge P_{\psi(\tau')}||_1. \tag{4.2}$$

where ζ satisfies $d(\psi(\tau), \psi(\tau')) \geq \zeta \sum_{i=1}^m \{\tau_i \neq \tau_i'\}$ for all $\tau, \tau \in \{0,1\}^m$. Here $d(\Sigma_1, \Sigma_2) := \inf_\Sigma (L(\Sigma_1, \Sigma) + L(\Sigma_2, \Sigma))$, the infimum being over all covariance matrices Σ. Also, for $\Sigma \in \mathcal{M}(\alpha)$, P_Σ denotes the probability measure $\otimes_{i=1}^n N(0, \Sigma)$.

CZZ's proof is described in the next section. The covariance matrices $\psi(\tau)$ in CZZ's construction can be viewed as perturbations of the identity matrix, which is an *interior* element of the parameter space $\mathcal{M}(\alpha)$. We show, in Section 4.4, that (4.1) can also be proved by the use of Assouad's inequality with another construction $\phi(\tau)$ whose members are perturbations of a matrix T which can be considered to be near the *boundary* (as opposed to the interior) of $\mathcal{M}(\alpha)$. Specifically, we use (4.2) with the map $\phi : \{0,1\}^m \to \mathcal{M}(\alpha)$ where each $\phi(\tau)$ is a perturbation of the matrix $T = (t_{ij})$ with $t_{ii} = 1$ and $t_{ij} = \gamma|i-j|^{-\alpha-1}$, for some small, positive constant γ.

In Section 4.5, we show how the inequalities from Chapter 3, can also be used to prove (4.1). Recall the following minimax lower bounds from Chapter 3,

$$R_n(\alpha) \geq \frac{\eta}{2}\left(1 - \frac{\log 2 + \log M_1(\epsilon; F) + \epsilon^2}{\log N}\right), \tag{4.3}$$

and

$$R_n(\alpha) \geq \frac{\eta}{2}\left(1 - \frac{1}{N} - \sqrt{\frac{(1+\epsilon^2)M_2(\epsilon; F)}{N}}\right). \tag{4.4}$$

Here $F \subset \mathcal{M}(\alpha)$ is η-separated i.e., it satisfies $d(A_1, A_2) \geq \eta$ for all $A_1, A_2 \in F$ with $A_1 \neq A_2$. Also N denotes the cardinality of F, and $M_1(\epsilon; F)$ and $M_2(\epsilon; F)$ denote the smallest number of probability measures needed to cover $\{P_A : A \in F\}$ up to

ϵ^2 in the Kullback-Leibler divergence and the chi-squared divergence respectively. In Section 4.5, we prove (4.1) by applying these inequalities with F chosen to be a well-separated subset of perturbations of T.

The inequality (4.3), due to Yang and Barron (1999), was intended by them to be used in situations where the (global) covering numbers of the entire parameter space are available. For this covariance matrix estimation problem however, the covering numbers of the parameter space $\mathcal{M}(\alpha)$ are unknown and hence, can not be used to bound the local covering number $M_1(\epsilon; F)$. Instead, we bound $M_1(\epsilon; F)$ and $M_2(\epsilon; F)$ from above directly without recourse to global covering bounds. This use of (4.3) in a situation where the global covering numbers are unknown is new.

Before proceeding, we put this problem in the decision theoretic setting considered in Chapter 2 by taking $\Theta = \mathcal{M}(\alpha)$, the action space to consist of all covariance matrices and the loss function, $L(\Sigma_1, \Sigma_2) = ||\Sigma_1 - \Sigma_2||^2$. The distance function $d(\Sigma_1, \Sigma_2)$ has, by triangle inequality, the following simple lower bound:

$$d(\Sigma_1, \Sigma_2) \geq \frac{1}{2} \inf_{\Sigma} \left(||\Sigma_1 - \Sigma|| + ||\Sigma_2 - \Sigma||\right)^2 \geq \frac{1}{2} ||\Sigma_1 - \Sigma_2||^2. \tag{4.5}$$

Throughout this chapter, we shall use c to denote a positive constant that depends on α alone (and hence has no relation to n or p) and whose specific value may change from place to place.

4.2 The proof of CZZ

Working with the assumption $p \geq 2n^{1/(2\alpha+1)}$, CZZ applied (4.2) to $m = n^{1/(2\alpha+1)}$ and $\psi : \{0,1\}^m \to \mathcal{M}(\alpha)$ defined by

$$\psi(\tau) := I_{p\times p} + c\ m^{-(\alpha+1)} \sum_{k=1}^{m} \tau_k B(k,m), \tag{4.6}$$

where $B(k,m) := (b_{ij})$ with b_{ij} taking the value 1 if either $(i = k, k+1 \leq j \leq 2m)$ or $(j = k, k+1 \leq i \leq 2m)$ and the value 0 otherwise and c is a constant (depending on α alone) that is small enough so that $\psi(\tau) \in \mathcal{M}(\alpha)$ for every $\tau \in \{0,1\}^m$.

To control each of the terms appearing in right hand side of (4.2), CZZ proved the following pair of inequalities:

$$d\left(\psi(\tau), \psi(\tau')\right) \geq c\Upsilon(\tau,\tau')m^{-2\alpha-1} \text{ and } \min_{\Upsilon(\tau,\tau')=1} ||P_{\psi(\tau)} \wedge P_{\psi(\tau')}||_1 \geq c$$

The required inequality (4.1) is a direct consequence of the application of Assouad's inequality (4.2) with the above pair of inequalities.

4.3 Finite Parameter Subset Construction

In this section, a finite subset of matrices in $\mathcal{M}(\alpha)$ are described whose elements are perturbations of the matrix T defined by $t_{ii} = 1$ and $t_{ij} = \gamma|i-j|^{-\alpha-1}$ where γ is a positive real number to be specified shortly. In subsequent sections, different proofs of (4.1) based on this construction are provided.

Fix a positive integer $k \leq p/2$ and partition T as

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{12}^T & T_{22} \end{bmatrix},$$

where T_{11} is $k \times k$ and T_{22} is $(p-k) \times (p-k)$. For each $\tau \in \{0,1\}^k$, consider the following matrix

$$T(\tau) := \begin{bmatrix} T_{11} & S(\tau)T_{12} \\ T_{12}^T S(\tau) & T_{22} \end{bmatrix}, \qquad \text{where } S(\tau) := \operatorname{diag}(\tau_1, \dots, \tau_k).$$

Lemma 4.3.1. *If $0 < \gamma \sum_{l \geq 1} l^{-\alpha-1} < 1/6$, then the eigenvalues of $T(\tau)$ lie in the interval $(2/3, 4/3)$ for every $\tau \in \{0,1\}^k$.*

Proof. Fix $\tau \in \{0,1\}^k$. The assumption on γ ensures that $T(\tau)$ is diagonally dominated. We shall denote the $(i,j)^{\text{th}}$ entry of $T(\tau)$ by $t_{ij}(\tau)$. Let λ be an eigenvalue of $T(\tau)$ and assume that $x \neq 0$ satisfies $T(\tau)x = \lambda x$. This can be rewritten as $(\lambda - t_{ii}(\tau))x_i = \sum_{j:j \neq i} t_{ij}(\tau)x_j$ for every i. Using this for the index i_0 for which $|x_{i_0}| = \max_j |x_j|$ (note that this implies that $x_{i_0} \neq 0$ because $x \neq 0$) and noting that $t_{ii}(\tau) = 1$ for all i, we get

$$|\lambda - 1||x_{i_0}| \leq \sum_{j:j \neq i_0} |t_{ij}(\tau)||x_j| \leq |x_{i_0}| 2\gamma \sum_{l \geq 1} l^{-\alpha-1}.$$

Thus if γ is chosen as in the statement of the lemma, we would obtain that $|\lambda - 1| < 1/3$ or $\lambda \in (2/3, 4/3)$. □

For use in the subsequent sections, we need the following two results which provide lower bounds for $d(T(\tau), T(\tau'))$ and upper bounds for divergences between $P_{T(\tau)}$ and $P_{T(\tau')}$ respectively.

