On Bayes Risk Lower Bounds

Xi Chen
New York University
xichen@nyu.edu

Adityanand Guntuboyina
UC Berkeley
aditya@stat.berkeley.edu

Yuchen Zhang
UC Berkeley
yuczhang@eecs.berkeley.edu

Abstract

This paper provides a general technique to lower bound the Bayes risk for arbitrary loss functions and prior distributions in the standard abstract decision theoretic setting. A lower bound on the Bayes risk not only serves as a lower bound on the minimax risk but also characterizes the fundamental limitations of the statistical difficulty of a decision problem under a given prior. Our bounds are based on the notion of \( f \)-informativity (Csiszár, 1972) of the underlying class of probability measures and the prior. Application of our bounds requires upper bounds on the \( f \)-informativity and we derive new upper bounds on \( f \)-informativity for a class of \( f \) functions which lead to tight Bayes risk lower bounds. Our technique leads to generalizations of a variety of classical minimax bounds. As applications, we present Bayes risk lower bounds for several concrete estimation problems, including Gaussian location models, Bayesian Lasso, generalized linear models and principle component analysis for spiked covariance models.

1 Introduction

Consider a standard decision-theoretic setting where \( \Theta \) and \( \mathcal{A} \) are the parameter and action spaces respectively and \( L(\theta, a) : \Theta \times \mathcal{A} \mapsto [0, \infty) \) is a non-negative loss function. We observe data \( X \) taking values in a sample space \( \mathcal{X} \). The distribution of \( X \) depends on the unknown parameter \( \theta \) and is denoted by \( P_\theta \) (\( P_\theta \) is a probability measure on \( \mathcal{X} \)). The class of probability measures \( \{P_\theta : \theta \in \Theta\} \) is denoted by \( P \). (Nonrandomized) decision rules are functions mapping \( \mathcal{X} \) to \( \mathcal{A} \). The risk of a decision rule \( \mathcal{d} \) is defined by \( \mathbb{E}_\theta L(\theta, \mathcal{d}(X)) \), where \( \mathbb{E}_\theta \) denotes expectation taken under the assumption that \( X \) is distributed according to \( P_\theta \). The minimax risk over the parameter space \( \Theta \) with respect to the loss \( L \) is defined by,

\[
R_{\text{minimax}}(L; \Theta) := \inf_{\mathcal{d}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \mathcal{d}(X)).
\]  

where the infimum is over all decision rules \( \mathcal{d} \).

The minimax risk plays an important role in statistical decision theory. It acts as a benchmark for the risk of any decision rule. Indeed, a common way of evaluating the performance of a given decision rule is to compare its maximum possible risk to the minimax risk of the problem. Unfortunately, it is difficult, in most problems, to determine the minimax risk exactly. One therefore attempts to obtain good lower bounds instead. Such a minimax lower bound that is tight up to a universal constant factor is known as the minimax rate and is believed to capture the fundamental difficulty of the decision problem in terms of the worst-case performance. The ideas behind minimax lower bounds are now well-understood and are summarized, for example, in Guntuboyina (2011a); Tsybakov (2010); Yu (1997).

A frequent criticism of the practice of using the minimax risk as the benchmark in decision-theoretic problems is the fact that the minimax risk can often be too pessimistic since it looks for the best decision to guard against the worst possible risk over the entire parameter space. When a
prior on the parameter $\theta$ is available, a much more natural benchmark would be the Bayes risk for the problem under the aforementioned prior. For a proper prior $w$ (i.e., $w$ is a probability measure on $\Theta$), the Bayes risk with respect to $w$ is defined by

$$R_{\text{Bayes}}(w, L; \Theta) := \inf_\theta \int_{\Theta} E_\theta L(\theta, \hat{a}(X)) w(d\theta).$$

(2)

The Bayes risk is always a lower bound of the minimax risk and, depending on the strength of the prior $w$, can be a lot smaller than the minimax risk.

