Lower bounds for the minimax risk using f-divergences, and applications

Adityanand Guntuboyina

***Abstract*—Lower bounds involving f-divergences between the underlying probability measures are proved for the minimax risk in estimation problems. Our proofs just use simple convexity facts. Special cases and straightforward corollaries of our bounds include well known inequalities for establishing minimax lower bounds such as Fano's inequality, Pinsker's inequality and inequalities based on global entropy conditions. Two applications are provided: a new minimax lower bound for the reconstruction of convex bodies from noisy support function measurements and a different proof of a recent minimax lower bound for the estimation of a covariance matrix.**

***Index Terms*—f-divergences; Fano's inequality; Minimax lower bounds; Pinsker's inequality; Reconstruction from support functions.**

I. Introduction

Consider an estimation problem in which we want to estimate $\theta \in \Theta$ based on an observation X from $\{P_\theta, \theta \in \Theta\}$ where each P_θ is a probability measure on a sample space $\mathcal{X}$. Suppose that estimators are allowed to take values in $\mathcal{A} \supseteq \Theta$ and that the loss function is of the form $\ell(\rho)$ where ρ is a metric on $\mathcal{A}$ and $\ell : [0, \infty) \to [0, \infty)$ is a nondecreasing function. The minimax risk for this problem is defined by

$$R := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \ell(\rho(\theta, \hat{\theta}(X))),$$

where the infimum is over all measurable functions $\hat{\theta} : \mathcal{X} \to \mathcal{A}$ and the expectation is taken under the assumption that X is distributed according to P_θ.

In this article, we are concerned with the problem of obtaining lower bounds for the minimax risk R. Such bounds are useful in assessing the quality of estimators for θ. The standard approach to these bounds is to obtain a reduction to the more tractable problem of bounding from below the minimax risk of a multiple hypothesis testing problem. More specifically, one considers a finite subset F of the parameter space Θ and a real number η such that $\rho(\theta, \theta') \geq \eta$ for $\theta, \theta' \in F, \theta \neq \theta'$ and employs the inequality $R \geq \ell(\eta/2) r$, where

$$r := \inf_T \sup_{\theta \in F} P_\theta \{T \neq \theta\}, \tag{1}$$

the infimum being over all estimators T taking values in F. The proof of this inequality relies on the triangle inequality satisfied by the metric ρ and can be found, for example, in [1, Page 1570, Proof of Theorem 1] (Let us note, for the convenience of the reader, that the notation employed by Yang and Barron [1] differs from ours in that they use d for the metric ρ, $\epsilon_{n,d}$ for our η and $N_{\epsilon_n,d}$ for the finite set F. Also the proof in [1] involves a positive constant A which can be taken to be 1 for our purposes. The constant A arises because Yang and Barron [1] do not require that d is a metric but rather require it to satisfy a weaker local triangle inequality which involves the constant A.)

The next step is to note that r is bounded from below by Bayes risks. Let w be a probability measure on F. The Bayes risk $\bar{r}_w$ corresponding to the prior w is defined by

$$\bar{r}_w := \inf_T \sum_{\theta \in F} w_\theta P_\theta \{T \neq \theta\}, \tag{2}$$

where $w_\theta := w\{\theta\}$ and the infimum is over all estimators T taking values in F. When w is the discrete uniform probability measure on F, we simply write $\bar{r}$ for $\bar{r}_w$. The trivial inequality $r \geq \bar{r}_w$ implies that lower bounds for $\bar{r}_w$ are automatically lower bounds for r.

The starting point for the results described in this paper is Theorem II.1, which provides a lower bound for $\bar{r}_w$ involving f-divergences of the probability measures $P_\theta, \theta \in F$. The f-divergences ([2]–[5]) are a general class of divergences between probability measures which include many common divergences/distances like the Kullback Leibler divergence, chi-squared divergence, total variation distance, Hellinger distance etc. For a *convex* function $f : [0, \infty) \to \mathbb{R}$ satisfying $f(1) = 0$, the f-divergence between two probabilities P and Q is given by

$$D_f(P||Q) := \int f\left(\frac{dP}{dQ}\right) dQ$$

if P is absolutely continuous with respect to Q and ∞ otherwise.

Our proof of Theorem II.1 presented in section II is extremely simple. It just relies on the convexity of the function f and the standard result that $\bar{r}_w$ has the following exact expression:

$$\bar{r}_w = 1 - \int_{\mathcal{X}} \max_{\theta \in F} \{w_\theta p_\theta(x)\} \, d\mu(x), \tag{3}$$

After acceptance of this manuscript, Professor Alexander Gushchin pointed out that Theorem II.1 appears in his paper [14]. Specifically, in a different notation, inequality (5) appears as Theorem 1 and inequality (4) appears in Section 4.3 in [14]. The proof of Theorem II.1 presented in section II is different from that in [14]. Also, except for Theorem II.1 and the observation that Fano's inequality is a special case of Theorem II.1 (see Example II.4), there is no other overlap between this paper and [14].

Some of the material in this paper was presented at the IEEE International Symposium on Information Theory, Austin, TX, June 2010.

A. Guntuboyina is with the Department of Statistics, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA. e-mail: adityanand.guntuboyina@yale.edu

where p_θ denotes the density of P_θ with respect to a common dominating measure μ (for example, one can take $\mu := \sum_{\theta \in F} P_\theta$).

We show that Fano's inequality is a special case (see Example II.4) of Theorem II.1, obtained by taking $f(x) = x \log x$. Fano's inequality is used extensively in the non-parametric statistics literature for obtaining minimax lower bounds, important works being [1], [6]–[11]. In the special case when F has only two points, Theorem II.1 gives a sharp inequality relating the total variation distance between two probability measures to f-divergences (see Corollary II.3). When $f(x) = x \log x$, Corollary II.3 implies an inequality due to Topsøe [12] from which Pinsker's inequality can be derived. Thus Theorem II.1 can be viewed as a generalization of both Fano's inequality and Pinsker's inequality.

The bound given by Theorem II.1 involves the quantity $J_f := \inf_Q \sum_{\theta \in F} D_f(P_\theta || Q)/|F|$, where the infimum is over all probability measures Q and $|F|$ denotes the cardinality of the finite set F. It is usually not possible to calculate J_f exactly and in section III, we provide upper bounds for J_f. The main result of this section, Theorem III.1, provides an upper bound for J_f based on approximating the set of $|F|$ probability measures $\{P_\theta, \theta \in F\}$ by a smaller set of probability measures. This result is motivated by and a generalization to f-divergences of a result of Yang and Barron [1] for the Kullback-Leibler divergence.

In section IV, we use the inequalities proved in sections II and III to obtain minimax lower bounds involving only global metric entropy attributes. Of all the lower bounds presented in this paper, Theorem IV.1, the main result of section IV, is the most application-ready method. In order to apply this in a particular situation, one only needs to determine suitable bounds on global covering and packing numbers of the parameter space Θ and the space of probability measures $\{P_\theta, \theta \in \Theta\}$ (see section V for an application).

Although the main results of sections II and III hold true for all f-divergences, Theorem IV.1 is stated only for the Kullback-Leibler divergence, chi-squared divergence and the divergences based on $f(x) = x^l - 1$ for $l > 1$. The reason behind this is that Theorem IV.1 is intended for applications where it is usually the case that the underlying probability measures P_θ are product measures and divergences such as the Kullback-Leibler divergence and chi-squared divergence can be computed for product probability measures.

The inequalities given by Theorem IV.1 for the chi-squared divergence and divergences based on $f(x) = x^l - 1$ for $l > 1$ are new while the inequality for the Kullback-Leibler divergence is due to Yang and Barron [1]. There turn out to be qualitative differences between these inequalities in the case of estimation problems involving finite dimensional parameters where the inequality based on chi-squared divergence gives minimax lower bounds having the optimal rate while the one based on the Kullback-Leibler divergence only results in sub-optimal lower bounds. We shall explain this happening in section IV by means of elementary examples.

We shall present two applications of our bounds. In section V, we shall prove a new lower bound for the minimax risk in the problem of estimation/reconstruction of a d-dimensional convex body from noisy measurements of its support function in n directions. In section VI, we shall provide a different proof of a recent result by Cai, Zhang and Zhou [13] on covariance matrix estimation.

II. Lower bounds for the testing risk $\bar{r}_w$

We shall prove a lower bound for $\bar{r}_w$ defined in (2) in terms of f-divergences. We shall assume that the $N := |F|$ probability measures $P_\theta, \theta \in F$ are all dominated by a sigma finite measure μ with densities $p_\theta, \theta \in F$. In terms of these densities, $\bar{r}_w$ has the exact expression given in (3). A trivial consequence of (3) that we shall often use in the sequel is that $\bar{r} \leq 1 - 1/N$ (recall that $\bar{r}$ is $\bar{r}_w$ in the case when w is the uniform probability measure on F).

Theorem II.1. *Let w be a probability measure on F. Define $T : \mathcal{X} \to F$ by $T(x) := \arg\max_{\theta \in F} \{w_\theta p_\theta(x)\}$, where $w_\theta := w\{\theta\}$. For every convex function $f : [0, \infty) \to \mathbb{R}$ and every probability measure Q on $\mathcal{X}$, we have*

$$\sum_{\theta \in F} w_\theta D_f(P_\theta || Q) \geq W f\left(\frac{1 - \bar{r}_w}{W}\right) + (1 - W) f\left(\frac{\bar{r}_w}{1 - W}\right), \tag{4}$$

where $W := \int_{\mathcal{X}} w_{T(x)} dQ(x)$. In particular, taking w to be the uniform probability measure, we get that

$$\sum_{\theta \in F} D_f(P_\theta || Q) \geq f\left(N(1 - \bar{r})\right) + (N - 1) f\left(\frac{N\bar{r}}{N - 1}\right). \tag{5}$$

The proof of this theorem relies on a simple application of the convexity of f and it is presented below.

