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

Adityanand Guntuboyina

***Abstract*—A new lower bound involving f-divergences between the underlying probability measures is proved for the minimax risk in estimation problems. The proof just uses the convexity of the function f and is extremely simple. Special cases and straightforward corollaries of our bound 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*—Minimax lower bounds; f-divergences; Fano's inequality; 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 $w(\rho)$ where ρ is a metric on $\mathcal{A}$ and $w : [0,\infty) \to [0,\infty)$ is a nondecreasing function. The minimax risk for this problem is defined by

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

where the infimum is over all measurable functions $\hat{\theta} : \mathcal{X} \to \mathcal{A}$ and the expectation is with respect to $X \sim P_\theta$.

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 $R \geq w(\eta/2)r$, where

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

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], [2].

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

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

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

where $w_\theta := w\left\{\theta\right\}$ and the infimum is over all estimators T taking values in F. When w is the discrete uniform probability measure, we simply write $\bar{r}$ for $\bar{r}_w$. We shall also write β_w for $1 - \bar{r}_w$ and β for $1 - \bar{r}$. The trivial inequality $r \geq \bar{r}_w$ implies that lower bounds for $\bar{r}_w$ are automatically lower bounds for r.

The main result of this paper (Theorem II.1) provides a new lower bound for $\bar{r}_w$ (or an upper bound for β_w) involving f-divergences of the probability measures $P_\theta, \theta \in F$. The f-divergences ([3]–[6]) 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 convexity of the function f and the standard result that β_w has the following exact expression:

$$\beta_w = \int_{\mathcal{X}} \max_{\theta \in F} \left\{w_\theta p_\theta(x)\right\} d\mu(x), \tag{3}$$

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 [2], [7]–[12]. 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 [13] 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. These bounds are generalizations of the usual bounds for J_f when $f(x) = x \log x$ ([2], [11], [14]). In section IV, we use these inequalities to obtain minimax lower bounds involving only global entropy conditions. One of these bounds involves the Kullback-Leibler divergence and is due to Yang and Barron [2]. The other bound involves the chi-squared divergence and is new. We use this chi-squared divergence based inequality to observe the somewhat surprising fact that global entropy conditions provide the optimal minimax lower bound even in finite dimensional cases.

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 [15] on covariance matrix estimation.

II. Main Result

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 $p_\theta, \theta \in F$. The quantity $\beta_w := 1 - \bar{r}_w$ has the exact expression given in (3).

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*

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

where $W := \int_{\mathcal{X}} w_{T(x)} dQ(x)$. The same inequality holds if we replace β_w by $1 - \bar{r}_w$ or by $1 - r$. In particular, taking w to be the uniform probability measure, we get that

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

Moreover, the same inequality also holds if $\bar{r}$ is replaced by r.

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

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

which is bounded by $W f(\beta_w/W)$ from below by Jensen's inequality. It follows similarly that

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

This completes the proof of inequality (4). The convexity of f now implies that the right hand side of (4) is non-decreasing as a function of β_w and hence (4) also holds if β_w is replaced by $1 - r$. When w is the uniform probability measure of the discrete set F, it is obvious that W equals $1/N$ and this leads to inequality (5). ∎

Inequality (5) gives an implicit lower bound for the minimax risk r. It can be turned into into an explicit lower bound. This is the content of the following corollary. We assume differentiability for convenience.

Corollary II.2. *Let $f : [0, \infty) \to \mathbb{R}$ be a differentiable convex function and let $g : [0, 1] \to \mathbb{R}$ be defined as*

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

Then for every $a \geq 1/N$, we have

$$r \geq \bar{r} \geq 1 - \frac{\inf_Q \sum_{\theta \in F} D_f(P_\theta||Q) + a g'(a) - g(a)}{g'(a)} \tag{8}$$

where the infimum is over all probability measures Q.