Lemma 4.3.2. *For every $\tau, \tau' \in \{0,1\}^k$, we have*

$$d(T(\tau), T(\tau')) \geq c\, k^{-2\alpha-1} \Upsilon(\tau, \tau') \qquad \textit{with } \Upsilon(\tau, \tau') = \sum_{i=1}^{k} \{\tau_i \neq \tau_i'\}. \tag{4.7}$$

Proof. Fix $\tau, \tau' \in \{0,1\}^k$ with $\tau \neq \tau'$. According to inequality (4.5), $d(T(\tau), T(\tau')) \geq ||T(\tau) - T(\tau')||^2/2$. To bound the spectral norm of $T(\tau) - T(\tau')$ from below, let v denote the $p \times 1$ vector $(0_k, 1_k, 0_{p-2k})^T$, where 0_k and 0_{p-2k} denote the $k \times 1$ and $(2p-k) \times 1$ vectors of zeros respectively and 1_k denotes the vector of ones. Clearly $||v||^2 = k$ and $(T(\tau) - T(\tau'))v$ is of the form $(u, 0)^T$ with $u = (u_1, \dots, u_k)^T$ given by $u_r = (\tau_r - \tau_r') \sum_{s=1}^k t_{r,k+s}$. Thus

$$\begin{aligned}|u_r| &= \{\tau_r \neq \tau_r'\} \sum_{s=1}^{k} \gamma |r-k-s|^{-\alpha-1} \\ &\geq \{\tau_r \neq \tau_r'\} \sum_{i=k}^{2k-1} \gamma i^{-\alpha-1} \geq c\gamma k^{-\alpha} \{\tau_r \neq \tau_r'\}.\end{aligned}$$

Therefore,

$$|| \left(T(\tau) - T(\tau')\right) v||^2 \geq \sum_{r=1}^{k} u_r^2 \geq c^2 \gamma^2 k^{-2\alpha} \Upsilon(\tau, \tau').$$

Using a new constant c for $c^2\gamma^2$ and noting that $||v||^2 = k$, we obtain the required inequality (4.7). □

The Frobenius norm of a matrix A is defined by $||A||_F := \sqrt{\sum_{i,j} a_{ij}^2}$. The following result gives an upper bound for divergences (chi-squared, Kullback-Leibler and total variation) between $P_{T(\tau)}$ and $P_{T(\tau')}$ in terms of the Frobenius norm of the difference $T(\tau) - T(\tau')$. It is based on a more general result, Theorem 4.6.1 (stated and proved in the Appendix), that relates the chi-squared divergence between two zero mean normal distrbutions to the Frobenius norm of the difference of their covariance matrices.

Lemma 4.3.3. *For every $\tau, \tau' \in \{0,1\}^k$, the following inequalities hold, provided*

$||T(\tau) - T(\tau')||_F^2 \leq 2/9$,

$$D_2(P_{T(\tau)}||P_{T(\tau')}) \leq \exp\left(\frac{3n}{2}||T(\tau) - T(\tau')||_F^2\right) - 1, \tag{4.8}$$

$$D_1(P_{T(\tau)}||P_{T(\tau')}) \leq \frac{3n}{2}||T(\tau) - T(\tau')||_F^2 \tag{4.9}$$

and

$$\left|\left|P_{T(\tau)} \wedge P_{T(\tau)}\right|\right|_1 \geq 1 - \sqrt{\frac{3n}{4}}||T(\tau) - T(\tau')||_F. \tag{4.10}$$

Moreover, the Frobenius norm $||T(\tau) - T(\tau)||_F$ *has the following bound:*

$$||T(\tau) - T(\tau')||_F^2 \leq \frac{2^{2\alpha+3}\gamma}{2\alpha+1}\sum_{i=1}^{k}\{\tau_i \neq \tau_i'\}(k-i+1)^{-2\alpha-1}. \tag{4.11}$$

Proof. Fix $\tau, \tau' \in \{0,1\}^k$ with $\tau \neq \tau'$. The proof of (4.8) is provided below. Inequalities (4.9) and (4.10) follow from (4.8) because

$$D_1(P_{T(\tau)}||P_{T(\tau')}) \leq \log\left(1 + D_2(P_{T(\tau)}||P_{T(\tau')})\right),$$

which is a consequence of Jensen's inequality and

$$\left|\left|P_{T(\tau)} \wedge P_{T(\tau')}\right|\right|_1 \geq 1 - \sqrt{\frac{D_1(P_{T(\tau)}||P_{T(\tau')})}{2}},$$

which is a consequence of Pinsker's inequality. Let χ^2 denote the chi-squared divergence between two zero mean normal distributions with covariance matrices $T(\tau)$ and $T(\tau')$ respectively. Because $P_{T(\tau)}$ is the n-fold product of the p-variate normal distribution with mean zero and covariance matrix $T(\tau)$, it follows from the formula for D_2 in terms of the marginal chi-squared divergences that

$$D_2(P_{T(\tau)}||P_{T(\tau')}) = (1+\chi^2)^n - 1.$$

From inequality (4.16) in Theorem (4.6.1), we have

$$\chi^2 \leq \exp\left(\frac{||T(\tau) - T(\tau')||_F^2}{\lambda_{\min}^2(T(\tau'))}\right) - 1$$

provided $||T(\tau) - T(\tau')||_F^2 \leq 2/9$. The following conditions that were required in Theorem 4.6.1 for (4.16) to hold:

$$2\Sigma_1^{-1} > \Sigma_2^{-1} \text{ and } 2||\Sigma_1 - \Sigma_2||_F^2 \leq \lambda_{\min}^2(\Sigma_2)$$

are satisfied for $\Sigma_1 = T(\tau)$ and $\Sigma_2 = T(\tau')$, provided $||T(\tau) - T(\tau')||_F^2 \leq 2/9$, because all the eigenvalues of $T(\tau)$ and $T(\tau')$ lie in $(2/3, 4/3)$. This proves inequalities (4.8), (4.9) and (4.10).

For (4.11), note that, by definition of Frobenius norm,

$$||T(\tau) - T(\tau')||_F^2 \leq 2\sum_{i=1}^{k}\{\tau_i \neq \tau_i'\}\sum_{j=k+1}^{p} t_{ij}^2 \leq 2\gamma\sum_{i=1}^{k}\{\tau_i \neq \tau_i'\}\sum_{j\geq k-i+1} j^{-2\alpha-2}.$$

Now, by the elementary inequality $j^{-2\alpha-2} \leq \int_j^{j+1}(x/2)^{-2\alpha-2}dx$ for $j \geq 1$, we obtain

$$\sum_{j\geq k-i+1} j^{-2\alpha-2} \leq 2^{2\alpha+2}\int_{k-i+1}^{\infty} x^{-2\alpha-2}dx = \frac{2^{2\alpha+2}}{2\alpha+1}(k-i+1)^{-2\alpha-1}.$$

The preceding two inequalities imply (4.11) and the proof is complete. □

Remark 4.3.1. *From the bound* (4.11)*, it follows that*

$$||T(\tau) - T(\tau')||_F^2 \le \frac{2^{2\alpha+3}\gamma}{2\alpha+1} \sum_{j \ge 1} j^{-2\alpha+1}.$$

We shall assume for the remainder of this chapter that the constant γ satisfies the condition in Lemma 4.3.1 and is chosen small enough so that $||T(\tau)-T(\tau')||_F^2 \le 2/9$ so that all the three inequalities (4.8)*,* (4.9) *and* (4.10) *hold for every $\tau, \tau' \in \{0,1\}^k$.*

4.4 Proof by Assouad's inequality

In this section, (4.1) is proved by the application of Assouad's inequality (4.2) to the matrices $T(\tau), \tau \in \{0,1\}^k$ described in the previous section.