Proof: We may assume that all the weights w_θ are strictly positive and that the probability measure Q has a density q with respect to μ. We start with a simple inequality for non-negative numbers $a_\theta, \theta \in F$ with $\tau := \arg\max_{\theta \in F} \{w_\theta a_\theta\}$. We first write

$$\sum_{\theta \in F} w_\theta f(a_\theta) = w_\tau f(a_\tau) + (1 - w_\tau) \sum_{\theta \neq \tau} \frac{w_\theta}{1 - w_\tau} f(a_\theta)$$

and then use the convexity of f to obtain that the quantity $\sum_\theta w_\theta f(a_\theta)$ is bounded from below by

$$w_\tau f(a_\tau) + (1 - w_\tau) f\left(\frac{\sum_{\theta \in F} w_\theta a_\theta - w_\tau a_\tau}{1 - w_\tau}\right).$$

We now fix $x \in \mathcal{X}$ such that $q(x) > 0$ and apply the inequality just derived to $a_\theta := p_\theta(x)/q(x)$. Note that in this case $\tau = T(x)$. We get that

$$\sum_{\theta \in F} w_\theta f\left(\frac{p_\theta(x)}{q(x)}\right) \geq A(x) + B(x), \tag{6}$$

where

$$A(x) := w_{T(x)} f\left(\frac{p_{T(x)}(x)}{q(x)}\right)$$

and

$$B(x) := (1 - w_{T(x)}) f\left(\frac{\sum_{\theta \in F} w_\theta p_\theta(x) - w_{T(x)} p_{T(x)}(x)}{(1 - w_{T(x)}) q(x)}\right).$$

Integrating inequality (6) with respect to the probability measure Q, we get that the term $\sum_{\theta \in F} w_\theta D_f(P_\theta || Q)$ is bounded from below by

$$\int_{\mathcal{X}} A(x) q(x) d\mu(x) + \int_{\mathcal{X}} B(x) q(x) d\mu(x).$$

Let Q' be the probability measure on $\mathcal{X}$ having the density $q'(x) := w_{T(x)} q(x)/W$ with respect to μ. Clearly

$$\int_{\mathcal{X}} A(x) q(x) d\mu(x) = W \int_{\mathcal{X}} f\left(\frac{p_{T(x)}(x)}{q(x)}\right) q'(x) d\mu(x),$$

which, by Jensen's inequality, is larger than or equal to $Wf((1-\bar{r}_w)/W)$. It follows similarly that

$$\int_{\mathcal{X}} B(x) q(x) d\mu(x) \geq (1-W) f\left(\frac{\bar{r}_w}{1-W}\right).$$

This completes the proof of inequality (4). When w is the uniform probability measure on the finite set F, it is obvious that W equals $1/N$ and this leads to inequality (5). ■

Let us denote the function of $\bar{r}$ on the right hand side of (5) by g:

$$g(a) := f(N(1-a)) + (N-1) f\left(\frac{Na}{N-1}\right). \quad (7)$$

Inequality (5) provides an implicit lower bound for $\bar{r}$. This is because $\bar{r} \in [0, 1-1/N]$ and g is non-increasing on $[0, 1-1/N]$ (as can be seen in the proof of the next corollary in the case when f is differentiable; if f is not differentiable, one needs to work with right and left derivatives which exist for convex functions).

The convexity of f also implies trivially that g is convex, which can be used to convert the implicit bound (5) into an explicit lower bound. This is the content of the following corollary. We assume differentiability for convenience; to avoid working with one-sided derivatives.

Corollary II.2. *Suppose that $f : [0, \infty)$ is a differentiable convex function and that g is defined as in* (7). *Then, for every $a \in [0, 1-1/N]$, we have*

$$r \geq \bar{r} \geq a + \frac{\inf_Q \sum_{\theta \in F} D_f(P_\theta || Q) - g(a)}{g'(a)}, \quad (8)$$

where the infimum is over all probability measures Q.

Proof: Fix a probability measure Q. Inequality (5) says that $\sum_{\theta \in F} D_f(P_\theta || Q) \geq g(\bar{r})$. The convexity of f implies that g is also convex and hence, for every $a \in [0, 1-1/N]$, we can write

$$\sum_{\theta \in F} D_f(P_\theta || Q) \geq g(\bar{r}) \geq g(a) + g'(a)(\bar{r} - a). \quad (9)$$

Also,

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

Because g is convex, we have $g'(a) \leq g'(1-1/N) = 0$ for $a \leq 1-1/N$ (this proves that g is non-increasing on $[0, 1-1/N]$). Therefore, by rearranging (9), we obtain (8). ■

Let us now provide an intuitive understanding of inequality (5). When the probability measures $P_\theta, \theta \in F$ are tightly packed i.e., when they are close to one another, it is hard to distinguish between them (based on the observation X) and hence, the testing Bayes risk $\bar{r}$ will be large. On the other hand, when the probability measures are well spread out, it is easy to distiguish between them and therefore, $\bar{r}$ will be small. Indeed, $\bar{r}$ takes on its maximum value of $1 - 1/N$ when the probability measures $P_\theta, \theta \in F$ are all equal to one another and it takes on its smallest value of 0 when $\max p_\theta = \sum p_\theta$ i.e., when $P_\theta, \theta \in F$ are all mutually singular.

Now, one way of measuring how packed/spread out the probability measures $P_\theta, \theta \in F$ are is to consider the quantity $\inf_Q \sum_{\theta \in F} D_f(P_\theta || Q)$, which is small when the probabilities are tightly packed and large when they are spread out. It is therefore reasonable to expect a connection between this quantity and $\bar{r}$. Inequality (5) makes this connection explicit and precise. The fact that the function g in (7) is non-increasing means that when $\inf_Q \sum_{\theta \in F} D_f(P_\theta || Q)$ is small, the lower bound on $\bar{r}$ implied by (5) is large and when $\inf_Q \sum_{\theta \in F} D_f(P_\theta || Q)$ is large, the lower bound on $\bar{r}$ is small.

Theorem II.1 implies the following corollary which provides sharp inequalities between total variation distance and f-divergences. The total variation distance between two probability measures is defined as *half* the L^1 distance between their densities.

Corollary II.3. *Let P_1 and P_2 be two probability measures on a space $\mathcal{X}$ with total variation distance V. For every convex function $f : [0, \infty) \to \mathbb{R}$, we have*

$$\inf_Q (D_f(P_1 || Q) + D_f(P_2 || Q)) \geq f(1+V) + f(1-V), \quad (10)$$

where the infimum is over all probability measures Q. Moreover this inequality is sharp in the sense that for every $V \in [0, 1]$, the infimum of the left hand side of (10) *over all probability measures P_1 and P_2 with total variation distance V equals the right hand side of* (10).

Proof: In the setting of Theorem II.1, suppose that $F = \{1, 2\}$ and that the two probability measures are P_1 and P_2 with densities p_1 and p_2 respectively. Since $2\max(p_1, p_2)$ equals $p_1 + p_2 + |p_1 - p_2|$, it follows that $2\bar{r}$ equal $1 - V$. Inequality (10) is then a direct consequence of inequality (5).

The following example shows that (10) is sharp. Fix $V \in [0, 1]$. Consider the space $\mathcal{X} = \{1, 2\}$ and define the probabilities P_1 and P_2 by $P_1\{1\} = P_2\{2\} = (1+V)/2$ and of course $P_1\{2\} = P_2\{1\} = (1-V)/2$. Then the total variation distance between P_1 and P_2 equals V. Also if we take Q to be the uniform probability measure $Q\{1\} = Q\{2\} = 1/2$, then one sees that $D_f(P_1 || Q) + D_f(P_2 || Q)$ equals $f(1+V) + f(1-V)$ which is same as the right hand side in (10). ■

What we have actually shown in the above proof is that inequality (10) is sharp for the space $\mathcal{X} = \{1, 2\}$. However, the result holds in more general spaces as well. For example, if the space is such that there exist two disjoint nonempty subsets A_1 and A_2 and two probability measures ν_1 and ν_2 concentrated on A_1 and A_2 respectively, then we can define $P_1 := \nu_1(1+V)/2 + \nu_2(1-V)/2$ and $P_2 := \nu_1(1-V)/2 + \nu_2(1+V)/2$ so that $V(P_1, P_2) = V$ and (10) becomes an equality (with $Q = \nu_1/2 + \nu_2/2$).

There exist many inequalities in the literature relating the f-divergence of two probability measures to their total variation distance. We refer the reader to [15] for the sharpest results in this direction and for earlier references. Inequality (10), which is new, can be trivially converted into an inequality between $D_f(P_1||P_2)$ and V by taking $Q = P_2$. The resulting inequality will not be sharp however and hence will be inferior to the inequalities in [15]. As stated, inequality (10) is a sharp inequality relating not $D_f(P_1||P_2)$ but a symmetrized form of f-divergence between P_1 and P_2 to their total variation distance.

In the remainder of this section, we shall apply Theorem II.1 and Corollary II.3 to specific f-divergences.

Example II.4 (Kullback-Leibler Divergence)**.** Let $f(x) := x \log x$. Then $D_f(P||Q)$ becomes the Kullback-Leibler divergence $D(P||Q)$ between P and Q. The quantity $\sum_{\theta\in F} D(P_\theta||Q)$ is minimized when $Q = \bar{P} := (\sum_{\theta\in F} P_\theta)/N$. This is a consequence of the following identity which is sometimes referred to as the *compensation identity*, see for example [12, Page 1603]:

$$\sum_{\theta\in F} D(P_\theta||Q) = \sum_{\theta\in F} D(P_\theta||\bar{P}) + ND(\bar{P}||Q).$$

Using inequality (5) with $Q = \bar{P} = (\sum_{\theta\in F} P_\theta)/N$, we obtain

$$\frac{1}{N}\sum_{\theta\in F} D(P_\theta||\bar{P}) \geq (1-\bar{r})\log(N(1-\bar{r})) + \bar{r}\log\left(\frac{N\bar{r}}{N-1}\right).$$

The quantity on the left hand side is known as the Jensen-Shannon divergence. It is also Shannon's mutual information [16, 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 above inequality is stronger than the version of Fano's inequality commonly used in nonparametric statistics. It is implicit in [17, Proof of Theorem 1] and is explicitly stated in a slightly different form in [18, Theorem 3]. The proof in [17] is based on the Fano's inequality from information theory [16, Theorem 2.10.1]. To obtain the usual form of Fano's inequality as used in statistics, we turn to inequality (8). For $a_0 := (N-1)/(2N-1) \leq 1 - 1/N$ and the function g in (7), it can be checked that

$$g(a_0) = \frac{N^2}{2N-1}\log N + N\log\left(\frac{N}{2N-1}\right)$$

and $g'(a_0) = -N\log N$. Using inequality (8) with $a = a_0$, we get that

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

Since $\log((2N-1)/N) \leq \log 2$, we have obtained

$$r \geq \bar{r} \geq 1 - \frac{\log 2 + \frac{1}{N}\sum_{\theta\in F} D(P_\theta||\bar{P})}{\log N}, \qquad (11)$$

which is the commonly used version of Fano's inequality.

By taking $f(x) = x\log x$ in Corollary II.3, we get that

$$D(P_1||\bar{P}) + D(P_2||\bar{P}) \geq (1+V)\log(1+V) + (1-V)\log(1-V).$$

This inequality relating the Jensen-Shannon divergence between two probability measures (also known as capacitory discrimination) to their total variation distance is due to Topsøe [12, Equation (24)]. Our proof is slightly simpler than Topsøe's. Topsøe [12] also explains how to use this inequality to deduce Pinsker's inequality with sharp constant: $D(P_1||P_2) \geq 2V^2$. Thus, Theorem II.1 can be considered as a generalization of both Fano's inequality and Pinsker's inequality to f-divergences.