Proof: Fix a probability measure Q. We first note that inequality (5) gives that $\sum_{\theta \in F} D_f(P_\theta||Q) \geq g(\beta)$. The

convexity of f implies that g is also convex and hence we can write

$$g(\beta) \geq g(a) + g'(a)(\beta - a) \text{ for every } a \in [0,1].$$

Also,

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

Because g is convex, we have $g'(a) \geq g'(1/N) = 0$ for $a \geq 1/N$. Therefore, for each $a \geq 1/N$, we get that

$$\beta \leq \frac{\sum_{\theta \in F} D_f(P_\theta||Q) + ag'(a) - g(a)}{g'(a)}.$$

The proof is complete. ∎

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*

$$\inf_Q \left[D_f(P_1||Q) + D_f(P_2||Q)\right] \geq f(1+V) + f(1-V) \quad (9)$$

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 (9) *over all probability measures P_1 and P_2 with total variation distance V equals the right hand side of* (9)*.*

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 2β and $2\bar{r}$ equal $1+V$ and $1-V$ respectively. Inequality (9) is then a direct consequence of inequality (5).

The following example shows that (9) 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 (9). ∎

Remark II.1. *There exist many inequalities in the literature relating the f-divergence of probability measures to the total variation distance. We refer the reader to [16] for the sharpest results in this direction and for earlier references. Inequality* (9) *does appear to be new however.*

In the reminder 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 [13, 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. 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 [19, Theorem 2.10.1]. To obtain the usual form of Fano's inequality as used in statistics, we turn to inequality (8)*. For $a_0 := N/(2N-1) \geq 1/N$ and the function g in* (7)*, it can be checked that $g'(a_0) = N\log N$ and $a_0 g'(a_0) - g(a_0) \leq N\log 2$. It follows from inequality* (8) *that*

$$r \geq \bar{r} \geq 1 - \frac{\log 2 + \frac{1}{N}\sum_{\theta\in F} D(P_\theta||\bar{P})}{\log N}, \quad (10)$$

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 [13, Equation (24)]. Our proof is slightly simpler than Topsøe's. Topsøe [13] also explains how to use this inequality to deduce the sharp constant Pinsker's inequality: $D(P_1||P_2) \geq 2V^2$. 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 $\chi^2(P||Q) := \int p^2/q - 1$. Invoking inequality* (5)*, we get that*

$$r \geq \bar{r} \geq 1 - \frac{1}{\sqrt{N}}\sqrt{\frac{\inf_Q \sum_{\theta\in F}\chi^2(P_\theta||Q)}{N}} - \frac{1}{N}. \quad (11)$$

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

$$\inf_Q \left(\chi^2(P_1||Q) + \chi^2(P_2||Q)\right) \geq 2V^2. \quad (12)$$

The weaker inequality $\chi^2(P_1||\bar{P}) + \chi^2(P_2||\bar{P}) \geq 2V^2$ can be found in [13, 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. In this case, 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 $(\sum_{\theta\in F}\sqrt{p_\theta})^2$. Applying inequality* (5) *with*

this Q and after some simplification, we get the following minimax lower bound

$$r \geq \bar{r} \geq 1 - \frac{1}{N} - \frac{h^2}{2} - \frac{h\sqrt{2-h^2}}{\sqrt{N}},$$

where $h^2 := \sum_{\theta,\theta'} H^2(P_\theta, P_{\theta'})/N^2$.

Similar calculations can be carried out to deduce from Corollary II.3 that for every two probability measures P_1 and P_2,

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

This inequality is usually attributed to Le Cam [20].

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. Inequality* (5) *results in the lower bound*

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

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. Using inequality* (5)*, we get the following bound:*

$$\inf_Q \sum_{\theta\in F} D_f(P_\theta||Q) \geq N^l(1-\bar{r})^l - N,$$

which gives us the following inequality

$$r \geq \bar{r} \geq 1 - \left(\frac{1}{N^{l-1}} + \frac{\inf_Q \sum_{\theta\in F} D_f(P_\theta||Q)}{N^l}\right)^{1/l} \quad (13)$$

When $l = 2$, inequality (13) *results in a bound that is weaker than inequality* (11) *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

$$V \leq \sqrt{1 - \exp\left(-\inf_Q \{D(Q||P_1) + D(Q||P_2)\}\right)}. \quad (14)$$

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 (14)*, can also be found in [21, 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 the Jensen-Shannon divergence 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:

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

In the case of Kullback-Leibler divergence, this inequality has been frequently used in the literature (see for example [11] and [14]). Again, in the case of Kullback-Leibler divergence, Yang and Barron [2] provided an improved bound for J_f. They showed that for any finite set $\{Q_\alpha : \alpha \in G\}$ of probability measures, the following inequality holds

$$\frac{1}{N}\sum_{\theta\in F} D(P_\theta||\bar{P}) \leq \log|G| + \max_{\theta\in F}\min_{\alpha\in G} D(P_\theta||Q_\alpha) \quad (15)$$

An important aspect of inequality (15) 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 [2] for a detailed discussion of these issues.