It might seem natural to apply Assouad's inequality to $m = k$ and $\phi(\tau) = T(\tau)$. But this would not yield (4.1). The reason is that the matrices $T((0,\ldots,0,0))$ and $T((0,\ldots,0,1))$ are quite far away from each other which leads to the affinity term $\min_{\Upsilon(\tau,\tau')=1} \left|\left|P_{T(\tau)} \wedge P_{T(\tau')}\right|\right|_1$ being rather small.

The bound (4.1) can be proved by taking $m = k/2$ and applying Assouad's inequality to $\phi(\theta) = T((\theta_1,\ldots,\theta_m,0,\ldots,0)), \theta \in \{0,1\}^m$. The right hand side of (4.2) can then be bounded from below by the following inequalities:

$$d(\phi(\theta),\phi(\theta')) \ge c\Upsilon(\theta,\theta')k^{-2\alpha-1} \text{ and } \min_{\Upsilon(\theta,\theta')=1} ||P_{\phi(\theta)} \wedge P_{\phi(\theta')}||_1 \ge 1 - \sqrt{\frac{n}{ck^{2\alpha+1}}}.$$

The first inequality above directly follows from Lemma 4.3.2. The second inequality is a consequence of inequalities (4.10) and (4.11) in Lemma 4.3.3. Indeed, (4.10) bounds the affinity in terms of the Frobenius norm $||\phi(\theta)-\phi(\theta')||_F$ and, using (4.11), this Frobenius norm can be bounded, for $\theta, \theta' \in \{0,1\}^m$ with $\Upsilon(\theta,\theta') = 1$, in the following way (note that $m = k/2$)

$$||\phi(\theta) - \phi(\theta')||_F^2 \leq \frac{2^{2\alpha+3}\gamma}{2\alpha+1} \sum_{i=1}^{m} \{\theta_i \neq \theta_i'\}(k-i+1)^{-2\alpha-1} \leq \frac{1}{c}(k-m+1)^{-2\alpha-1} \leq \frac{k^{-2\alpha-1}}{c}.$$

Assouad's inequality (4.2) with the above pair of inequalities gives

$$R_n(\alpha) \geq ck^{-2\alpha}\left(1 - \sqrt{\frac{n}{ck^{2\alpha+1}}}\right) \qquad \text{for every } k \text{ with } 1 \leq k \leq p/2.$$

By choosing $k = (2n/c)^{1/(2\alpha+1)}$ (note that this choice of k would require the assumption $p \geq c_2 n^{1/(2\alpha+1)}$ for a constant c_2), we obtain (4.1).

4.5 Proofs using Inequalities (4.3) and (4.4)

We provide proofs of (4.1) using the inequalities (4.3) and (4.4). The set F will be chosen to be a sufficiently well-separated subset of $\{T(\tau) : \tau \in \{0,1\}^k\}$. By the Varshamov-Gilbert lemma (see for example Massart, 2007, Lemma 4.7), there exists a subset W of $\{0,1\}^k$ with $|W| \geq \exp(k/8)$ such that $\Upsilon(\tau, \tau') = \sum_i \{\tau_i \neq \tau_i'\} \geq k/4$ for all $\tau, \tau' \in W$ with $\tau \neq \tau'$. We take $F := \{T(\tau) : \tau \in W\}$ so that, by construction, $N = |F| = |W| \geq \exp(k/8)$.

According to Lemma 4.3.2, $d(T(\tau), T(\tau')) \geq ck^{-2\alpha}$ whenever $\Upsilon(\tau, \tau') \geq k/4$ which implies that F is an η-separated subset of $\mathcal{M}(\alpha)$ with $\eta := ck^{-2\alpha}$.

To bound the covering numbers of $P_A, A \in F$, let us fix $1 \leq l < k$ and define, for each $u \in \{0,1\}^{k-l+1}$,

$$S(u) := T(0, \dots, 0, u_1, \dots, u_{k-l+1}) \text{ and } Q_u := P_{S(u)}.$$

The inequality (4.11) gives

$$||T(\tau) - S(\tau_l, \dots, \tau_k)||_F^2 \leq \frac{1}{c}\sum_{i=1}^{l}(k-i+1)^{-2\alpha-1} \leq \frac{1}{c}(k-l)^{-2\alpha} \tag{4.12}$$

and thus, by (4.9), the following inequalities hold if $u = (\tau_1, \dots, \tau_k)$:

$$D_1\left(P_{T(\tau)}||Q_u\right) \leq \frac{n}{c}(k-l)^{-2\alpha} \text{ and } D_2\left(P_{T(\tau)}||Q_u\right) \leq \exp\left(\frac{n}{c}(k-l)^{-2\alpha}\right) - 1$$

It follows therefore that $Q_u, u \in \{0,1\}^{k-l+1}$ covers $P_A, A \in F$ up to ϵ_1^2 in Kullback-Leibler divergence and up to ϵ_2^2 in chi-squared divergence where $\epsilon_1^2 := n(k-l)^{-2\alpha}/c$ and $\epsilon_2^2 := \exp\left(n(k-l)^{-2\alpha}/c\right) - 1$. As a direct consequence, $M_1(\epsilon_1; F) \leq 2^{k-l+1}$ and $M_2(\epsilon_2; F) \leq 2^{k-l+1}$. Therefore, from (4.3),

$$R_n(\alpha) \geq ck^{-2\alpha}\left[1 - \frac{1}{ck}\left(k - l + \frac{n}{(k-l)^{2\alpha}}\right)\right] \tag{4.13}$$

and from (4.4),

$$R_n(\alpha) \geq ck^{-2\alpha}\left[1 - \exp\left(\frac{-k}{8}\right) - \exp\left(\frac{1}{c}\left(\frac{n}{(k-l)^{2\alpha}} + (k-l)\right) - \frac{k}{16}\right)\right] \tag{4.14}$$

for every $k \leq p/2$ and $1 \leq l < k$.

Each of the above two inequalities imply (4.1). Indeed, taking $k - l = n^{1/(2\alpha+1)}$ and $k = 4n^{1/(2\alpha+1)}/c$ in (4.13) implies (4.1).

Also, by taking $k - l = n^{1/(2\alpha+1)}$ and $k = (32/c)(1+B)n^{1/(2\alpha+1)}$ for $B \geq 0$ in (4.14), we get

$$R_n(\alpha) \geq ck^{-2\alpha}\left[1 - 2\exp\left(\frac{-2B}{c}n^{1/(2\alpha+1)}\right)\right] \geq ck^{-2\alpha}\left[1 - 2\exp\left(\frac{-2B}{c}\right)\right]$$

from which (4.1) is obtained by taking $B = (c \log 4)/2$.

Note that the choice of k necessitates that $p \geq c_2 n^{1/(2\alpha+1)}$ for a large enough constant c_2.

4.6 Appendix: Divergences between Gaussians

In this section, we shall prove a bound on the chi-squared divergence (which, in turn, implies bounds on the Kullback-Leibler divergence and testing affinity) between two zero mean gaussians by the Frobenius norm of the difference of their covariance matrices. The Frobenius norm of a matrix A is defined as

$$||A||_F := \sqrt{\sum_{i,j} a_{ij}^2} = \sqrt{\operatorname{tr}(AA^T)} = \sqrt{\operatorname{tr}(A^T A)}.$$

Two immediate consequences of the above definition are:

1. $||A||_F = ||UA||_F = ||AU||_F$ for every orthogonal matrix U.
2. $||A||_F^2 \min_i d_i^2 \leq ||DA||_F^2 \leq ||A||_F^2 \max_i d_i^2$ for every diagonal matrix D with diagonal entries d_i. Exactly the same relation holds if DA is replaced by AD.