Example II.5 (Chi-Squared Divergence)**.** Let $f(x) = x^2 - 1$. Then $D_f(P||Q)$ becomes the chi-squared divergence $\chi^2(P||Q) := \int p^2/q - 1$. The function g can be easily seen to satisfy

$$g(a) = \frac{N^3}{N-1}\left(1 - \frac{1}{N} - a\right)^2 \geq N^2\left(1 - \frac{1}{N} - a\right)^2.$$

Because $\bar{r} \leq 1 - 1/N$, we can invert the inequality $\inf_Q \sum_{\theta\in F}\chi^2(P_\theta||Q) \geq g(\bar{r})$ to obtain

$$r \geq \bar{r} \geq 1 - \frac{1}{N} - \frac{1}{\sqrt{N}}\sqrt{\frac{\inf_Q \sum_{\theta\in F}\chi^2(P_\theta||Q)}{N}}. \qquad (12)$$

Also it follows from Corollary II.3 that for every two probability measures P_1 and P_2,

$$\inf_Q\left(\chi^2(P_1||Q) + \chi^2(P_2||Q)\right) \geq 2V^2. \qquad (13)$$

The weaker inequality $\chi^2(P_1||\bar{P}) + \chi^2(P_2||\bar{P}) \geq 2V^2$ can be found in [12, Equation (11)].

Example II.6 (Hellinger Distance)**.** Let $f(x) = 1 - \sqrt{x}$. Then $D_f(P||Q) = 1 - \int\sqrt{pq}d\mu = H^2(P,Q)/2$, where $H^2(P,Q) = \int(\sqrt{p} - \sqrt{q})^2 d\mu$ is the square of the Hellinger distance between P and Q. It can be shown, using the Cauchy-Schwarz inequality, that $\sum_{\theta\in F} D_f(P_\theta||Q)$ is minimized when Q has a density with respect to μ that is proportional to $(\sum_{\theta\in F}\sqrt{p_\theta})^2$. Indeed if $u := \sum_{\theta\in F}\sqrt{p_\theta}$, then

$$\sum_{\theta\in F} D_f(P_\theta||Q) = N - \int\sqrt{qu^2}d\mu \geq N - \sqrt{\int u^2 d\mu},$$

by the Cauchy-Schwarz inequality with equality when q is proportional to u^2. The inequality (5) can then be simplified to

$$\sqrt{1-\bar{r}} + \sqrt{(N-1)\bar{r}} \geq \sqrt{\frac{\int u^2 d\mu}{N}}. \qquad (14)$$

We now observe that

$$\int u^2 d\mu = N + \sum_{\theta\neq\theta'}\int\sqrt{p_\theta p_{\theta'}}d\mu = N^2 - \frac{1}{2}\sum_{\theta\neq\theta'} H^2(P_\theta, P_{\theta'}).$$

We let $h^2 := \sum_{\theta,\theta'} H^2(P_\theta, P_{\theta'})/N^2$ so that $\int u^2 d\mu = N^2(1 - h^2/2)$. As a consequence, we have $\int u^2 d\mu \leq N^2$. Also note that $\int u^2 d\mu \geq \int(\sum_\theta p_\theta)d\mu = N$. Therefore, the right hand side of the inequality (14) lies between 1 and $\sqrt{N}$. On the other hand, it can be checked that, as a function of $\bar{r}$, the left hand side of (14) is strictly increasing from 1 at $\bar{r} = 0$ to $\sqrt{N}$ at $\bar{r} = 1 - 1/N$. It therefore follows that inequality (14) is equivalent to $\bar{r} \geq \check{r}$ where $\check{r} \in [0, 1 - 1/N]$ is the solution to the equation obtained by replacing the inequality in (14) with an equality.

This equation can be solved in the usual way by squaring etc., until we get a quadratic equation in $\bar{r}$ which can be solved resulting in two solutions. One of the two solutions can be discarded by continuity considerations (the solution has to be continuous in $\int u^2 d\mu/N$) and the fact that $\bar{r} \leq 1 - 1/N$. The other solution equals $\check{r}$ and is given by

$$\check{r} = 1 - \frac{1}{N} - \frac{N-2}{N}\frac{h^2}{2} - \frac{\sqrt{N-1}}{N}\sqrt{h^2(2-h^2)}.$$

We have thus shown that

$$r \geq \bar{r} \geq 1 - \frac{1}{N} - \frac{N-2}{N}\frac{h^2}{2} - \frac{\sqrt{N-1}}{N}\sqrt{h^2(2-h^2)}.$$

In the case when $N = 2$ and $F = \{1, 2\}$, it is clear that $h^2 = (H^2(P_1, P_2) + H^2(P_2, P_1))/4 = H^2(P_1, P_2)/2$. Also since $2\bar{r}$ equals $1 - V$, where V denotes the total variation distance between P_1 and P_2, the above inequality implies that for every pair of probability measures P_1 and P_2, we have

$$V \leq H(P_1, P_2)\sqrt{1 - \frac{H^2(P_1, P_2)}{4}}.$$

This inequality is usually attributed to Le Cam [19].

Example II.7 (Total Variation Distance)**.** Let $f(x) = |x - 1|/2$. Then $D_f(P||Q)$ becomes the total variation distance between P and Q. The function g satisfies

$$g(\bar{r}) = \frac{1}{2}|N(1-\bar{r}) - 1| + \frac{N-1}{2}\left|\frac{N\bar{r}}{N-1} - 1\right|.$$

Since $\bar{r} \leq 1 - 1/N$, we have $N(1 - \bar{r}) \geq 1$ and $N\bar{r}/(N - 1) \leq 1$ so that the above expression for $g(\bar{r})$ simplifies to $N - 1 - N\bar{r}$. Inequality (5), therefore, results in

$$r \geq \bar{r} \geq 1 - \frac{1}{N} - \frac{\inf_Q \sum_{\theta \in F} V_\theta}{N}.$$

where V_θ denotes the total variation distance between P_θ and Q.

Example II.8. Let $f(x) = x^l - 1$ where $l > 1$. The case $l = 2$ has already been considered in Example II.5. The function g has the expression

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

It therefore follows that $\inf_Q \sum_{\theta \in F} D_f(P_\theta||Q) \geq g(\bar{r}) \geq N^l(1 - \bar{r})^l - N$ which results in the inequality

$$r \geq \bar{r} \geq 1 - \left(\frac{1}{N^{l-1}} + \frac{\inf_Q \sum_{\theta \in F} D_f(P_\theta||Q)}{N^l}\right)^{1/l}. \quad (15)$$

When $l = 2$, inequality (15) results in a bound that is weaker than inequality (12) although for large N, the two bounds are almost the same.

Example II.9 (*Reverse* Kullback-Leibler divergence)**.** Let $f(x) = -\log x$ so that $D_f(P||Q) = D(Q||P)$. Then from Corollary II.3, we get that for every two probability measures P_1 and P_2,

$$\inf_Q \{D(Q||P_1) + D(Q||P_2)\} \geq \log\left(\frac{1}{1 - V^2}\right).$$

This can be rewritten to get

$$V \leq \sqrt{1 - \exp\left(-\inf_Q \{D(Q||P_1) + D(Q||P_2)\}\right)}. \quad (16)$$

Unlike Example II.4, it is not true that $D(Q||P_1) + D(Q||P_2)$ is minimized when $Q = \bar{P}$. This is easy to see because $D(\bar{P}, P_1) + D(\bar{P}, P_2)$ is finite only when $P_1 << P_2$ and $P_2 << P_1$. By taking $Q = P_1$ and $Q = P_2$, we get that

$$V \leq \sqrt{1 - \exp\left(-\min\left(D(P_1||P_2), D(P_2||P_1)\right)\right)}.$$

The above inequality, which is clearly weaker than inequality (16), can also be found in [20, Proof of Lemma 2.6].

III. Bounds for J_f

In order to apply the minimax lower bounds of the previous section in practical situations, we must be able to bound the quantity $J_f := \inf_Q \sum_{\theta \in F} D_f(P_\theta||Q)/N$ from above. We shall provide such bounds in this section. It should be noted that for some functions f, it may be possible to calculate J_f directly. For example, the quantity $\inf_Q \sum_{\theta \in F} H^2(P_\theta, Q)$ can be written in terms of pairwise Hellinger distances (Example II.6) and may be calculated exactly for certain probability measures P_θ. This is not the case for most functions f however.

The following is a simple upper bound for J_f which, in the case when $f(x) = x \log x$ or Kullback-Leibler divergence, has been frequently used in the literature (see for example [10] and [21]).

$$J_f \leq \frac{1}{N}\sum_{\theta \in F} D_f(P_\theta||\bar{P}) \leq \frac{1}{N^2}\sum_{\theta, \theta' \in F} D_f(P_\theta||P_{\theta'}) \leq \max_{\theta, \theta' \in F} D_f(P_\theta||P_{\theta'}).$$

We observed in section II that J_f measures the *spread* of the probability measures $P_\theta, \theta \in F$ i.e., how tightly packed/spread out they are. It should be clear that the simple bound $\max_{\theta, \theta'} D_f(P_\theta||P_{\theta'})$ does not adequately describe this aspect of $P_\theta, \theta \in F$ and it is therefore desirable to look for alternative upper bounds for J_f that capture the notion of spread in a better way.

In the case of the Kullback-Leibler divergence, Yang and Barron [1, Page 1571] provided such an upper bound for J_f. They showed that for any finite set $\{Q_\alpha : \alpha \in G\}$ of probability measures,

$$\frac{1}{N}\sum_{\theta \in F} D(P_\theta||\bar{P}) \leq \log|G| + \max_{\theta \in F}\min_{\alpha \in G} D(P_\theta||Q_\alpha). \quad (17)$$

Let us now take a closer look at this beautiful inequality of Yang and Barron [1]. The $|G|$ probability measures $Q_\alpha, \alpha \in G$ can be viewed as an approximation of the N probability measures $P_\theta, \theta \in F$. The term $\max_\theta \min_\alpha D(P_\theta||Q_\alpha)$ then denotes the approximation error, measured via the Kullback-Leibler divergence. The right hand side of inequality (17) can therefore be made small if it is possible to choose not too many probability measures Q_α which well approximate the given set of probability measures P_θ.

It should be clear how the upper bound (17) measures the spread of the probability measures $P_\theta, \theta \in F$. If the probabilities are tightly packed, it is possible to approximate them well with a smaller set of probabilities and then the bound will be small. On the other hand, if $P_\theta, \theta \in F$ are well spread out, we need more probability measures for approximation and consequently the bound will be large.

Another important aspect of inequality (17) is that it can be used to obtain lower bounds for R depending only on global metric entropy properties of the parameter space Θ and the space of probability measures $\{P_\theta, \theta \in \Theta\}$ (see section IV). On the other hand, the evaluation of inequalities resulting from the use of $J_f \leq \max_{\theta,\theta'} D(P_\theta||P_{\theta'})$ requires knowledge of both metric entropy and the existence of certain special localized subsets. We refer the reader to [1] for a detailed discussion of these issues.