In the following theorem, we shall generalize inequality (15) to f-divergences. 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 $f : [0,\infty) \to \mathbb{R}$, the following inequality holds*

$$J_f \leq \frac{1}{N}\sum_{\theta\in F}\int_{\mathcal{X}} \frac{q_{j(\theta)}}{M} f\left(\frac{Mp_\theta}{q_{j(\theta)}}\right) d\mu + \left(1 - \frac{1}{M}\right) f(0). \quad (16)$$

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 $\theta \in F$,

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

Inequality (16) now follows as a consequence of the inequality $J_f \leq \sum_{\theta \in F} D_f(P_\theta||\bar{Q})/N$. ∎

We shall now apply this result to specific f-divergences.

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* (16)*, 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 (15) *would now follow if we choose $j(\theta) := \arg\min_{\alpha \in G} D(P_\theta||Q_\alpha)$. Hence Theorem III.1 is indeed a generalization of* (15)*.*

Example III.3. *Let $f(x) = x^l - 1$ for $l > 1$. Applying inequality* (16)*, 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

$$J_f \leq M^{l-1} \left\{ \max_{\theta \in F} \min_{\alpha \in G} D_f(P_\theta||Q_\alpha) + 1 \right\} - 1. \tag{17}$$

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

$$M \left\{ \max_{\theta \in F} \min_{\alpha \in G} \chi^2(P_\theta||Q_\alpha) + 1 \right\} - 1. \tag{18}$$

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* (16)*, 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

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

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 $N(\eta)$ be a positive 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.

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

$$\frac{\log 2 + \log M_{KL}(\epsilon; \Theta) + \epsilon^2}{\log N(\eta)} \tag{19}$$

$$\frac{1}{N(\eta)} + \sqrt{\frac{(1 + \epsilon^2) M_C(\epsilon, \Theta)}{N(\eta)}} \tag{20}$$

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

$$\left\{ \frac{1}{(N(\eta))^{l-1}} + \frac{(1 + \epsilon^2)(M_l(\epsilon, \Theta))^{l-1}}{(N(\eta))^{l-1}} \right\}^{1/l}. \tag{21}$$

In the sequel, by inequality (20), we mean the inequality $R \geq \sup_{\eta > 0, \epsilon > 0} w(\eta/2)(1 - \star)$ with $\star$ representing (20). Similarly for inequalities (19) and (21).

Proof: We shall give the proof of inequality (20). 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 $R \geq w(\eta/2) r$, where r is defined as in (1). Inequality (11) 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 (18) 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 (19) is due to Yang and Barron [2, 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 (19) which applies to every estimation problem. Inequalities (20) and (21) 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 (10) 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. On the other hand, the application of Theorem IV.1 does not require the construction of such a special subset F.

Yang and Barron [2] have successfully applied inequality (19) to obtain 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 [2] for examples. In some of these examples, inequality (20) can also be applied to get optimal lower bounds. In section V, we shall apply inequality (20) to obtain a new minimax lower bound in the problem of reconstructing convex bodies from noisy support function measurements.