Theorem 4.6.1. *The chi-squared divergence χ^2 between two normal distributions with mean 0 and covariance matrices Σ_1 and Σ_2 satisfies*

$$\chi^2 \leq \left(1 - \frac{||\Delta||_F^2}{\lambda_{\min}^2(\Sigma_2)}\right)_+^{-1/2} - 1 \qquad \textit{provided } 2\Sigma_1^{-1} > \Sigma_2^{-1}. \tag{4.15}$$

where $\Delta := \Sigma_1 - \Sigma_2$ and $|| \cdot ||_F$ denotes the Frobenius norm. Moreover, if $2||\Delta||_F^2 \leq \lambda_{\min}^2(\Sigma_2)$, then

$$\chi^2 \leq \exp\left(\frac{||\Delta||_F^2}{\lambda_{\min}^2(\Sigma_2)}\right) - 1. \tag{4.16}$$

Proof. When $2\Sigma_1^{-1} > \Sigma_2^{-1}$, it can checked by a routine calculation that

$$\chi^2 = \left| I - \left(\Sigma_2^{-1/2} \Delta \Sigma_2^{-1/2} \right)^2 \right|^{-1/2} - 1$$

where $\Delta = \Sigma_1 - \Sigma_2$ and $|\cdot|$ denotes determinant. Let $\lambda_1, \dots, \lambda_p$ be the eigenvalues of the symmetric matrix $\Sigma_2^{-1/2} \Delta \Sigma_2^{-1/2}$. Then $\chi^2 = [(1-\lambda_1^2)\dots(1-\lambda_p^2)]^{-1/2} - 1$ and consequently, by an elementary inequality, $\chi^2 \le (1 - \sum_i \lambda_i^2)_+^{-1/2} - 1$. Observe that $\sum_i \lambda_i^2 = \left|\left| \Sigma_2^{-1/2} \Delta \Sigma_2^{-1/2} \right|\right|_F^2$. Suppose that $\Sigma_2 = U \Lambda U^T$ for an orthogonal matrix U and a positive definite diagonal matrix Λ. Then $\Sigma_2^{-1/2} = U \Lambda^{-1/2} U^T$ and by properties of the Frobenius norm, we have

$$\sum_{i=1}^p \lambda_i^2 = \left|\left| \Sigma_2^{-1/2} \Delta \Sigma_2^{-1/2} \right|\right|_F^2 = \left|\left| \Lambda^{-1/2} U^T \Delta U \Lambda^{-1/2} \right|\right|_F^2 \le \frac{||U^T \Delta U||_F^2}{\lambda_{\min}^2(\Sigma_2)} = \frac{||\Delta||_F^2}{\lambda_{\min}^2(\Sigma_2)}.$$

This completes the proof of (4.15). The inequality (4.16) is a consequence of the elementary inequality $1 - x \ge e^{-2x}$ for $0 \le x \le 1/2$. □

Bibliography

Cai, T. T., C.-H. Zhang, and H. H. Zhou (2010). Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics 38*, 2118–2144.

Massart, P. (2007). *Concentration inequalities and model selection. Lecture notes in Mathematics*, Volume 1896. Berlin: Springer.

Yang, Y. and A. Barron (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics 27*, 1564–1599.

Chapter 5

Estimation of Convex Sets

5.1 Introduction

In this chapter, we study the problem of estimating a compact, convex set from noisy support function measurements. We use techniques described in Chapter 3 to prove a minimax lower bound. We also construct an estimator that achieves the lower bound up to multiplicative constants.

The support function h_K of a compact, convex subset K of $\mathbb{R}^d$ $(d \geq 2)$ is defined for u in the unit sphere, $S^{d-1} := \{x : \sum_i x_i^2 = 1\}$ by $h_K(u) := \sup_{x \in K} \langle x, u \rangle$, where $\langle x, u \rangle = \sum_i x_i u_i$. The support function is a fundamental quantity in convex geometry and a key fact (Schneider, 1993, Section 1.7 or Rockafellar, 1970, Section 13) is that $K = \cap_{u \in S^{d-1}} \left\{ x \in \mathbb{R}^d : \langle x, u \rangle \leq h_K(u) \right\}$ which, in particular, implies that K is uniquely determined by h_K.

We consider the problem of estimating K based on observations $(u_1, Y_1), \ldots, (u_n, Y_n)$ under the following three assumptions:

1. $Y_i = h_K(u_i) + \xi_i$ where $\xi_1, \ldots, \xi_n$ are independent normal random variables with mean zero and known variance σ^2,

2. $u_1, \dots, u_n$ are independently distributed according to the uniform distribution on S^{d-1},

3. $u_1, \dots, u_n$ are independent of $\xi_1, \dots, \xi_n$.

We summarize the history of this problem and provide motivation for its study in the next section.

We prove upper and lower bounds for the minimax risk

$$R(n) = R(n; \sigma, \Gamma) := \inf_{\hat{K}} \sup_{K \in \mathcal{K}^d(\Gamma)} \mathbb{E}_K \ell^2(K, \hat{K})$$

with

$$\ell^2(K, K') := \int_{S^{d-1}} (h_K(u) - h_{K'}(u))^2 d\nu(u),$$

where $\mathcal{K}^d(\Gamma)$ denotes the set of all compact, convex sets contained in the ball of radius Γ centered at the origin, and ν denotes the uniform probability measure on S^{d-1}. We assume that σ and Γ are known so that estimators in the definition of $R(n)$ are allowed to depend on them.

Specifically, we show that, in every dimension $d \geq 2$, the minimax risk $R(n)$ is bounded from above and below by constant multiples (which depend on d, σ and Γ) of $n^{-4/(d+3)}$. The lower bound is proved in Section 5.3 using an inequality from Chapter 3, and the upper bound is proved in Section 5.4.

A word on notation: In this chapter, by a constant, we mean a positive quantity that depends on the dimension d alone. We shall denote such constants by c, C, c_1, c' etc. and by δ_0 and ϵ_0. We are never explicit about the precise value of these constants and their value may change with every occurence.

5.2 Background

In two and three dimensions, the problem of recovering a compact, convex set from noisy support function measurements was studied in the context of certain engineering applications. For example, Prince and Willsky (1990), who were the first to propose the regression model $Y_i = h_K(u_i) + \xi_i$ for this problem, were motivated by application to Computed Tomography. Lele, Kulkarni, and Willsky (1992) showed how solutions to this problem can be applied to target reconstruction from resolved laser-radar measurements in the presence of registration errors. Gregor and Rannou (2002) considered applications to Projection Magnetic Resonance Imaging.

Additional motivation for studying this problem comes from the fact that it has a similar flavour to well-studied regression problems. For example,

1. It is essentially a nonparametric function estimation problem where the true function is assumed to be the support function of a compact, convex set i.e., there is an implicit convexity-based constraint on the true regression function. Regression and density esimation problems with explicit such constraints e.g., log-concave density estimation and convex regression have received much attention.

2. The model $Y_i = \max_{x \in K} \langle x, u \rangle + \xi_i$ can also be viewed as a variant of the usual linear regression model where the dependent variable is modeled as the maximum of linear combinations of the explanatory variables over a set of parameter values and the interest lies in estimating the convex hull of the set of parameters. While we do not know if this maximum regression model has been used outside the context of convex set estimation, the idea of combining linear functions of independent variables into nonlinear algorithmic prediction models for the response variable is familiar (as in neural networks).

The least squares estimator has been the most commonly used estimator for this problem. It is defined as

$$\hat{K}_{ls} := \operatorname*{argmin}_{L} \sum_{i=1}^{n} \left(Y_i - h_L(u_i)\right)^2, \tag{5.1}$$

where the minimum is taken over all compact, convex subsets L. The minimizer here is not unique and one can always take it to be a polyhedron. This estimator, for $d = 2$, was first proposed by Prince and Willsky (1990), who assumed that $u_1, \ldots, u_n$ are evenly spaced on the unit circle and that the error variables $\xi_1, \ldots, \xi_n$ are normal with mean zero. They also proposed an algorithm for computing it based on quadratic programming. Lele et al. (1992) extended this algorithm to include the case of non-evenly spaced $u_1, \ldots, u_n$ as well. Recently, Gardner and Kiderlen (2009) proposed an algorithm for computing a minimizer of the least squares criterion for every dimension $d \geq 2$ and every sequence $u_1, \ldots, u_n$.