The goal of this section is to generalize inequality (17) to f-divergences. The main result is given below. In section IV, we shall use this theorem along with the results of the previous section to come up with minimax lower bounds involving global entropy properties.

Theorem III.1. *Let $Q_\alpha, \alpha \in G$ be $M := |G|$ probability measures having densities $q_\alpha, \alpha \in G$ with respect to μ and let $j : F \to G$ be a mapping from F to G. For every convex function $f : [0, \infty) \to \mathbb{R}$, we have*

$$J_f \leq \frac{1}{N} \sum_{\theta \in F} \int_{\mathcal{X}} \frac{q_{j(\theta)}}{M} f\left(\frac{M p_\theta}{q_{j(\theta)}}\right) d\mu + \left(1 - \frac{1}{M}\right) f(0). \quad (18)$$

Proof: Let $\bar{Q} := \sum_{\alpha \in G} Q_\alpha / M$ and $\bar{q} := \sum_{\alpha \in G} q_\alpha / M$. Clearly for each $\theta \in F$, we have

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

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 $\bar{q} \geq q_{j(\theta)}/M$, we get that for every $\theta \in F$,

$$D_f(P_\theta||\bar{Q}) \leq \int_{\mathcal{X}} \frac{q_{j(\theta)}}{M} \left[f\left(\frac{M p_\theta}{q_{j(\theta)}}\right) - f(0) \right] d\mu + f(0).$$

Inequality (18) now follows as a consequence of the inequality $J_f \leq \sum_{\theta \in F} D_f(P_\theta||\bar{Q})/N$. ■

In the following examples, we shall demonstrate that Theorem III.1 is indeed a generalization of the bound (17) to f-divergences. We shall also see that Theorem III.1 results in inequalities that have the same qualitative structure as (17), at least for the convex functions f of interest such as $x^l - 1, l > 1$ and $(\sqrt{x} - 1)^2$.

Example III.2 (Kullback-Leibler divergence)**.** Let $f(x) = x \log x$. In this case, J_f equals $\sum_{\theta \in F} D(P_\theta||\bar{P})/N$ and invoking inequality (18), we get that

$$\frac{1}{N} \sum_{\theta \in F} D(P_\theta||\bar{P}) \leq \log M + \frac{1}{N} \sum_{\theta \in F} D(P_\theta||Q_{j(\theta)}).$$

Inequality (17) now follows if we choose $j(\theta) := \arg\min_{\alpha \in G} D(P_\theta||Q_\alpha)$. Hence Theorem III.1 is indeed a generalization of (17).

Example III.3. Let $f(x) = x^l - 1$ for $l > 1$. Applying inequality (18), we get that

$$J_f \leq M^{l-1} \left(\frac{1}{N} \sum_{\theta \in F} D_f(P_\theta||Q_{j(\theta)}) + 1 \right) - 1.$$

By choosing $j(\theta) = \arg\min_{\alpha \in G} D_f(P_\theta||Q_\alpha)$, we get that

$$J_f \leq M^{l-1} \left(\max_{\theta \in F} \min_{\alpha \in G} D_f(P_\theta||Q_\alpha) + 1 \right) - 1. \quad (19)$$

In particular, in the case of the chi-squared divergence i.e., when $l = 2$, the quantity $J_f = \inf_Q \sum_{\theta \in F} \chi^2(P_\theta||Q)/N$ is bounded from above by

$$M \left(\max_{\theta \in F} \min_{\alpha \in G} \chi^2(P_\theta||Q_\alpha) + 1 \right) - 1. \quad (20)$$

Just like (17), each of the above two inequalities is also a function of the number of probability measures Q_α and the approximation error which is now measured in terms of the chi-squared divergence.

Example III.4 (Hellinger distance)**.** Let $f(x) = (\sqrt{x} - 1)^2$ so that $D_f(P||Q) = H^2(P, Q)$, the square of the Hellinger distance between P and Q. Using inequality (18), we get that

$$J_f \leq 2 - \frac{1}{\sqrt{M}} \left(2 - \frac{1}{N} \sum_{\theta \in F} H^2(P_\theta, Q_{j(\theta)}) \right).$$

If we now choose $j(\theta) := \arg\min_{\alpha \in G} H^2(P_\theta, Q_\alpha)$, then we get

$$J_f \leq 2 - \frac{1}{\sqrt{M}} \left(2 - \max_{\theta \in F} \min_{\alpha \in G} H^2(P_\theta, Q_\alpha) \right).$$

Notice, once again, the trade-off between M and the approximation error which is measured in terms of the Hellinger distance.

IV. Bounds involving global entropy

In this section, we shall apply the results of the previous two sections to obtain lower bounds for the minimax risk R depending only on global metric entropy properties of the parameter space. The theorem is stated below, but we shall need to establish some notation first.

1) For $\eta > 0$, let $N(\eta) \geq 1$ be a real number for which there exists a finite subset $F \subseteq \Theta$ with cardinality $\geq N(\eta)$ satisfying $\rho(\theta, \theta') \geq \eta$ whenever $\theta, \theta' \in F$ and $\theta \neq \theta'$. In other words, $N(\eta)$ is a lower bound on the η-packing number of the metric space (Θ, ρ).
2) For a convex function $f : [0, \infty) \to \mathbb{R}$ satisfying $f(1) = 0$, a subset $S \subseteq \Theta$ and a positive real number ϵ, let $M_f(\epsilon, S)$ be a positive real number for which there exists a finite set G with cardinality $\leq M_f(\epsilon, S)$ and probability measures $Q_\alpha, \alpha \in G$ such that $\sup_{\theta \in S} \min_{\alpha \in G} D_f(P_\theta||Q_\alpha) \leq \epsilon^2$. In other words, $M_f(\epsilon, S)$ is an upper bound on the ϵ-covering number of the space $\{P_\theta : \theta \in S\}$ when distances are measured by the square root of the f-divergence. For purposes of clarity, we write $M_{KL}(\epsilon, S)$, $M_C(\epsilon, S)$ and $M_l(\epsilon, S)$ for

$M_f(\epsilon, S)$ when the function f equals $x \log x$, $x^2 - 1$ and $x^l - 1$ and respectively.

We note here that the probability measures $Q_\alpha, \alpha \in G$ in the definition of $M_f(\epsilon, S)$ do not need to be included in the set $\{P_\theta, \theta \in \Theta\}$ and the set G just denotes the index set and need not have any relation to S or Θ.

Theorem IV.1. *The minimax risk R satisfies the inequality $R \geq \sup_{\eta>0,\epsilon>0} \ell(\eta/2)(1-\star)$ where $\star$ stands for any of the following quantities*

$$\frac{\log 2 + \log M_{KL}(\epsilon, \Theta) + \epsilon^2}{\log N(\eta)} \tag{21}$$

$$\frac{1}{N(\eta)} + \sqrt{\frac{(1+\epsilon^2) M_C(\epsilon, \Theta)}{N(\eta)}} \tag{22}$$

and for $l > 1, l \neq 2$,

$$\left(\frac{1}{N(\eta)^{l-1}} + \frac{(1+\epsilon^2) M_l(\epsilon, \Theta)^{l-1}}{N(\eta)^{l-1}} \right)^{1/l}. \tag{23}$$

In the sequel, by inequality (22), we mean the inequality $R \geq \sup_{\eta>0,\epsilon>0} \ell(\eta/2)(1-\star)$ with $\star$ representing (22) and similarly for inequalities (21) and (23).

Proof: We shall give the proof of inequality (22). The remaining two inequalities are proved in a similar manner. Fix $\eta > 0$. By the definition of $N(\eta)$, one can find a finite subset $F \subset \Theta$ with cardinality $|F| \geq N(\eta)$ such that $\rho(\theta, \theta') \geq \eta$ for $\theta, \theta' \in F$ and $\theta \neq \theta'$. We then employ the inequality $R \geq \ell(\eta/2) r$, where r is defined as in (1). Inequality (12) can now be used to obtain

$$r \geq 1 - \frac{1}{\sqrt{|F|}} \sqrt{\frac{\inf_Q \sum_{\theta \in F} \chi^2(P_\theta || Q)}{|F|} - \frac{1}{|F|}}.$$

We now fix $\epsilon > 0$ and use the definition of $M_C(\epsilon, F)$ to get a finite set G with cardinality $\leq M_C(\epsilon, F)$ and probability measures $Q_\alpha, \alpha \in G$ such that $\sup_{\theta \in S} \min_{\alpha \in G} \chi^2(P_\theta || Q_\alpha) \leq \epsilon^2$. We then use inequality (20) to get that

$$\inf_Q \frac{1}{|F|} \sum_{\theta \in F} \chi^2(P_\theta || Q) \leq M_C(\epsilon, F)\left(1+\epsilon^2\right) - 1.$$

The proof is complete by the trivial observation $M_C(\epsilon, F) \leq M_C(\epsilon, \Theta)$. ∎

The inequality (21) is due to Yang and Barron [1, Proof of Theorem 1]. In their paper, Yang and Barron mainly considered the problem of estimation from n independent and identically distributed observations. However their method results in inequality (21) which applies to every estimation problem. Inequalities (22) and (23) are new.

Note that the lower bounds for R given in Theorem IV.1 all depend only on the quantities $N(\eta)$ and $M_f(\epsilon, \Theta)$, which describe packing/covering properties of the entire parameter space Θ. Consequently, these inequalities only involve global metric entropy properties. This is made possible by the use of inequalities in Theorem III.1. In applications of Fano's inequality (11) with the standard bound $J_f \leq \max_{\theta,\theta' \in F} D(P_\theta || P_{\theta'})$ as well as in the application of other popular methods for obtaining minimax lower bounds like Le Cam's method or Assouad's lemma, one needs to construct the finite subset F of the parameter space in a very special way: the parameter values in F should be reasonably separated in the metric ρ and also, the probability measures $P_\theta, \theta \in F$ should be close in some probability metric. In contrast, the application of Theorem IV.1 does not require the construction of such a special subset F.

Yang and Barron [1] have successfully applied inequality (21) to achieve minimax lower bounds of the optimal rate for many nonparametric density estimation and regression problems where $N(\eta)$ and $M_{KL}(\epsilon, \Theta)$ can be deduced from standard results in approximation theory for function classes. We refer the reader to [1] for examples. In some of these examples, inequality (22) can also be applied to get optimal lower bounds. In section V, we shall employ inequality (22) to obtain a new minimax lower bound in the problem of reconstructing convex bodies from noisy support function measurements.

But prior to that, let us assess the performance of inequality (22) in certain standard parametric estimation problems. In these problems, an interesting contrast arises between the two minimax lower bounds (21) and (22): the inequality (21) only results in a sub-optimal lower bound on the minimax risk (this observation, due to Yang and Barron [1, Page 1574], is also explained in Example IV.2 below) while (22) produces rate-optimal lower bounds.