Although inequality (19) works well in nonparametric problems, for *not very rich* parameter spaces like finite dimensional spaces and analytical densities, inequality (19) only results in a sub-optimal lower bound, as observed by Yang and Barron [2, Page 1574]. This can be seen as follows: In typical estimation problems of a one-dimensional parameter, restricted to be in a bounded parameter space Θ, from n independent and identically distributed observations with squared error loss, one has $N(\eta) = c_1/\eta$ and $M_{KL}(\epsilon;\Theta) = c_2\sqrt{n}/\epsilon$ for positive constants c_1 and c_2. Thus, in this case, by (19), the minimax risk R_n satisfies the inequality

$$R_n \geq \sup_{\eta>0,\epsilon>0} \frac{\eta^2}{4}\left(1 - \frac{\log 2 + \log(c_2\sqrt{n}/\epsilon) + \epsilon^2}{\log(c_1/\eta)}\right).$$

This inequality is optimized when $\epsilon = 1/\sqrt{2}$ and then we get

$$R_n \geq \sup_{\eta>0} \frac{\eta^2}{4}\left(1 - \frac{\log(\sqrt{2}c_2\sqrt{n}) + 1/2 + \log 2}{\log(c_1/\eta)}\right). \quad (22)$$

We now note that when $\eta = c/\sqrt{n}$ for a constant c, the quantity inside the parantheses on the right hand side of (22) converges to 0 as n goes to ∞. This means that inequality (19) 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 (20) gives $R_n \geq c/n$ for a positive constant c. Typically, one has $M_C(\epsilon,\Theta) = c_3\sqrt{n}/\sqrt{\log(1+\epsilon^2)}$ for a positive constant c_3. Inequality (20) then gives that

$$R_n \geq \sup_{\eta>0,\epsilon>0} \frac{\eta^2}{4}\left(1 - \frac{\eta}{c_1} - \sqrt{\frac{c_3}{c_1}}\sqrt{\eta\sqrt{n}}\sqrt{\frac{1+\epsilon^2}{\sqrt{\log(1+\epsilon^2)}}}\right).$$

Taking $\epsilon = 1$ and $\eta = c_4/\sqrt{n}$, we get

$$R_n \geq \frac{c_4^2}{4n}\left(1 - \frac{c_4}{c_1\sqrt{n}} - \sqrt{\frac{2c_3c_4}{c_1\sqrt{\log 2}}}\right) \quad (23)$$

Hence by choosing c_4 small, we get that $R_n \geq c/n$ for all large n.

We have thus demonstrated that inequality (20) works even for finite dimensional parameter spaces, a scenario in which inequality (19) fails. Of course, obtaining optimal minimax rates in finite dimensional problems is quite easy in most situations. Often even two point priors give the optimal rate. But the point here is that even in finite dimensional situations, global metric entropy conditions are enough for obtaining optimal minimax lower bounds. This goes against the usual claim that one needs more than the global entropy condition in order to come up with optimal lower bounds for parametric problems [2, Page 1574].

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 (20). 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 [22, 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 [23] (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 $u_1, \ldots, u_n$. More precisely, their problem was to estimate K from observations $Y_1, \ldots, Y_n$ drawn according to the model $Y_i = h_K(u_i) + \xi_i, i = 1, \ldots, n$ where $\xi_1, \ldots, \xi_n$ are independent random variables with mean zero and finite variance. They constructed a convex body (estimator) $\hat{K}_n = \hat{K}_n(Y_1, \ldots, Y_n)$ having the property that, for *nice* sequences $\{u_i, i \geq 1\}$, the L^2 norm $||h_K - h_{\hat{K}_n}||_2$ (see (24) 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 [23, Theorem 6.2]).

We shall show here that when the errors $\xi_1, \ldots, \xi_n$ are gaussian, it is impossible in a 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 [23] 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 $R > 0$, let $\mathcal{K}^d(R)$ denote the set of all convex bodies in $\mathbb{R}^d$ that are contained in the closed ball of radius R centered at the origin. 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 $Y_1, \ldots, Y_n$. The loss functions that we shall use are the L^p norms for $p \in [1, \infty]$ defined by

$$||h_K - \hat{h}||_p := \left(\int_{S^{d-1}} (h_K(u) - \hat{h}(u))^p du\right)^{1/p} \quad (24)$$

for $p \in [1,\infty)$ and $||h_K - \hat{h}||_\infty := \sup_{u \in S^{d-1}} |h_K(u) - \hat{h}(u)|$. We shall consider the minimax risk of the problem of estimating h_K from $Y_1, \ldots, Y_n$ when K is assumed to belong to $\mathcal{K}^d(R)$ i.e., we are interested in the following quantity

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

The following is the main theorem of this section.