In addition to the least squares estimator, Prince and Willsky (1990) and Lele et al. (1992) also proposed estimators (in the case $d = 2$) designed to take advantage of certain forms of prior knowledge, when available, about the true compact, convex set. These estimators are all based on a least squares minimization.

Fisher, Hall, Turlach, and Watson (1997) proposed estimators for $d = 2$ that are not based on the least squares criterion. They assumed that the support function h_K, viewed as a function on the unit circle or on the interval $(-\pi, \pi]$, is smooth and estimated it using periodic versions of standard nonparametric regression techniques such as local regression, kernel smoothing and splines. They suggested a way to convert the estimator of h_K into an estimator for K using a formula, which works for smooth h_K, for the boundary of K in terms of h_K. Hall and Turlach (1999) added a corner-finding technique to the method of Fisher et al. (1997) to estimate two-dimensional convex sets with certain types of corners.

There are relatively fewer theoretical results in the literature. Fisher et al. (1997, Theorem 4.1) stated a theorem without proof which appears to imply consistency and certain rates of convergence for their estimator under certain smoothness assumptions on the support function of the true compact, convex set K. Gardner, Kiderlen, and Milanfar (2006) proved consistency of the least squares estimator and also derived rates of convergence. They worked with the following assumptions:

1. $u_1, u_2, \dots$ are deterministic satisfying

$$\max_{u \in S^{d-1}} \min_{1 \le i \le n} ||u - u_i|| = O(n^{-1/(d-1)}) \qquad \text{as } n \to \infty,$$

2. $\xi_1, \xi_2, \dots$ are independent normal with mean zero and variance σ^2,

3. K is contained in a ball of radius Γ centered at the origin with $\Gamma \ge \sigma^{15/2}$.

Their Theorem 6.2 showed that $\ell^2(K, \hat{K}_{ls}) = O_{d,\sigma,\Gamma}(\beta_n)$ as n approaches ∞ almost surely, where

$$\beta_n := \begin{cases} n^{-4/(d+3)} & \text{when } d = 2, 3, 4 \\ n^{-1/2} (\log n)^2 & \text{when } d = 5 \\ n^{-2/(d-1)} & \text{when } d \ge 6. \end{cases} \tag{5.2}$$

Here $O_{d,\sigma,\Gamma}$ is the usual big-O notation where the constant involved depends on d, σ and Γ. Gardner et al. (2006, Theorem 6.2) provided explicit expressions for the dependence of the constant with respect to σ and Γ (but not d) which we have not shown here because our interest only lies in the dependence on n.

As part of our proof of the upper bound for the minimax risk, we construct an estimator with improved rates.

5.3 Lower Bound

The following theorem shows that $R(n)$ is at least $n^{-4/(d+3)}$ up to a multiplicative constant that depends only on d, σ and Γ.

Theorem 5.3.1. *There exist two positive constants c and C depending only on d (and independent of n, σ and Γ) such that*

$$R(n) \geq c\sigma^{8/(d+3)}\Gamma^{2(d-1)/(d+3)}n^{-4/(d+3)} \qquad \textit{whenever } n \geq C(\sigma/\Gamma)^2. \tag{5.3}$$

For the proof, we put this problem in the general decision-theoretic framework of Chapter 3 and use an inequality proved in Section 3.8. Let $\Theta = \mathcal{K}^d(\Gamma)$ and the action space $\mathcal{A}$ consist of all possible compact, convex subsets of $\mathbb{R}^d$. The loss function equals $L(K, K') = \ell^2(K, K')$. For $K \in \Theta$, let P_K denote the joint distribution of $(u_1, Y_1), \ldots, (u_n, Y_n)$. It may be recalled that a subset F of Θ is called η-separated if

$$\inf_{K \in \mathcal{A}} \left(\ell^2(K_1, K) + \ell^2(K_2, K)\right) \geq \eta \qquad \text{for all } K_1, K_2 \in F \text{ with } K_1 \neq K_2.$$

We use the following global minimax lower bound proved in Chapter 3 (see Section 3.8):

$$R(n) \geq \frac{\eta}{2}\left(1 - \frac{1}{N(\eta)} - \sqrt{\frac{(1+\epsilon^2)M_2(\epsilon;\Theta)}{N(\eta)}}\right) \qquad \text{for every } \eta > 0 \text{ and } \epsilon > 0, \tag{5.4}$$

where $N(\eta)$ is the size of a maximal η-separated subset of Θ and $M_2(\epsilon;\Theta)$ is the number of probability measures needed to cover $\{P_\theta, \theta \in \Theta\}$ up to ϵ^2 in the chi-squared divergence.

Proof. For the application of (5.4), we only need a lower bound for $N(\eta)$ and an upper bound for $M_2(\epsilon; \Theta)$. We start with $N(\eta)$. By the triangle inequality, we have

$$\ell^2(K_1, K) + \ell^2(K_2, K) \geq \frac{1}{2}\left(\ell(K_1, K) + \ell(K_2, K)\right)^2 \geq \frac{1}{2}\ell^2(K_1, K_2)$$

for every K_1, K_2 and K. It follows therefore that $N(\eta) \geq \tilde{N}(\sqrt{2\eta}; \ell)$, where $\tilde{N}(\delta; \ell)$ denotes the δ-packing number of $\mathcal{K}^d(\Gamma)$ under the metric ℓ i.e., the size of a maximal subset $F \subset \mathcal{K}^d(\Gamma)$ such that $\ell(K_1, K_2) \geq \delta$ for $K_1, K_2 \in F$ with $K_1 \neq K_2$.

Bronshtein (1976, Theorem 4 and Remark 1) proved that there exist positive constants c' and δ_0 depending only on d such that the δ-packing number of $\mathcal{K}^d(\Gamma)$ under the Hausdorff metric is at least $\exp\left(c'(\Gamma/\delta)^{(d-1)/2}\right)$ whenever $\delta \leq \Gamma\delta_0$. The Hausdorff distance is defined as $\ell_H(K, K') := \sup_{u \in S^{d-1}} |h_K(u) - h_{K'}(u)|$ and is clearly larger than $\ell(K, K')$.

It turns out that Bronshtein's result is true for the metric ℓ as well. This has not been proved anywhere in the literature however. We provide a proof in the Appendix (Theorem 5.5.1) by modifying Bronshtein's proof appropriately and using Varshamov-Gilbert lemma. Therefore, from Theorem 5.5.1, we have

$$\log N(\eta) \geq \log \tilde{N}(\sqrt{2\eta}; \ell) \geq c' \left(\frac{\Gamma}{\sqrt{\eta}}\right)^{(d-1)/2} \qquad \text{for } \eta \leq \Gamma^2\delta_0^2/2. \tag{5.5}$$

Let us now turn to $M_2(\epsilon; \Theta)$. For $K, K' \in \mathcal{K}^d(\Gamma)$, the chi-squared divergence $D_2(P_K || P_{K'})$ satisfies

$$1 + D_2(P_K || P_{K'}) = \left(\int_{S^{d-1}} \exp\left(\frac{(h_K(u) - h_{K'}(u))^2}{\sigma^2}\right) du\right)^n \leq \exp\left(\frac{n\ell_H^2(K, K')}{\sigma^2}\right).$$

As a result,

$$D_2(P_K || P_{K'}) \leq \epsilon^2 \qquad \text{whenever } \ell_H(K, K') \leq \epsilon' := \sigma\sqrt{\log(1+\epsilon^2)}/\sqrt{n}. \tag{5.6}$$