Our intention here is to demonstrate, with the help of the subsequent three examples, that inequality (22) works even for finite dimensional parametric estimation problems, a scenario in which it is already known [1, Page 1574] that inequality (21) fails. Of course, obtaining optimal minimax rates in such problems is facile in most situations. For example, a two-points argument based on Hellinger distance gives the optimal rate, as is widely recognized since Le Cam [22]. But the point here is that even in finite dimensional situations, global metric entropy features are adequate for obtaining rate-optimal minimax lower bounds. This is contrary to the usual claim that in order to establish rate-optimal lower bounds in parametric settings, one needs more information than global entropy characteristics [1, Page 1574].

In each of the ensuing 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 distributed 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 shall work with the squared error loss so that $\ell(x) = x^2$, ρ is the Euclidean distance on the real line and $N(\eta)$ can be taken to c_1/η for $\eta \leq \eta_0$ where c_1 and η_0 are positive constants depending on the bounded parameter space alone. We shall 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.

Example IV.2. Suppose that m_θ equals the normal distribution with mean θ and variance 1. It can be readily verified that, for $\theta, \theta' \in \Theta$, one has

$$D(P_\theta || P_{\theta'}) = \frac{n}{2} |\theta - \theta'|^2$$

and

$$\chi^2(P_\theta||P_{\theta'}) = \exp\left(n|\theta - \theta'|^2\right) - 1.$$

It follows directly that $D(P_\theta||P_{\theta'}) \le \epsilon^2$ if and only if $|\theta-\theta'| \le \sqrt{2}\epsilon/\sqrt{n}$ and $\chi^2(P_\theta||P_{\theta'}) \le \epsilon^2$ if and only if $|\theta - \theta'| \le \sqrt{\log(1+\epsilon^2)}/\sqrt{n}$. As a result, we can take

$$M_{KL}(\epsilon, \Theta) = \frac{c_2\sqrt{n}}{\epsilon} \text{ and } M_C(\epsilon, \Theta) = \frac{c_2\sqrt{n}}{\sqrt{\log(1+\epsilon^2)}} \quad (24)$$

for $\epsilon \le \epsilon_0$. Now, inequality (21) says that the minimax risk R_n satisfies

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

The function $\epsilon \mapsto \epsilon^2 - \log \epsilon$ is minimized on $[0, \epsilon_0]$ at, say, $\epsilon = \epsilon_1$ and we then get

$$R_n \ge \sup_{\eta \le \eta_0} \frac{\eta^2}{4}\left(1 - \frac{\log n + c_3}{2\log c_1 + 2\log(1/\eta)}\right), \quad (25)$$

where c_3 is a function of c_2 and ϵ_1. We now note that when $\eta = c/\sqrt{n}$ for a constant c, the quantity inside the parantheses on the right hand side of (25) converges to 0 as n goes to ∞. This means that inequality (21) only gives lower bounds of inferior order for R_n, the optimal order being, of course, $1/n$.

On the other hand, we shall show below that inequality (22) gives $R_n \ge c/n$ for a positive constant c. Indeed, inequality (22) says that

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

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

$$R_n \ge \frac{c_3^2}{4n}\left(1 - \frac{c_3}{c_1\sqrt{n}} - c_4\sqrt{c_3}\right), \quad (26)$$

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

Example IV.3. 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 $[0, \theta]$. It is then elementary to check that the chi-squared divergence between P_θ and $P_{\theta'}$ equals $(\theta'/\theta)^n - 1$ if $\theta \le \theta'$ and ∞ otherwise. It follows accordingly that $\chi^2(P_\theta||P_{\theta'}) \le \epsilon^2$ provided

$$0 \le \theta' - \theta \le \frac{\theta\log(1+\epsilon^2)}{n}. \quad (27)$$

Because the parameter space is a compact interval bounded away from zero, in order to ensure (27), it is enough to require that $0 \le \theta' - \theta \le c_2\log(1+\epsilon^2)/n$. Therefore, we can take

$$M_C(\epsilon, \Theta) = \frac{c_3 n}{\log(1+\epsilon^2)}$$

for $\epsilon \le \epsilon_0$. Inequality (22) now implies that

$$R_n \ge \sup_{\eta \le \eta_0, \epsilon \le \epsilon_0} \frac{\eta^2}{4}\left(1 - \frac{\eta}{c_1} - \sqrt{\eta n}\sqrt{\frac{c_3(1+\epsilon^2)}{c_1\log(1+\epsilon^2)}}\right).$$

Taking $\epsilon = \epsilon_0$ and $\eta = c_4/n$, we get that

$$R_n \ge \frac{c_4^2}{4n^2}\left(1 - \frac{c_4}{nc_1} - \sqrt{c_4}c_5\right),$$

where c_5 depends only on c_1, c_3 and ϵ_0. Hence by choosing c_4 sufficiently small, we get that $R_n \ge c/n^2$ 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 IV.4. Suppose that m_θ denotes the uniform distribution on the interval $[\theta, \theta+1]$. We shall argue that $M_C(\epsilon, \Theta)$ can be chosen to be

$$M_C(\epsilon, \Theta) = \frac{c_2}{(1+\epsilon^2)^{1/n} - 1} \quad (28)$$

for a positive constant c_2 at least for large n. 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 $\epsilon \le \epsilon_0$. 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 $\alpha \le \theta \le \alpha + \epsilon'$. 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 $M_C(\epsilon, \Theta)$ can be taken to be the number of probability measures Q_α, which is the same as the number of points in G. We thus have (28). It can be checked by elementary calculus (Taylor expansion, for example) that the inequality

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

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

$$M_C(\epsilon, \Theta) \le \frac{2nc_2}{2\epsilon^2 - (1-1/n)\epsilon^4}.$$

From inequality (22), we get that for every $\eta \le \eta_0$ and $\epsilon \le \min(\epsilon_0, \sqrt{2})$,

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

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

The fact that inequality (22) produced optimal lower bounds for the minimax risk in each of the above three examples is reassuring but not really exciting because, as we mentioned before, there are other simpler methods of obtaining such bounds in these examples. We presented them as simple toy examples to evaluate the performance of (22), to present a difference between (21) and (22) (which provides a justification for using divergences other than the Kullback-Leibler divergence for lower bounds) and also to stress the fact that global packing and covering characteristics are enough to obtain optimal minimax lower bounds. In order to convince the reader of the effectiveness of (22) in more involved situations, we now apply it to obtain the optimal minimax rate in a d-dimensional normal mean estimation problem. We are grateful to an anonymous referee for communicating this example to us. Another non-trivial application of (22) is presented in the next section.

Example IV.5. 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 so that $\ell(x) = x^2$ and ρ is the Euclidean distance on $\mathbb{R}^d$.

We shall use inequality (22) to show that the minimax risk R for this problem is larger than or equal to a constant multiple of $d\sigma^2$ when $\Gamma \geq \sigma\sqrt{d}$. We begin by arguing that we can take

$$N(\eta) = \left(\frac{\Gamma}{\eta}\right)^d, M_C(\epsilon, \Theta) = \left(\frac{3\Gamma}{\sigma\sqrt{\log(1+\epsilon^2)}}\right)^d \tag{29}$$

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

For $N(\eta)$, we first note that the η-packing number of the metric space (Θ, ρ) is bounded from below by its η-covering number. Now, for any η-covering set, the space Θ is contained in the union of the balls of radius η with centers in the covering set and hence the volume of Θ must be smaller than the sum of the volumes of these balls. Therefore, the number of points in the η-covering set must be at least $(\Gamma/\eta)^d$. Since this is true for every η-covering set, it follows that the η-covering number and hence the η-packing number is not smaller than $(\Gamma/\eta)^d$.

For $M_C(\epsilon, \Theta)$, we first observe that for $\theta, \theta' \in \Theta$, the chi-squared divergence between P_θ and $P_{\theta'}$ can be easily computed (because they are normal distributions with the same covariance matrix) to be $\chi^2(P_\theta || P_{\theta'}) = \exp\left(\rho^2(\theta, \theta')/\sigma^2\right) - 1$. Therefore $\chi^2(P_\theta || P_{\theta'}) \leq \epsilon^2$ if and only if $\rho(\theta, \theta') \leq \epsilon' := \sigma\sqrt{\log(1+\epsilon^2)}$. As a result, $M_C(\epsilon, \Theta)$ can be taken to be any upper bound on the ϵ'-covering number of (Θ, ρ). The ϵ'-covering number, as noted previously, is bounded from above by the ϵ'-packing number. Now, for any ϵ'-packing set, the balls of radius $\epsilon'/2$ with centers in the packing set are all disjoint and their union is contained in the ball of radius $\Gamma + (\epsilon'/2)$ centered at the origin. Consequently, the sum of the volumes of these balls is smaller than the volume of the ball of radius $\Gamma + (\epsilon'/2)$ centered at the origin. Therefore, the number of points in the ϵ'-packing set is at most $(1 + (2\Gamma/\epsilon'))^d \leq (3\Gamma/\epsilon')^d$ provided $\epsilon' \leq \Gamma$. Since this is true for every ϵ'-packing set, it follows that the ϵ'-packing number and hence the ϵ'-covering number is not larger than $(3\Gamma/\epsilon')^d$.

We can thus apply inequality (22) with (29) to get that, for every $\eta > 0$ and $\epsilon > 0$ such that $\sigma\sqrt{\log(1+\epsilon^2)} \leq \Gamma$, we have

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

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

$$R \geq \sup_{\eta > 0} \frac{\eta^2}{4}\left(1 - \left(\frac{\eta}{\Gamma}\right)^d - \left(\frac{18e\eta^2}{\sigma^2 d}\right)^{d/4}\right).$$

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

$$R \geq \frac{c_1^2 \sigma^2 d}{4}\left(1 - c_1^d - (18ec_1^2)^{d/4}\right).$$

We can therefore choose c_1 small enough (independent of d) to obtain that $R \geq cd\sigma^2$. Note that, up to constants, this lower bound is optimal for R because $\mathbb{E}\rho^2(X, \theta) = d\sigma^2$.

V. Reconstruction of convex bodies from noisy support function measurements

In this section, we shall present a novel application of the global minimax lower bound (22). Let $d \geq 2$ and let K be a convex body in $\mathbb{R}^d$, i.e., K is compact, convex and has a non-empty interior. The support function of K, $h_K : S^{d-1} \to \mathbb{R}$, is defined by

$$h_K(u) := \sup\{\langle x, u\rangle : x \in K\} \text{ for } u \in S^{d-1},$$

where $S^{d-1} := \left\{x \in \mathbb{R}^d : \sum_i x_i^2 = 1\right\}$ is the unit sphere. We direct the reader to [23, Section 1.7] or [24, Section 13] for basic properties of support functions. An important property is that the support function uniquely determines the convex body, i.e., $h_K = h_L$ if and only if $K = L$.