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

$$r_n(p,R) \geq c\sigma^{4/(d+3)} R^{(d-1)/(d+3)} n^{-2/(d+3)}, \qquad (25)$$

for a constant c that is independent of n.

Remark V.1. *Gardner, Kiderlen and Milanfar [23] showed that the least squares estimator converges at the rate given by the right hand side of* (25) *for dimensions $d = 2, 3, 4$. Thus the lower bound given by* (25) *is optimal for dimensions $d = 2, 3, 4$.*

We shall use inequality (20) to prove (25). First, let us put the support function estimation problem in the general estimation setting of the last section. Let $\Theta := \{h_K : K \in \mathcal{K}^d(R)\}$ 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.

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

In order to apply inequality (20), 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 $\mathcal{K}^d(R)$ under the L^p norm. When $p = \infty$, Bronshtein [24, Theorem 4 and Remark 1] proved that there exist positive constants c' and η_0 depending only on d such that $\exp\left(c'(\eta/R)^{(1-d)/2}\right)$ 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 $\mathcal{K}^d(R)$.

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 Varshamov-Gilbert's lemma (see for example [25, Lemma 4.7]), we provide, in Theorem VII.1, a proof of this fact. Therefore from Theorem VII.1, we can take

$$\log N(\eta) = c' \left(\frac{R}{\eta}\right)^{(d-1)/2} \quad \text{for } \eta \leq \eta_0, \qquad (26)$$

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

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

$$||f - g||_\infty \leq \epsilon' \Longrightarrow \chi^2(P_f||P_g) \leq \epsilon^2. \qquad (27)$$

where $\epsilon' := \sigma\sqrt{\log(1+\epsilon^2)}/\sqrt{n}$. Let $W_{\epsilon'}$ be the smallest W for which there exist sets $K_1, \ldots, K_W$ in $\mathcal{K}^d(R)$ having the property that for every set $K \in \mathcal{K}^d(R)$, 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 (27) that $M_C(\epsilon, \Theta)$ can be taken to be a number larger than $W_{\epsilon'}$. Bronshtein [24, Theorem 3 and Remark 1] showed that there exist positive constants c'' and ϵ_0 depending only on d such that

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

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

$$\log M_C(\epsilon, \Theta) = c'' \left(\frac{R\sqrt{n}}{\sigma\sqrt{\log(1+\epsilon^2)}}\right)^{(d-1)/2}. \qquad (28)$$

We are now ready to prove inequality (25). We shall define two quantities

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

and

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

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 $u(n)/\sqrt{n} \to 0$. It can be easily checked that the quantity

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

equals

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

Inequality (25) now follows if we choose c sufficiently small, say such that $c^{(d-1)/2} = c'/(2 + 2c'')$.

VI. A COVARIANCE MATRIX ESTIMATION EXAMPLE

In the previous section, we have used the global minimax lower bound (20). However, in some situations, the global entropy numbers might be difficult to bound. In such cases, inequalities (19) and (20) are, of course, not applicable and we are unaware of the use of inequality (15) in conjuction with Fano's inequality (10) in the literature. The standard examples use (10) with the bound $J_f \leq \min_{\theta,\theta' \in F} D(P_\theta||P_{\theta'})$ while the examples in [2] 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 [15] can also be proved using inequalities (10) and (15).

Cai, Zhang and Zhou [15] considered n independent $p \times 1$ random vectors $X_1, \ldots, X_n$ 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 [15] showed that when p is large compared to n, it is impossible to estimate Σ from $X_1, \ldots, X_n$ in the spectral norm at a rate faster than $n^{-\alpha/(2\alpha+1)}$. More precisely, they showed that when $p \geq Cn^{1/(2\alpha+1)}$,

$$R_n(\alpha) := \inf_{\hat{\Sigma}} \sup_{\Sigma \in \Theta} \mathbb{E}_\Sigma ||\hat{\Sigma} - \Sigma|| \geq c\, n^{-\alpha/(2\alpha+1)} \tag{29}$$

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 [15] used Assouad's lemma for the proof of the inequality (29). We shall use inequalities (10) and (15). Moreover, the choice of the finite subset F that we use is different from the one used in [15, Equation (17)]. This makes our approach different from the general method, due to Yu [1], of replacing Assouad's lemma by Fano's inequality.