Let $W_{\epsilon'}$ be the ϵ'-covering number for $\mathcal{K}^d(\Gamma)$ in the Hausdorff metric i.e., it is the smallest W for which there exist sets $K_1, \ldots, K_W$ in $\mathcal{K}^d(\Gamma)$ having the property that for every set $L \in \mathcal{K}^d(\Gamma)$, there exists a K_j such that $\ell_H(L, K_j) \leq \epsilon'$. Bronshtein (1976, Theorem 3 and Remark 1) showed that there exist positive constants c'' and ϵ_0 depending only on d such that $\log W_{\epsilon'}$ is at most $c''(\Gamma/\epsilon')^{(d-1)/2}$ whenever $\epsilon' \leq \Gamma\epsilon_0$. Consequently, from (5.6), we obtain

$$\log M_2(\epsilon; \Theta) \leq c'' \left(\frac{\Gamma\sqrt{n}}{\sigma\sqrt{\log(1+\epsilon^2)}}\right)^{(d-1)/2} \qquad \text{if } \log(1+\epsilon^2) \leq n\Gamma^2\epsilon_0^2/\sigma^2. \tag{5.7}$$

We are now ready to apply (5.4). Let us define the following two quantities

$$\eta(n) := c\ \sigma^{8/(d+3)}\Gamma^{2(d-1)/(d+3)}n^{-4/(d+3)} \text{ and } \alpha(n) := \left(\frac{\Gamma\sqrt{n}}{\sigma}\right)^{(d-1)/(d+3)},$$

where c is a positive constant that depends on d alone and will be specified shortly. Also let $\epsilon^2(n) = \exp(\alpha^2(n)) - 1$. By (5.5) and (5.7), we have

$$\log N(\eta) \geq c'c^{-(d-1)/4}\alpha^2(n) \text{ and } \log M_2(\epsilon; \Theta) \leq c''\alpha^2(n),$$

provided

$$\eta(n) \leq \Gamma^2\delta_0^2/2 \text{ and } \alpha^2(n) \leq n\Gamma^2\epsilon_0^2/\sigma^2. \tag{5.8}$$

Inequality (5.4) with $\eta = \eta(n)$ and $\epsilon = \epsilon(n)$ gives the following lower bound for $R(n)$:

$$\frac{\eta(n)}{2}\left[1 - \exp\left(-\alpha^2(n)c'c^{-(d-1)/4}\right) - \exp\left(\frac{\alpha^2(n)}{2}(1 + c'' - c'c^{-(d-1)/4})\right)\right].$$

If we choose c so that $c'c^{-(d-1)/4} = 2(1 + c'')$, then

$$R(n) \geq \frac{\eta(n)}{2}\left(1 - \exp\left(-\frac{1+c''}{2}\alpha^2(n)\right)\right).$$

If the condition $(1+c'')\alpha^2(n) \geq 2\log 4$ holds, then the above inequality implies $R(n) \geq \eta(n)/4$. This condition as well as (5.8) hold provided $n \geq C(\sigma/\Gamma)^2$ for a large enough C. □

Remark 5.3.1. *In the above proof, our assumptions about the design unit vectors $u_1, \dots, u_n$ were only used via*

$$D_2(P_K||P_{K'}) \leq \exp\left(\frac{n\ell_H^2(K,K')}{\sigma^2}\right) - 1.$$

This inequality is actually true for every joint distribution of $(u_1, \dots, u_n)$ as long as they are independent of the errors $\xi_1, \dots, \xi_n$. Consequently, $c\, n^{-4/(d+3)}$ is a lower bound for the minimax risk for any arbitrary choice of the design unit vectors provided they are independent of $\xi_1, \dots, \xi_n$.

5.4 Upper Bound

The following theorem shows that $R(n)$ is at most $n^{-4/(d+3)}$ up to a multiplicative constant that depends only on d, σ and Γ.

Theorem 5.4.1. *There exist two positive constants c and C depending only on d (and independent of n, σ and Γ) such that*

$$R(n) \leq \frac{c(\Gamma^2/\sigma^2)}{1-e^{-\Gamma^2/(2\sigma^2)}}\sigma^{8/(d+3)}\Gamma^{2(d-1)/(d+3)}n^{-4/(d+3)} \qquad \textit{if } n \geq C(\sigma/\Gamma)^2. \tag{5.9}$$

For each finite subset F of $\mathcal{K}^d(\Gamma)$, let us define the least squares estimator $\hat{K}_F$ by

$$\hat{K}_F := \operatorname*{argmin}_{L\in F} \sum_{i=1}^{n} \left(Y_i - h_L(u_i)\right)^2.$$

We shall show that, if F is chosen appropriately, then $\sup_{K\in\mathcal{K}^d(\Gamma)} \mathbb{E}_K \ell^2(K, \hat{K}_F)$ is bounded from above by the right hand side of (5.9).

Our proof is based on a general estimation result described next. This general result is an adaptation of a technique due to Li (1999) and Barron, Li, Huang, and Luo (2008) for obtaining risk bounds for penalized likelihood estimators.

5.4.1 A general estimation result

Consider an estimation problem in which we want to estimate $\theta \in \Theta$, under a loss function L, based on an observation X whose distribution P_θ depends on the unknown θ. We assume that P_θ has a density p_θ with respect to a common dominating measure μ.

Let $\hat{\theta}(X) := \arg\max_{\theta' \in F} p_{\theta'}(X)$ denote the maximum likelihood estimator over a finite subset F of Θ. The following method of obtaining an upper bound is based on an idea of Li (1999) and Barron et al. (2008). For every $\theta \in \Theta$, $\theta^* \in F$ and $\alpha > 0$, we can write

$$\begin{aligned} L(\theta, \hat{\theta}(X)) &= \log\left(e^{L(\theta,\hat{\theta}(X))}\right) \\ &\leq \log\left(e^{L(\theta,\hat{\theta}(X))}\left(\frac{p_{\hat{\theta}(X)}(X)}{p_{\theta^*}(X)}\right)^{\alpha}\right) \\ &= \log\left(e^{L(\theta,\hat{\theta}(X))}\left(\frac{p_{\hat{\theta}(X)}(X)}{p_{\theta}(X)}\right)^{\alpha}\right) + \alpha \log\left(\frac{p_\theta(X)}{p_{\theta^*}(X)}\right) \end{aligned}$$

Taking expectation with respect to X under the probability measure P_θ on both sides and using Jensen's inequality, we obtain

$$\begin{aligned} \mathbb{E}_\theta L\left(\theta, \hat{\theta}(X)\right) &\leq \log \mathbb{E}_\theta\left(e^{L(\theta,\hat{\theta}(X))}\left(\frac{p_{\hat{\theta}(X)}(X)}{p_{\theta}(X)}\right)^{\alpha}\right) + \alpha D_1(P_\theta || P_{\theta^*}) \\ &\leq \log\left[\sum_{\theta' \in F} e^{L(\theta,\theta')} \mathbb{E}_\theta \left(\frac{p_{\theta'}(X)}{p_\theta(X)}\right)^{\alpha}\right] + \alpha D_1(P_\theta || P_{\theta^*}), \end{aligned}$$

where $D_1(P_\theta || P_{\theta^*})$ denotes the Kullback-Leibler divergence between P_θ and P_{θ^*}. Since this is true for any arbitrary $\theta^* \in F$, we get that

$$\mathbb{E}_\theta L\left(\theta, \hat{\theta}(X)\right) \leq \log \left[\sum_{\theta' \in F} e^{L(\theta,\theta')} \mathbb{E}_\theta \left(\frac{p_{\theta'}(X)}{p_\theta(X)}\right)^\alpha\right] + \alpha \min_{\theta^* \in F} D_1(P_\theta || P_{\theta^*}).$$

In particular, for the following choice of the loss function L,

$$L(\theta, \theta') := -\log \mathbb{E}_\theta \left(\frac{p_{\theta'}(X)}{p_\theta(X)}\right)^\alpha, \tag{5.10}$$

we would obtain

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \hat{\theta}(X)) \leq \log |F| + \alpha \sup_{\theta \in \Theta} \min_{\theta^* \in F} D_1(P_\theta || P_{\theta^*}). \tag{5.11}$$

Note that for $\alpha = 1/2$, the loss function (5.10) is known as the Bhattacharyya divergence (see Bhattacharyya, 1943).