Let $\{u_i, i \geq 1\}$ be a sequence of d-dimensional unit vectors. Gardner, Kiderlen and Milanfar [25] (see their paper for earlier references) considered the problem of reconstructing an unknown convex body K from noisy measurements of h_K in the directions $u_1, \dots, u_n$. More precisely, their problem was to estimate K from observations $Y_1, \dots, Y_n$ drawn according to the model $Y_i = h_K(u_i) + \xi_i, i = 1, \dots, n$ where $\xi_1, \dots, \xi_n$ are independent and identically distributed mean zero gaussian random variables. They constructed a convex body (estimator) $\hat{K}_n = \hat{K}_n(Y_1, \dots, Y_n)$ having the property that, for *nice* sequences $\{u_i, i \geq 1\}$, the L^2 norm $||h_K - h_{\hat{K}_n}||_2$ (see (30) below) converges to zero at the rate $n^{-2/(d+3)}$ for dimensions $d = 2, 3, 4$ and at a slower rate for dimensions $d \geq 5$ (see [25, Theorem 6.2]).

We shall show here that in the same setting, it is impossible in the minimax sense to construct estimators for K converging at a rate faster than $n^{-2/(d+3)}$. This implies that the least squares estimator in [25] is rate optimal for dimensions $d = 2, 3, 4$. We shall need some notation to describe our result.

Let $\mathcal{K}^d$ denote the set of all convex bodies in $\mathbb{R}^d$ and for $\Gamma > 0$, let $\mathcal{K}^d(\Gamma)$ denote the set of all convex bodies in $\mathbb{R}^d$ that are contained in the closed ball of radius Γ centered at the origin so that $\mathcal{K}^d(1)$ denotes the set of all convex bodies contained in the unit ball, which we shall denote by B. Note that estimating K is equivalent to estimating the function h_K because the support function uniquely determines the convex body. Thus we shall focus on the problem of estimating h_K.

An estimator for h_K is allowed to be a bounded function on S^{d-1} that depends on the data $Y_1, \dots, Y_n$. The loss functions that we shall use are the L^p norms for $p \in [1, \infty]$ defined by

$$||h_K - \hat{h}||_p := \left(\int_{S^{d-1}} |h_K(u) - \hat{h}(u)|^p du\right)^{1/p} \tag{30}$$

for $p \in [1, \infty)$ and $||h_K - \hat{h}||_\infty := \sup_{u \in S^{d-1}} |h_K(u) - \hat{h}(u)|$. For convex bodies K and L and $p \in [1, \infty]$, we shall also write

$\delta_p(K, L)$ for $||h_K - h_L||_p$ and refer to δ_p as the L^p distance between K and L.

We shall consider the minimax risk of the problem of estimating h_K from $Y_1, \dots, Y_n$ when K is assumed to belong to $\mathcal{K}^d(\Gamma)$ i.e., we are interested in the quantity

$$r_n(p, \Gamma) := \inf_{\hat{h}} \sup_{K \in \mathcal{K}^d(\Gamma)} \mathbb{E}_K ||h_K - \hat{h}(Y_1, \dots, Y_n)||_p.$$

The following is the main theorem of this section.

Theorem V.1. *Fix $p \in [1, \infty)$ and $\Gamma > 0$. Suppose the errors $\xi_1, \dots, \xi_n$ are independent normal random variables with mean zero and variance σ^2. Then the minimax risk $r_n(p, \Gamma)$ satisfies*

$$r_n(p, \Gamma) \geq c\sigma^{4/(d+3)}\Gamma^{(d-1)/(d+3)}n^{-2/(d+3)}, \qquad (31)$$

for a constant c that is independent of n.

Remark V.1. In the case when $p = 2$, Gardner, Kiderlen and Milanfar [25] showed that the least squares estimator converges at the rate given by the right hand side of (31) for dimensions $d = 2, 3, 4$. Thus, at least for $p = 2$, the lower bound given by (31) is optimal for dimensions $d = 2, 3, 4$.

We shall use inequality (22) to prove (31). First, let us put the support function estimation problem in the general estimation setting of the last section. Let $\Theta := \{h_K : K \in \mathcal{K}^d(\Gamma)\}$ and let $\mathcal{A}$ be the collection of all bounded functions on the unit sphere S^{d-1}. The metric ρ on $\mathcal{A}$ is just the L^p norm and $\ell(x) = x$.

Finally, let $\mathcal{X} = \mathbb{R}^n$ and for $f \in \Theta$, let P_f be the n-variate normal distribution with mean vector $(f(u_1), \dots, f(u_n))$ and variance-covariance matrix $\sigma^2 I_n$, where I_n is the identity matrix of order n.

In order to apply inequality (22), we need to determine $N(\eta)$ and $M_C(\epsilon, \Theta)$. The quantity $N(\eta)$ is a lower bound on the η-packing number of the set $\mathcal{K}^d(\Gamma)$ under the L^p norm. When $p = \infty$, Bronshtein [26, Theorem 4 and Remark 1] proved that there exist positive constants c' and η_0 depending only on d such that $\exp\left(c'(\eta/\Gamma)^{(1-d)/2}\right)$ is a lower bound for the η-packing number of Θ for $\eta \leq \eta_0$. It is a standard fact that $p = \infty$ corresponds to the Hausdorff metric on $\mathcal{K}^d(\Gamma)$.

It turns out that Bronshtein's result is actually true for every $p \in [1, \infty]$ and not just for $p = \infty$. However, to the best of our knowledge, this has not been proved anywhere in the literature. By modifying Bronshtein's proof appropriately and using the Varshamov-Gilbert lemma (see for example [27, Lemma 4.7]), we provide, in Theorem VII.1, a proof of this fact. Therefore from Theorem VII.1, we can take

$$\log N(\eta) = c'\left(\frac{\Gamma}{\eta}\right)^{(d-1)/2} \quad \text{for } \eta \leq \eta_0, \qquad (32)$$

where c' and η_0 are constants depending only on d and p.

Now let us turn to $M_C(\epsilon, \Theta)$. For $f, g \in \Theta$, P_f and P_g are normal distributions with the same covariance matrix and hence the chi-squared divergence between P_f and P_g can be seen to be

$$\chi^2(P_f||P_g) = \exp\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(f(u_i) - g(u_i))^2\right] - 1$$
$$\leq \exp\left[\frac{n||f-g||_\infty^2}{\sigma^2}\right] - 1.$$

It follows that

$$||f - g||_\infty \leq \epsilon' \Longrightarrow \chi^2(P_f||P_g) \leq \epsilon^2. \qquad (33)$$

where $\epsilon' := \sigma\sqrt{\log(1+\epsilon^2)}/\sqrt{n}$. Let $W_{\epsilon'}$ be the smallest W for which there exist sets $K_1, \dots, K_W$ in $\mathcal{K}^d(\Gamma)$ having the property that for every set $K \in \mathcal{K}^d(\Gamma)$, there exists a K_j such that the Hausdorff distance between K and K_j is less than or equal to ϵ'. It must be clear from (33) that $M_C(\epsilon, \Theta)$ can be taken to be a number larger than $W_{\epsilon'}$. Bronshtein [26, Theorem 3 and Remark 1] showed that there exist positive constants c'' and ϵ_0 depending only on d such that

$$\log W_{\epsilon'} \leq c''\left(\frac{\Gamma}{\epsilon'}\right)^{(d-1)/2} \quad \text{for } \epsilon' \leq \epsilon_0.$$

Hence for all ϵ such that $\log(1+\epsilon^2) \leq n\epsilon_0^2/\sigma^2$, we can take

$$\log M_C(\epsilon, \Theta) = c''\left(\frac{\Gamma\sqrt{n}}{\sigma\sqrt{\log(1+\epsilon^2)}}\right)^{(d-1)/2}. \qquad (34)$$

We are now ready to prove inequality (31). We shall define two quantities

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

and

$$u(n) := \left(\frac{\Gamma\sqrt{n}}{\sigma}\right)^{(d-1)/(d+3)}.$$

where $c = c(d, p)$ will be specified shortly. Also let $\epsilon(n)$ be such that $\log(1 + \epsilon^2(n)) = u^2(n)$. Clearly as $n \to \infty$, we have $\eta(n) \to 0$, $u(n) \to \infty$ and $u(n)/\sqrt{n} \to 0$. As a result $\eta(n) \leq \eta_0$ and $u^2(n) \leq n\epsilon_0^2/\sigma^2$ for large n and therefore from (32) and (34), we get that

$$\log N(\eta(n)) = c'\left(\frac{\Gamma}{\eta(n)}\right)^{(d-1)/2} = \frac{c'}{c^{(d-1)/2}}u^2(n).$$

and

$$\log M_C(\epsilon(n), \Theta) = c''\left(\frac{\Gamma\sqrt{n}}{\sigma u(n)}\right)^{(d-1)/2} = c''u^2(n).$$

We now apply inequality (22) (recall that $\ell(x) = x$) to obtain that $r_n(p, \Gamma)$ is bounded from below by

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

for all large n. If we now choose c so that $c^{(d-1)/2} = c'/(2 + 2c'')$, we get that

$$r_n(p, \Gamma) \geq \frac{\eta(n)}{2}\left[1 - \frac{1}{N(\eta(n))} - \exp\left(\frac{-u^2(n)}{2}(1 + c'')\right)\right].$$

Now observe that as $n \to \infty$, the quantity $\eta(n)$ goes to 0 and hence $N(\eta(n))$ goes to ∞. Further, as we have already noted, $u(n)$ goes to ∞. It follows hence that $r_n(p, \Gamma) \geq \eta(n)/4$ for all large n. By choosing c even smaller, we can make inequality (31) true for all n.

VI. A COVARIANCE MATRIX ESTIMATION EXAMPLE

In the previous section, we have used the global minimax lower bound (22). However, in some situations, the global entropy numbers might be difficult to bound. In such cases, inequalities (21) and (22) are, of course, not applicable and we are unaware of the use of inequality (17) in conjuction with Fano's inequality (11) in the literature. The standard examples use (11) with the bound $J_f \leq \min_{\theta,\theta' \in F} D(P_\theta || P_{\theta'})$ while the examples in [1] all deal with the case when global entropies are available. In this section, we shall demonstrate how a recent minimax lower bound due to Cai, Zhang and Zhou [13] can also be proved using inequalities (11) and (17).

Cai, Zhang and Zhou [13] considered n independent $p \times 1$ random vectors $X_1, \dots, X_n$ distributed according to $N_p(0, \Sigma)$. Suppose that the entries of the $p \times p$ covariance matrix $\Sigma = (\sigma_{ij})$ decay at a certain rate as we move away from the diagonal. Specifically, let us suppose that for a fixed positive constant $\alpha > 0$, the entries σ_{ij} of Σ satisfy the inequality $\sigma_{ij} \leq |i-j|^{-\alpha-1}$ for $i \neq j$. Cai, Zhang and Zhou [13] showed that when p is large compared to n, it is impossible to estimate Σ from $X_1, \dots, X_n$ in the spectral norm at a rate faster than $n^{-\alpha/(2\alpha+1)}$. More precisely, they showed that when $p \geq Cn^{1/(2\alpha+1)}$,

$$R_n(\alpha) := \inf_{\hat{\Sigma}} \sup_{\Sigma \in \Theta} \mathbb{E}_\Sigma ||\hat{\Sigma} - \Sigma|| \geq c\ n^{-\alpha/(2\alpha+1)}, \tag{35}$$

where c and C denote positive constants depending only on α. Here Θ denotes the collection of all covariance matrices $\Sigma = (\sigma_{ij})$ satisfying $\sigma_{ij} \leq |i-j|^{-\alpha-1}$ for $i \neq j$ and the norm $||.||$ is the spectral norm (largest eigenvalue).