Throughout, K denotes a constant depending 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 $a_{ij} = 1/(K|i-j|^{\alpha+1})$ for $i \neq j$. For K 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 $\tau_1, \ldots, \tau_k$. Clearly, $A(\tau) \in \Theta$ for all $\tau \in \{0,1\}^k$. We shall need the following two lemmas in order to prove inequality (29).

Lemma VI.1. *For $\tau, \tau' \in \{0,1\}^k, \tau \neq \tau'$, the following inequality holds*

$$||A(\tau) - A(\tau')|| \geq \frac{1}{Kk^\alpha} \sqrt{\frac{\xi(\tau,\tau')}{k}}, \tag{30}$$

where $\xi(\tau,\tau') := \sum_{r=1}^k \{\tau_r \neq \tau_r'\}$ 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 $u = (u_1, \ldots, u_k)^T$. Moreover $u_r = \sum_{s=1}^k (\tau_r - \tau_r') a_{r,k+s}$ and hence

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

Therefore,

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

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

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

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

Proof: The key is to note that one has the inequality $D\left(N(0, A(\tau)) || N(0, A(\tau'))\right) \leq K ||A(\tau) - A(\tau')||_F^2$, 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 [15, Proof of Lemma 6]. We can now bound

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

The proof is complete. ∎

Varshamov-Gilbert's lemma (see for example [25, Lemma 4.7]) asserts the existence of a subset W of $\{0,1\}^k$ with $|W| \geq \exp(k/8)$ such that $\xi(\tau,\tau') \geq k/4$ for all $\tau, \tau' \in W$ with $\tau \neq \tau'$. Let $F := \{A(\tau) : \tau \in W\}$. From inequality (10) and Lemma VI.1, we get that

$$R_n(\alpha) \geq \frac{1}{K} \frac{1}{k^\alpha} \left(1 - \frac{\log 2 + \frac{1}{|W|} \sum_{A \in F} D(P_A || \bar{P})}{k/8}\right) \tag{31}$$

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 $N(0, A(0, \ldots, 0, t_1, \ldots, t_{k-m+1}))$ probability measure. Applying inequality (15), 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

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

Using the above in (31), we get

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

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 (29), it is enough to take $k - m = n^{1/(2\alpha+1)}$ and $k = 4Kn^{1/(2\alpha+1)}$. The condition $2k \le p$ will be satisfied if $p \ge Cn^{1/(2\alpha+1)}$ for a large enough C. It is elementary to check that with these choices of k and m, inequality (29) 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 $N(\eta; p, R)$ of $\mathcal{K}^d(R)$ under the L^p metric is at least $\exp\left(c(\eta/R)^{(1-d)/2}\right)$ for a positive c and sufficiently small η. This means that there exist at least $\exp\left(c(\eta/R)^{(1-d)/2}\right)$ sets in $\mathcal{K}^d(R)$ separated by at least η in the L^p metric. This result was needed in the proof of Theorem V.1. Bronshtein [24, 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 $\eta \le \eta_0$, we have*

$$N(\eta; p, R) \ge \exp\left(C\left(\frac{R}{\eta}\right)^{(d-1)/2}\right). \tag{32}$$

Proof: Observe that by scaling, it is enough to prove for the case $R = 1$. We loosely follow Bronshtein [24, Proof of Theorem 4]. We write $d(x,y)$ for the Euclidean distance between two points x and y in $\mathbb{R}^d$. 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 x and y are two points in S^{d-1} with $d(x,y) > 2\sqrt{2}\sqrt{\epsilon}$, then the sets T_x and T_y are disjoint.