5.4.2 Application of the general result

We apply inequality (5.11) to our problem with $\Theta = \mathcal{K}^d(\Gamma)$ and P_K, the joint distribution of $(u_1, Y_1), \ldots, (u_n, Y_n)$. Also, let p_K denote the density of P_K with respect to the dominating measure $(\nu \otimes \text{Leb})^n$ where Leb denotes Lebesgue measure on the real line. It can be easily checked that (X below stands for the observation vector comprising of $u_i, Y_i, i = 1, \ldots, n$) for $K, K' \in \mathcal{K}^d(\Gamma)$, we have

$$\mathbb{E}_K \left(\frac{p_{K'}(X)}{p_K(X)}\right)^\alpha = \left(\int \exp\left(-\frac{\alpha(1-\alpha)}{2\sigma^2} \left(h_K(u) - h_{K'}(u)\right)^2\right) d\nu(u)\right)^n \tag{5.12}$$

and

$$D_1(P_K||P_{K'}) = \frac{n}{2\sigma^2}\int \left(h_K(u) - h_{K'}(u)\right)^2 d\nu(u) = \frac{n}{2\sigma^2}\ell^2(K, K'). \tag{5.13}$$

Therefore, inequality (5.11) implies that the risk

$$\mathbb{E}_K\left[-\log\left(\int \exp\left(-\frac{\alpha(1-\alpha)}{2\sigma^2}\left(h_K(u) - h_{\hat{K}_F}(u)\right)^2\right) d\nu(u)\right)\right] \tag{5.14}$$

of $\hat{K}_F$ is bounded from above by

$$\frac{\log|F|}{n} + \frac{\alpha}{2\sigma^2}\min_{K' \in F}\ell^2(K, K').$$

Because $-\log x \geq 1 - x$, the above upper bound also holds for the risk when the loss function is taken to be the power divergence $D_\alpha(P_{K'}||P_K)$, for $\alpha \in (0,1)$:

$$D_\alpha(P_{K'}||P_K) := \int \left(1 - \exp\left(-\frac{\alpha(1-\alpha)}{2\sigma^2}\left(h_K(u) - h_{K'}(u)\right)^2\right)\right) d\nu(u).$$

For $K, K' \in \mathcal{K}^d(\Gamma)$, the loss function $\ell^2(K, K')$ can be bounded from above by a multiple of $D_\alpha(K, K')$ for $\alpha \in (0,1)$. Indeed, for $K, K' \in \mathcal{K}^d(\Gamma)$, we have

$$\frac{\alpha(1-\alpha)\left(h_{K'}(u) - h_K(u)\right)^2}{2\sigma^2} \leq \frac{2\alpha(1-\alpha)\Gamma^2}{\sigma^2}$$

and since the convex function $x \mapsto e^{-x}$ lies below the chord joining the points $(0,1)$ and $(2\alpha(1-\alpha)\Gamma^2/\sigma^2, \exp(-2\alpha(1-\alpha)\Gamma^2/\sigma^2))$, it can be checked that

$$\ell^2(K, K') \leq \frac{4\Gamma^2}{1 - \exp(-2\alpha(1-\alpha)\Gamma^2/\sigma^2)} D_\alpha(K, K').$$

We have therefore shown that

$$\mathbb{E}_K \ell^2(K, \hat{K}_F) \le \frac{4\Gamma^2/\sigma^2}{1 - \exp(-2\alpha(1-\alpha)\Gamma^2/\sigma^2)} \left[\frac{\sigma^2}{n} \log |F| + \frac{\alpha}{2} \min_{K' \in F} \ell^2(K, K') \right]. \tag{5.15}$$

According to Bronshtein (1976, Theorem 3 and Remark 1), there exist positive constants c' and ϵ_0 depending only on d and a finite subset $F \subseteq \mathcal{K}^d(\Gamma)$ such that

$$\log |F| \le c' \left(\frac{\Gamma}{\epsilon} \right)^{(d-1)/2} \quad \text{and} \quad \sup_{K \in \mathcal{K}^d(\Gamma)} \min_{K' \in F} \ell^2(K, K') \le \epsilon^2$$

whenever $\epsilon \le \Gamma \epsilon_0$. With this choice of F and $\alpha = 1/2$, inequality (5.15) gives

$$\mathbb{E}_K \ell^2(K, \hat{K}_F) \le \frac{4\Gamma^2/\sigma^2}{1 - \exp(-2\alpha(1-\alpha)\Gamma^2/\sigma^2)} \left[\frac{c'\sigma^2}{n} \left(\frac{\Gamma}{\epsilon} \right)^{(d-1)/2} + \frac{\epsilon^2}{4} \right], \tag{5.16}$$

for every $\epsilon \le \Gamma \epsilon_0$. If we now choose

$$\epsilon := \sigma^{4/(d+3)} \Gamma^{(d-1)/(d+3)} n^{-2/(d+3)},$$

then $\epsilon \le \Gamma \epsilon_0$ provided $n \ge C(\sigma/\Gamma)^2$ for a large enough constant C depending only on d and the required inequality (5.9) follows from (5.16).

5.5 Appendix: A Packing Number Bound

In this section, we prove that the η-packing number $\tilde{N}(\delta; \ell)$ of $\mathcal{K}^d(\Gamma)$ under the ℓ metric is at least $\exp(c(\Gamma/\delta)^{(d-1)/2})$ for a positive c and sufficiently small η. This result was needed in the proof of our minimax lower bound. Bronshtein (1976, Theorem 4 and Remark 1) proved this for the Haussdorff metric ℓ_H which is larger than ℓ.

Theorem 5.5.1. *There exist positive constants δ_0 and c depending only on d such that*

$$\tilde{N}(\delta; \ell) \geq \exp\left(c\left(\frac{\Gamma}{\delta}\right)^{(d-1)/2}\right) \qquad \textit{whenever } \eta \leq \Gamma\eta_0.$$

The following lemma will be used in the proof of the above theorem. Let B denote the unit ball in $\mathbb{R}^d$.

Lemma 5.5.2. *For a fixed $0 < \eta \leq 1/8$ and a unit vector v, consider the following two subsets of the unit ball B:*

$$D(1) := B \textit{ and } D(0) := B \cap \{x : \langle x, v\rangle \leq 1 - \eta\}.$$

Then $\ell^2(D(0), D(1)) \geq c\eta^{(d+3)/2}$ for a positive constant c that depends only on d.

We first provide the proof of Theorem 5.5.1 using the above lemma, which will be proved subsequently.

Proof of Theorem 5.5.1. We observe that, by scaling, it is enough to prove for $\Gamma = 1$. We loosely follow Bronshtein (1976, Proof of Theorem 4). We fix $0 < \eta \leq 1/8$ and let $v_1, \ldots, v_m$ be unit vectors such that the Euclidean distance between v_i and v_j is at least $2\sqrt{2\eta}$ for $i \neq j$. Since the ϵ-packing number of the unit sphere under the Euclidean metric is $\geq c\epsilon^{1-d}$ for $0 < \epsilon < 1$, we assume that $m \geq c_1\eta^{(1-d)/2}$ for a positive constant c_1 that depends only on d.

For each $\tau \in \{0, 1\}^m$, we define the compact, convex set

$$K(\tau) := D_1(\tau_1) \cap \cdots \cap D_m(\tau_m)$$

where $D_j(\tau_j)$ equals $B \cap \{x : \langle x, v_j\rangle \leq 1 - \eta\}$ when $\tau_j = 0$ and B when $\tau_j = 1$, where B denotes the unit ball in $\mathbb{R}^d$. By the choice of $v_1, \ldots, v_m$, it follows that the sets

$B \cap \{x : \langle x, v_j \rangle > 1 - \eta\}$ are disjoint. As a result, we have

$$\ell^2(K(\tau), K(\tau')) = \sum_{i:\tau_i \neq \tau_i'} \ell^2(D_j(0), D_j(1)) = \Upsilon(\tau, \tau')\ell^2(D_1(0), D_1(1)),$$

for every $\tau, \tau' \in \{0,1\}^m$ where $\Upsilon(\tau, \tau') := \sum_i \{\tau_i \neq \tau_i'\}$ denotes the Hamming distance between τ and τ'. By Lemma 5.5.2, we get $\ell^2(K(\tau), K(\tau')) \geq c_2 \Upsilon(\tau, \tau')\eta^{(d+3)/2}$ where c_2 depends on d alone.