Cai, Zhang and Zhou [13] used Assouad's lemma for the proof of the inequality (35). We shall use inequalities (11) and (17). Moreover, the choice of the finite subset F that we use is different from the one used in [13, Equation (17)]. This makes our approach different from the general method, due to Yu [28], of replacing Assouad's lemma by Fano's inequality.

Throughout, Δ denotes a constant that depends on α alone. The value of the constant might vary from place to place.

Consider the matrix $A = (a_{ij})$ with $a_{ij} = 1$ for $i = j$ and $a_{ij} = 1/(\Delta |i-j|^{\alpha+1})$ for $i \neq j$. For Δ sufficiently large (depending on α alone), A is positive definite and belongs to Θ. Let us fix a positive integer $k \leq p/2$ and partition A as

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{12}^T & A_{22} \end{array} \right],$$

where A_{11} is $k \times k$ and A_{22} is $(p-k) \times (p-k)$. For each $\tau \in \mathbb{R}^k$, we define the matrix

$$A(\tau) := \left[\begin{array}{c|c} A_{11} & A_{12}(\tau) \\ \hline (A_{12}(\tau))^T & A_{22} \end{array} \right],$$

where $A_{12}(\tau)$ is the $k \times (p-k)$ matrix obtained by premultiplying A_{12} with the $k \times k$ diagonal matrix with diagonal entries $\tau_1, \dots, \tau_k$. Clearly, $A(\tau) \in \Theta$ for all $\tau \in \{0,1\}^k$. We shall need the following two lemmas in order to prove inequality (35).

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

$$||A(\tau) - A(\tau')|| \geq \frac{1}{\Delta k^\alpha} \sqrt{\frac{\Upsilon(\tau,\tau')}{k}}, \tag{36}$$

where $\Upsilon(\tau,\tau') := \sum_{r=1}^k \{\tau_r \neq \tau'_r\}$ denotes the Hamming distance between τ and τ'.

Proof: Fix $\tau, \tau' \in \{0,1\}^k$ with $\tau \neq \tau'$. Let v denote the $p \times 1$ vector $(0_k, 1_k, 0_{p-2k})^T$, where 0_k denotes the $k \times 1$ vector of zeros etc. Clearly $||v||^2 = k$ and $(A(\tau) - A(\tau'))v$ will be a vector of the form $(u, 0)^T$ for some $k \times 1$ vector $u = (u_1, \dots, u_k)^T$. Moreover $u_r = \sum_{s=1}^k (\tau_r - \tau'_r) a_{r,k+s}$ and hence

$$\begin{aligned} |u_r| &= \frac{\{\tau_r \neq \tau'_r\}}{\Delta} \sum_{s=1}^k \frac{1}{|r-k-s|^{\alpha+1}} \\ &\geq \frac{\{\tau_r \neq \tau'_r\}}{\Delta} \sum_{i=k}^{2k-1} \frac{1}{i^{\alpha+1}} \geq \frac{\{\tau_r \neq \tau'_r\}}{\Delta} \frac{1}{k^\alpha}. \end{aligned}$$

Therefore,

$$||\left(A(\tau) - A(\tau')\right) v||^2 \geq \sum_{r=1}^k u_r^2 \geq \frac{1}{\Delta^2 k^{2\alpha}} \Upsilon(\tau,\tau').$$

The proof is complete because $||v||^2 = k$. ∎

Lemma VI.2. *Let $1 \leq m < k, \tau \in \{0,1\}^k$ and $\tau' := (0, \dots, 0, \tau_m, \dots, \tau_k)$. Then*

$$D\left(N(0, A(\tau)) || N(0, A(\tau'))\right) \leq \frac{\Delta}{(k-m)^{2\alpha}}.$$

Proof: The key is to note that one has the inequality $D\left(N(0, A(\tau)) || N(0, A(\tau'))\right) \leq \Delta ||A(\tau) - A(\tau')||_F^2$, where $||A||_F := \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$ denotes the Frobenius norm. The proof of this assertion can be found in [13, Proof of Lemma 6]. We can now bound

$$\begin{aligned} ||A(\tau) - A(\tau')||_F^2 &\leq 2 \sum_{r=1}^{m-1} \tau_r^2 \sum_{j=1}^{p-k} a_{r,k+j}^2 \\ &\leq \Delta \sum_{r=1}^{m-1} \sum_{j=1}^{p-k} \frac{1}{|r-k-j|^{2\alpha+2}} \\ &\leq \Delta \sum_{r=1}^{m-1} \sum_{j=1}^{\infty} \frac{1}{|k-r+j|^{2\alpha+2}} \\ &\leq \Delta \sum_{r=1}^{m-1} \frac{1}{(k-r)^{2\alpha+1}} \leq \frac{\Delta}{(k-m)^{2\alpha}}. \end{aligned}$$

The proof is complete. ∎

The Varshamov-Gilbert lemma (see for example [27, Lemma 4.7]) asserts the existence of a subset W of $\{0,1\}^k$ with $|W| \geq \exp(k/8)$ such that $\Upsilon(\tau,\tau') \geq k/4$ for all $\tau, \tau' \in W$ with $\tau \neq \tau'$. Let $F := \{A(\tau) : \tau \in W\}$. From inequality (11) and Lemma VI.1, we get that

$$R_n(\alpha) \geq \frac{1}{\Delta} \frac{1}{k^\alpha} \left(1 - \frac{\log 2 + \frac{1}{|W|} \sum_{A \in F} D(P_A || \bar{P})}{k/8}\right), \tag{37}$$

where P_A denotes the n-fold product of the $N(0, A)$ probability measure and $\bar{P} := \sum_{A \in F} P_A/|W|$. Now for $1 \leq m < k$ and for $t \in \{0,1\}^{k-m+1}$, let Q_t denote the n-fold product of the $N(0, A(0, \dots, 0, t_1, \dots, t_{k-m+1}))$ probability measure. Applying inequality (17), we get the quantity $\sum_{A \in F} D(P_A || \bar{P})/|W|$ is bounded from above by

$$(k-m+1)\log 2 + \max_{A \in F} \min_{t \in \{0,1\}^{k-m+1}} D(P_A || Q_t).$$

Now we use Lemma VI.2 to obtain

$$\frac{1}{|W|} \sum_{A \in F} D(P_A || \bar{P}) \leq \Delta \left[(k-m) + \frac{n}{(k-m)^{2\alpha}} \right].$$

Using the above in (37), we get

$$R_n(\alpha) \geq \frac{1}{\Delta} \frac{1}{k^\alpha} \left[1 - \frac{\Delta}{k} \left((k-m) + \frac{n}{(k-m)^\alpha} \right) \right].$$

Note that the above lower bound for $R_n(\alpha)$ depends on k and m, which are constrained to satisfy $2k \leq p$ and $1 \leq m < k$. To get the best lower bound, we need to optimize the right hand side of the above inequality over k and m. It should be obvious that in order to prove (35), it is enough to take $k - m = n^{1/(2\alpha+1)}$ and $k = 4\Delta n^{1/(2\alpha+1)}$. The condition $2k \leq p$ will be satisfied if $p \geq C n^{1/(2\alpha+1)}$ for a large enough C. It is elementary to check that with these choices of k and m, inequality (35) is established.

VII. A Packing Number Lower Bound

In this section, we shall prove that for every $p \in [1, \infty]$ the η-packing number $N(\eta; p, \Gamma)$ of $\mathcal{K}^d(\Gamma)$ under the L^p metric is at least $\exp\left(c(\eta/\Gamma)^{(1-d)/2}\right)$ for a positive c and sufficiently small η. This means that there exist at least $\exp\left(c(\eta/\Gamma)^{(1-d)/2}\right)$ sets in $\mathcal{K}^d(\Gamma)$ separated by at least η in the L^p metric. This result was needed in the proof of Theorem V.1. Bronshtein [26, Theorem 4 and Remark 1] proved this for $p = \infty$ (the case of the Hausdorff metric).

Theorem VII.1. *Fix $p \in [1, \infty]$. There exist positive constants η_0 and C depending only on d and p such that for every $\eta \leq \eta_0$, we have*

$$N(\eta; p, \Gamma) \geq \exp\left(C \left(\frac{\Gamma}{\eta} \right)^{(d-1)/2} \right). \tag{38}$$

Proof: Observe that by scaling, it is enough to prove for the case $\Gamma = 1$. We loosely follow Bronshtein [26, Proof of Theorem 4]. Fix $\epsilon \in (0, 1)$. For each point $x \in S^{d-1}$, let S_x denote the supporting hyperplane to the unit ball B at x and let H_x be the hyperplane intersecting the sphere that is parallel to S_x and at a distance of ϵ from S_x. Let H_x^+ and H_x^- denote the two halfspaces bounded by H_x where we assume that H_x^+ contains the origin. Let $T_x := S^{d-1} \cap H_x^-$ and $A_x := B \cap H_x$, where B stands for the unit ball. It can be checked that the (Euclidean) distance between x and every point in T_x (and A_x) is less than or equal to $\sqrt{2}\sqrt{\epsilon}$. It follows that if the distance between two points x and y in S^{d-1} is strictly larger than $2\sqrt{2}\sqrt{\epsilon}$, then the sets T_x and T_y are disjoint.