By standard results, there exist positive constants C_1, depending only on d, and ϵ_0 such that for every $\epsilon \le \epsilon_0$, there exist $N \ge C_1(\sqrt{\epsilon})^{1-d}$ points $x_1, \ldots, x_N$ in S^{d-1} such that $d(x_i, x_j) > 2\sqrt{2}\sqrt{\epsilon}$ if $i \ne j$. From now on, we assume that $\epsilon \le \epsilon_0$. We then consider a mapping $\Phi : \{0,1\}^N \to \mathcal{K}^d(B)$, which is defined, for $w = (w_1, \ldots, w_N) \in \{0,1\}^N$, by

$$\Phi(w) := B \cap D_1(w_1) \cap D_2(w_2) \cap \cdots \cap D_N(w_N),$$

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

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

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

$$\delta_p\left(\Phi(w), \Phi(w')\right) \ge C_2\epsilon\left(\sqrt{\epsilon}\right)^{(d-1)/p}\left(\rho(w,w')\right)^{1/p}, \tag{33}$$

where C_2 depends only on d and p. The claim would be proved later. Assuming it is true, we can apply Varshamov-Gilbert's lemma (see for example [25, Lemma 4.7]). This lemma asserts the existence of a subset W of $\{0,1\}^N$ with $|W| \ge \exp(N/8)$ such that $\rho(w,w') \ge N/4$ for all $w, w' \in W$ with $w \ne w'$. Because $N \ge C_1(\sqrt{\epsilon})^{1-d}$, we get from (33) that for all $w, w' \in W$ with $w \ne w'$, we have

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

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

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

The constant C_4 only depends on d and p thereby proving (32).

It remains to prove the claim (33). Fix a point $x \in S^{d-1}$ and $\epsilon \in (0,1)$. We first observe that it is enough to prove that

$$\left(\delta_p(A_x, T_x)\right)^p \ge C_5\epsilon^p\left(\sqrt{\epsilon}\right)^{d-1}, \tag{34}$$

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 $w, w' \in W$ with $w \ne w'$, we can write

$$\left(\delta_p\left(\Phi(w), \Phi(w')\right)\right)^p = \sum_{i \in I}\left(\delta_p(A_{x_i}, T_{x_i})\right)^p, \tag{35}$$

where $I := \{1 \le i \le N : w_i \ne w_i'\}$. The equality (35) is a consequence of the fact that the points $x_1, \ldots, x_N$ are chosen so that $T_{x_1}, \ldots, T_{x_N}$ are disjoint.

We shall now prove the inequality (34) 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

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

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

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

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

above, we get that

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

We shall show that $\sin\beta \geq (\sqrt{\epsilon})/(2\sqrt{2})$ which would prove (34). 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 would give $\sin\beta \geq (\sqrt{\epsilon})/(2\sqrt{2})$. 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 (20) 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 (10) and (15) to give a different proof of a recent lower bound for covariance matrix estimation due to Cai, Zhang and Zhou [15].

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.

References

[1] 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.

[2] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, vol. 27, pp. 1564–1599, 1999.

[3] 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.

[4] I. Csiszar, "A note on Jensen's inequality," *Studia Scientarium Mathematicarum Hungarica*, vol. 1, pp. 185–188, 1966.

[5] ——, "Information-type measures of difference of probability distributions and indirect observations," *Studia Scientarium Mathematicarum Hungarica*, vol. 2, pp. 299–318, 1967.

[6] ——, "On topological properties of $f-$divergences," *Studia Scientarium Mathematicarum Hungarica*, vol. 2, pp. 329–339, 1967.

[7] 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.

[8] ——, "On estimate of the density function," *Zap. Nauchn. Semin. LOMI*, pp. 61–85, 1980.

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

[10] 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.

[11] L. Birgé, "Approximation dans les espaces metriques et theorie de l'estimation," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 65, pp. 181–237, 1983.

[12] ——, "On estimating a density using Hellinger distance and some other strange facts," *Probability Theory and Related Fields*, vol. 71, pp. 271–291, 1986.

[13] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1602–1609, 2000.

[14] 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.

[15] T. T. Cai, C. Zhang, and H. H. Zhou, "Optimal rates of convergence for covariance matrix estimation," 2009, (To appear in the Annals of Statistics).

[16] M. D. Reid and R. C. Williamson, "Generalized Pinsker inequalities," in *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

[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] T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[20] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag, 1986.

[21] A. Tsybakov, *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.

[22] R. T. Rockafellar, *Convex Analysis*. Princeton, New Jersey: Princeton Univ. Press, 1970.

[23] 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.

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

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