We recall the Varshamov-Gilbert lemma used in the previous chapter to assert the existence of a subset W of $\{0,1\}^m$ with $|W| \geq \exp(m/8)$ such that $\Upsilon(\tau, \tau') = \sum_i \{\tau_i \neq \tau_i'\} \geq m/4$ for all $\tau, \tau' \in W$ with $\tau \neq \tau'$.

Therefore, for every $\tau, \tau' \in W$ with $\tau \neq \tau'$, we get (note that $m \geq c_1 \eta^{(1-d)/2}$)

$$\ell^2(K(\tau), K(\tau')) \geq \frac{c_2}{4} m \eta^{(d+3)/2} \geq \frac{c_1 c_2}{4} \eta^2.$$

Taking $\delta := \eta\sqrt{c_1 c_2 / 4}$, we see that, whenever $\delta \leq \sqrt{c_1 c_2}/16$, $\{K(\tau), \tau \in W\}$ is a δ-packing subset of $\mathcal{K}^d(\Gamma)$ in the ℓ^2-metric of size M where

$$\log M \geq \frac{m}{8} \geq \frac{c_1}{8}\eta^{(1-d)/2} \geq c\delta^{(1-d)/2} \text{ with } c := \frac{c_1}{8}\left(\frac{2}{\sqrt{c_1 c_2}}\right)^{(1-d)/2}.$$

The proof is complete. □

For the proof of Lemma 5.5.2, we recall an elementary fact about spherical caps. For a unit vector x and a real number $0 < \delta < 1$, consider the spherical cap $\mathcal{S}(x; \delta)$ centered at x of radius δ consisting of all unit vectors whose Euclidean distance to x is at most δ. It can be checked that this spherical cap consists of precisely those unit vectors which form an angle of at most α with the vector x, where α is related to δ through

$$\cos\alpha = 1 - \frac{\delta^2}{2} \text{ and } \sin\alpha = \frac{\delta\sqrt{4-\delta^2}}{2}.$$

A standard result is that $\nu(\mathcal{S}(x;\delta))$ equals $c\int_0^\alpha \sin^{d-2} t\, dt$ where the constant c only depends on d. This integral can be bounded from below in the following simple way:

$$\int_0^\alpha \sin^{d-2} t\, dt \geq \int_0^\alpha \sin^{d-2} t \cos t\, dt \geq \frac{\sin^{d-1}\alpha}{d-1},$$

and for an upper bound, we note

$$\int_0^\alpha \sin^{d-2} t\, dt \leq \int_0^\alpha \frac{\cos t}{\cos\alpha} \sin^{d-2} t\, dt \leq \frac{\sin^{d-1}\alpha}{(d-1)\cos\alpha}.$$

We thus have $c_1 \sin^{d-1}\alpha \leq \nu(\mathcal{S}(x;\delta)) \leq c_2 \sin^{d-1}\alpha/\cos\alpha$ for constants c_1 and c_2 depending on d alone. Writing $\cos\alpha$ and $\sin\alpha$ in terms of δ and using the assumption that $0 < \delta \leq 1$, we obtain that

$$C_1\delta^{d-1} \leq \nu(\mathcal{S}(x;\delta)) \leq C_2\delta^{d-1}, \tag{5.17}$$

for positive constants C_1 and C_2 depending only on d.

Proof of Lemma 5.5.2. It can be checked that the support functions of $D(0)$ and $D(1)$ differ only for unit vectors in the spherical cap $\mathcal{S}(v, \sqrt{2\eta})$. This spherical cap consists of all unit vectors which form an angle of at most α with v where $\cos\alpha = 1-\eta$. In fact, if θ denotes the angle between an arbitrary unit vector u and v, it can be verified by elementary trigonometry that

$$h_{D(0)}(u) - h_{D(1)}(u) = \begin{cases} (1-\cos(\alpha-\theta)) & \text{if } 0 \leq \theta \leq \alpha, \\ 0 & \text{otherwise.} \end{cases} \tag{5.18}$$

For a fixed $0 < b \leq 1$, let $0 \leq \beta \leq \alpha$ denote the angle for which $1 - \cos(\alpha - \beta) = b\eta$. It follows from (5.18) that the difference in the support functions of $D(0)$ and $D(1)$ is at least $b\eta$ for all unit vectors in the spherical cap consisting of all unit vectors forming an angle of at most β with v. This spherical cap can be denoted by $\mathcal{S}(v, t)$ where t is given by $t^2 := 2(1 - \cos\beta)$. Therefore $\ell^2(D(0), D(1)) \geq b^2\eta^2\nu(\mathcal{S}(v, t))$. It is easy to check that $t^2 \leq 2(1 - \cos\alpha) \leq 2\eta$. Also, $t \geq \sin\beta$ and $\sin\beta$ can be bounded from below in the following way

$$1 - b\eta = \cos(\alpha - \beta) \leq \cos\alpha + \sin\alpha \sin\beta \leq 1 - \eta + \sqrt{2\eta}\sin\beta.$$

Thus $t \geq \sin\beta \geq (1 - b)\sqrt{\eta/2}$ and from (5.17), it follows that

$$\ell^2(D(0), D(1)) \geq c\eta^2 b^2 t^{d-1} \geq cb^2(1 - b)^{d-1}\eta^{(d+3)/2}$$

for all $0 < b \leq 1$. Choosing $b = 1/2$ will yield $\ell^2(D(0), D(1)) \geq c\eta^{(d+3)/2}$. □

Bibliography

Barron, A. R., J. Q. Li, C. Huang, and X. Luo (2008). The MDL principle, penalized likelihood, and statistical risk. In P. Grünwald, P. Myllymäki, I. Tabus, M. Weinberger, and B. Yu (Eds.), *Festschrift for Jorma Rissanen*. Tampere, Finland.

Bhattacharyya (1943). On a measure of divergence between two statistical populations defined by probability distributions. *Bulletin of the Calcutta Mathematical Society 35*, 99–109.

Bronshtein, E. M. (1976). ϵ-entropy of convex sets and functions. *Siberian Math. J. 17*, 393–398.

Fisher, N. I., P. Hall, B. A. Turlach, and G. S. Watson (1997). On the estimation of a convex set from noisy data on its support function. *Journal of the American Statistical Association 92*, 84–91.

Gardner, R. J. and M. Kiderlen (2009). A new algorithm for 3D reconstruction from support functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 556–562.

Gardner, R. J., M. Kiderlen, and P. Milanfar (2006). Convergence of algorithms for reconstructing convex bodies and directional measures. *Annals of Statistics 34*, 1331–1374.

Gregor, J. and F. R. Rannou (2002). Three-dimensional support function estimation and application for projection magnetic resonance imaging. *International Journal of Imaging Systems and Technology 12*, 43–50.

Hall, P. and B. A. Turlach (1999). On the estimation of a convex set with corners. *IEEE Transactions on Pattern Analysis and Machine Intelligence 21*, 225–234.

Lele, A. S., S. R. Kulkarni, and A. S. Willsky (1992). Convex-polygon estimation from support-line measurements and applications to target reconstruction from laser-radar data. *Journal of the Optical Society of America, Series A 9*, 1693–1714.

Li, J. Q. (1999). *Estimation of Mixture Models*. Ph. D. thesis, Yale University.

Massart, P. (2007). *Concentration inequalities and model selection. Lecture notes in Mathematics*, Volume 1896. Berlin: Springer.

Prince, J. L. and A. S. Willsky (1990). Reconstructing convex sets from support line measurements. *IEEE Transactions on Pattern Analysis and Machine Intelligence 12*, 377–389.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press.

Schneider, R. (1993). *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge: Cambridge Univ. Press.