By standard results (see for example [26, Proof of Theorem 4] where it is referred to as Mikhlin's result), there exist positive constants C_1, depending only on d, and ϵ_0 such that for every $\epsilon \leq \epsilon_0$, there exist $N \geq C_1(\sqrt{\epsilon})^{1-d}$ points $x_1, \dots, x_N$ in S^{d-1} such that the Euclidean distance between x_i and x_j is strictly larger than $2\sqrt{2}\sqrt{\epsilon}$ whenever $i \neq j$. From now on, we assume that $\epsilon \leq \epsilon_0$. We then consider a mapping $\Phi : \{0,1\}^N \to \mathcal{K}^d(1)$, which is defined, for $\tau = (\tau_1, \dots, \tau_N) \in \{0,1\}^N$, by

$$\Phi(\tau) := B \cap D_1(\tau_1) \cap D_2(\tau_2) \cap \dots \cap D_N(\tau_N),$$

where for $i = 1, \dots, N$,

$$D_i(0) := H_{x_i}^+ \text{ and } D_i(1) := B.$$

It must be clear that the Hausdorff distance between $\Phi(\tau)$ and $\Phi(\tau')$ is not less than ϵ (in fact, it is exactly equal to ϵ) if $\tau \neq \tau'$. Thus, $\left\{\Phi(\tau) : \tau \in \{0,1\}^N\right\}$ is an ϵ-packing set for $\mathcal{K}^d(1)$ under the Hausdorff metric. However, it is *not* an ϵ-packing set under the L^p metric. Indeed, the L^p distance between $\Phi(\tau)$ and $\Phi(\tau')$ is not necessarily larger than ϵ for all pairs $(\tau, \tau'), \tau \neq \tau'$. The L^p distance between $\Phi(\tau)$ and $\Phi(\tau')$ depends on the Hamming distance $\Upsilon(\tau, \tau') = \sum_i \{\tau_i \neq \tau'_i\}$ between τ and τ'. We make the claim that

$$\delta_p(\Phi(\tau), \Phi(\tau')) \geq C_2 \epsilon \left(\sqrt{\epsilon}\right)^{(d-1)/p} \left(\Upsilon(\tau, \tau')\right)^{1/p}, \tag{39}$$

where C_2 depends only on d and p. The claim will be proved later. Assuming it is true, we recall the Varshamov-Gilbert lemma from the previous section to assert the existence of a subset W of $\{0,1\}^N$ with $|W| \geq \exp(N/8)$ such that $\Upsilon(\tau, \tau') \geq N/4$ for all $\tau, \tau' \in W$ with $\tau \neq \tau'$. Because $N \geq C_1(\sqrt{\epsilon})^{1-d}$, we get from (39) that for all $\tau, \tau' \in W$ with $\tau \neq \tau'$, we have

$$\delta_p(\Phi(\tau), \Phi(\tau')) \geq C_3 \epsilon \text{ where } C_3 := C_2 \left(\frac{C_1}{4} \right)^{1/p}.$$

Taking $\eta := C_3 \epsilon$, we have obtained, for each $\eta \leq \eta_0 := C_3 \epsilon_0$, an η-packing subset of $\mathcal{K}^d(1)$ with size M, where

$$\log M \geq N/8 \geq \frac{C_1}{8} \left(\frac{1}{\sqrt{\epsilon}} \right)^{d-1} = C_4 \left(\frac{1}{\sqrt{\eta}} \right)^{d-1}.$$

The constant C_4 only depends on d and p thereby proving (38).

It remains to prove the claim (39). Fix a point $x \in S^{d-1}$ and $\epsilon \in (0, 1)$. We first observe that it is enough to prove that

$$\delta_p(A_x, T_x)^p \geq C_5 \epsilon^p \left(\sqrt{\epsilon}\right)^{d-1}, \tag{40}$$

for a constant C_5 depending on just d and p, where A_x and T_x are as defined in the beginning of the proof. This is because of the fact that for every $\tau, \tau' \in W$ with $\tau \neq \tau'$, we can write

$$\delta_p(\Phi(\tau), \Phi(\tau'))^p = \sum_{i \in I} \delta_p(A_{x_i}, T_{x_i})^p, \tag{41}$$

where $I := \{1 \leq i \leq N : \tau_i \neq \tau'_i\}$. The equality (41) is a consequence of the fact that the points $x_1, \dots, x_N$ are chosen so that $T_{x_1}, \dots, T_{x_N}$ are disjoint.

We shall now prove the inequality (40) which will complete the proof. Let u_0 denote the point in A_x that is closest to the origin. Also let u_1 be a point in $A_x \cap S^{d-1}$. Let α denote the angle between u_0 and u_1. Clearly, α does not depend on the choice of u_1 and $\cos \alpha = 1 - \epsilon$. Now let u be a fixed unit vector and let θ be the angle between the vectors u and u_0. By elementary geometry, we deduce that

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

Because the difference of support functions only depends on the angle θ, we can write, for a constant C_6 depending only on d, that

$$\delta_p(A_x, T_x)^p = C_6 \int_0^{\alpha} (1 - \cos(\alpha - \theta))^p \sin^{d-2}\theta d\theta.$$

Now suppose β is such that $\cos(\alpha - \beta) = 1 - \epsilon/2$. Then from above, we get that

$$\begin{aligned} \delta_p(A_x, T_x)^p &\geq C_6 \int_0^{\beta} (1 - \cos(\alpha - \theta))^p \sin^{d-2}\theta d\theta \\ &\geq C_6 \left(\frac{\epsilon}{2}\right)^p \int_0^{\beta} \sin^{d-2}\theta d\theta \\ &\geq C_6 \left(\frac{\epsilon}{2}\right)^p \int_0^{\beta} \sin^{d-2}\theta \cos\theta d\theta \\ &= \frac{C_6}{d-1} \left(\frac{\epsilon}{2}\right)^p \sin^{d-1}\beta. \end{aligned}$$

We shall show that $\sin\beta \geq (\sqrt{\epsilon})/(2\sqrt{2})$ which will prove (40). Recall that $\cos\alpha = 1 - \epsilon$. Thus

$$\begin{aligned} 1 - \frac{\epsilon}{2} &= \cos(\alpha - \beta) \\ &\leq \cos\alpha + \sin\alpha \sin\beta \\ &= 1 - \epsilon + \sqrt{1 - (1-\epsilon)^2} \sin\beta \\ &\leq 1 - \epsilon + \sqrt{2}\sqrt{\epsilon}\sin\beta, \end{aligned}$$

which when rearranged gives $\sin\beta \geq (\sqrt{\epsilon})/(2\sqrt{2})$. The proof is complete. ∎

VIII. Conclusion

By a simple application of convexity, we proved an inequality relating the minimax risk in multiple hypothesis testing problems to f-divergences of the probability measures involved. This inequality is an extension of Fano's inequality. As another corollary, we obtained a sharp inequality between total variation distance and f-divergences. We also indicated how to control the quantity J_f which appears in our lower bounds. This leads to important global lower bounds for the minimax risk. Two applications of our bounds are presented. In the first application, we used the bound (22) to prove a new lower bound (which turns to be rate-optimal) for the minimax risk of estimating a convex body from noisy measurements of the support function in n directions. In the second application, we employed inequalities (11) and (17) to give a different proof of a recent lower bound for covariance matrix estimation due to Cai, Zhang and Zhou [13].

Acknowledgment

The author is indebted to David Pollard for his insight and also for numerous stimulating discussions which led to many of the ideas in this paper; to Andrew Barron for his constant encouragement and willingness to discuss his own work on minimax bounds. Thanks are also due to Aditya Mahajan for pointing out to the author that inequality (5) has the extension (4) for the case of non-uniform priors w; to an anonymous referee for helpful comments, for pointing out an error and for Example IV.5; to Richard Gardner for comments that greatly improved the quality of the paper and to Alexander Gushchin for informing the author about his paper [14] and for sending him a scanned copy of it.

References

[1] Y. Yang and A. Barron, “Information-theoretic determination of minimax rates of convergence,” *Annals of Statistics*, vol. 27, pp. 1564–1599, 1999.
[2] S. M. Ali and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another,” *Journal of the Royal Statistical Society, Series B*, vol. 28, pp. 131–142, 1996.
[3] I. Csiszar, “A note on Jensen's inequality,” *Studia Scientarium Mathematicarum Hungarica*, vol. 1, pp. 185–188, 1966.
[4] ——, “Information-type measures of difference of probability distributions and indirect observations,” *Studia Scientarium Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.
[5] ——, “On topological properties of $f-$divergences,” *Studia Scientarium Mathematicarum Hungarica*, vol. 2, pp. 329–339, 1967.
[6] I. Ibragimov and R. Z. Has'minskii, “A problem of statistical estimation in Gaussian white noise,” *Dokl. Akad. Nauk SSSR*, vol. 236, pp. 1053–1055, 1977.
[7] ——, “On estimate of the density function,” *Zap. Nauchn. Semin. LOMI*, pp. 61–85, 1980.
[8] I. A. Ibragimov and R. Z. Has'minskii, *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag, 1981.
[9] R. Z. Has'minskii, “A lower bound on the risk of nonparametric estimates of densities in the uniform metric,” *Theory Probability and Its Applications*, vol. 23, pp. 794–798, 1978.
[10] L. Birgé, “Approximation dans les espaces metriques et theorie de l'estimation,” *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 65, pp. 181–237, 1983.
[11] ——, “On estimating a density using Hellinger distance and some other strange facts,” *Probability Theory and Related Fields*, vol. 71, pp. 271–291, 1986.
[12] F. Topsøe, “Some inequalities for information divergence and related measures of discrimination,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 1602–1609, 2000.
[13] T. T. Cai, C.-H. Zhang, and H. H. Zhou, “Optimal rates of convergence for covariance matrix estimation,” *Annals of Statistics*, vol. 38, pp. 2118–2144, 2010.
[14] A. A. Gushchin, “On Fano's lemma and similar inequalities for the minimax risk,” *Theor. Probability and Math. Statist.*, vol. 67, pp. 29–41, 2003.
[15] M. D. Reid and R. C. Williamson, “Generalized Pinsker inequalities,” in *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.
[16] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
[17] T. S. Han and S. Verdú, “Generalizing the Fano inequality,” *IEEE Transactions on Information Theory*, vol. 40, pp. 1247–1251, 1994.
[18] L. Birgé, “A new bound for multiple hypothesis testing,” *IEEE Transactions on Information Theory*, vol. 51, pp. 1611–1615, 2005.
[19] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.
[20] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.
[21] A. S. Nemirovski, “Topics in nonparametric statistics,” in *Lecture on Probability Theory and Statistics, École d'Été de Probabilitiés de Saint-flour XXVIII-1998*. Berlin, Germany: Springer-Verlag, 2000, vol. 1738, Lecture Notes in Mathematics.
[22] L. Le Cam, “Convergence of estimates under dimensionality restrictions,” *Annals of Statistics*, vol. 1, pp. 38–53, 1973.
[23] R. Schneider, *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge: Cambridge Univ. Press, 1993.
[24] R. T. Rockafellar, *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press, 1970.
[25] R. Gardner, M. Kiderlen, and P. Milanfar, “Convergence of algorithms for reconstructing convex bodies and directional measures,” *Annals of Statistics*, vol. 34, pp. 1331–1374, 2006.

[26] E. M. Bronshtein, "ϵ-entropy of convex sets and functions," *Siberian Math. J.*, vol. 17, pp. 393–398, 1976.

[27] P. Massart, *Concentration inequalities and model selection. Lecture notes in Mathematics*. Berlin: Springer, 2007, vol. 1896.

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

Adityanand Guntuboyina (S'10) received the B.Stat. and M.Stat. degrees from the Indian Statistical Institute, Kolkata, in 2004 and 2006 respectively and the M.A. degree in statistics from Yale University in 2008.

He is currently a PhD candidate in the Department of Statistics at Yale University. His dissertation advisor is Professor David Pollard. His main research interests include various areas of nonparametric statistics like high-dimensional nonparametric regression and bayesian nonparametrics. He is also interested in the concentration of measure phenomena, random matrices and respondent driven sampling.