Like in the case of the minimax risk, it is generally difficult to determine the Bayes risk exactly (except in special situations such as when $w$ is a conjugate prior). This paper focuses on obtaining lower bounds for the Bayes risk. Good lower bounds for the Bayes risk are important for several reasons: (a) they provide benchmarks for evaluating decision rules of interest under the prior, (b) they yield the “Bayes rate”, which, similar to the minimax rate, characterizes the fundamental difficulty and limitations of the decision theoretic problem under the prior and sheds light on the relationship between different model parameters (e.g., parameters of the prior, dimensionality of $\Theta$, number of samples etc.), and, (c) they automatically provide lower bounds for the minimax risk.

Before describing our results, let us briefly review existing results on lower bounds for the Bayes risk. To the best of our knowledge, we are not aware of any paper which deals with this problem for arbitrary priors in the abstract decision theoretic setting described above. Quite a few results exist however in the special case of regular finite dimensional estimation problems under (weighted/truncated) quadratic losses. The first results of this kind were established by Van Trees (1968) and Borovkov and Sakhanienko (1980) with extensions by Brown (1993); Brown and Gajek (1990); Sato and Akahira (1996); Takada (1999). A few additional papers dealt with even more specialized problems e.g., estimating Gaussian variance (Vidakovic and DasGupta, 1995), Gaussian white noise model (Brown and Liu, 1993) and scale models (Gajek and Kaluszka, 1994).

The papers on minimax lower bounds also contain some Bayes risk lower bounds because the main strategy for obtaining lower bounds for the minimax risk is to bound it from below by an appropriate Bayes risk. However, these Bayes risk bounds are only applicable to certain very special and unrealistic discrete priors.

Our aim here is to provide a general technique for obtaining Bayes risk lower bounds for arbitrary priors in the abstract setting described previously. Our intention is to obtain bounds that are tight up to constant multiplicative factors. Determining the correct constants in these bounds is intractable because of the generality considered here.

The rest of the paper is organized in the following way. After setting up the notation in Section 2, we prove our main result, Theorem 3.2, in Section 3 which gives a lower bound for $R_{\text{Bayes}}(w, L; \Theta)$ for arbitrary $w$, $L$ and $\Theta$. Theorem 3.2 states that

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{a \in A} w(B_t(a, L)) < T(w, \mathcal{P}) \right\}. \quad (3)$$

Here

$$B_t(a, L) := \{ \theta \in \Theta : L(\theta, a) < t \} \quad \text{for } a \in A \text{ and } t > 0$$

and $T(w, \mathcal{P})$ is a $[0, 1]$-valued quantity that only depends on the prior $w$ and the family of distributions $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$ (more details on $T(w, \mathcal{P})$ are described below while its definition is given in (12)). In words, the right hand side of (3) equals that the largest possible $t$ such that the maximum prior mass of any $t$-“ball”, $\max_{a \in A} B_t(a, L)$, is less than $T(w, \mathcal{P})$. A nice feature of our lower bound is that the roles of the loss function $L$ and the family of probability measures $\mathcal{P}$ occur separately on two sides of the inequality in (3).

Our proof of Theorem 3.2 is based on a reduction to the case when the loss function $L$ is zero-one valued. Bayes risk bounds in this special case can be proved by the data processing inequality from information theory and this is the content of our Theorem 3.3. Theorem 3.3, which appears to be new, has a variety of interesting connections to existing bounds in the minimax literature such as the classical Fano inequality (Cover and Thomas, 2006; Han and Verdú, 1994), generalized Fano inequalities (Birgé, 2005; Duchi and Wainwright, 2013; Guntuboyina, 2011b; Gushchin, 2003) and
often-used inequalities due to Le Cam (1973) and Assouad (1983). These connections are described in Section 4.

For the application of Theorem 3.2 to actual problems, we need to be able to control the terms \( \max_{a \in A} \psi(w(B_1(a, L))) \) and \( T(w, \mathcal{P}) \). The prior mass \( w(B_1(a, L)) \) can usually be controlled by volumetric arguments, but the computation of \( T(w, \mathcal{P}) \) is more involved. As will be clear from (12), \( T(w, \mathcal{P}) \) is defined in terms of the \( f \)-informativity (Csiszár, 1972) of the class of probability measures \( \mathcal{P} \) and the prior \( w \) where \( f \) ranges over convex functions defined on \((0, \infty)\). The \( f \)-informativity, which is further defined in terms of \( f \)-divergence, is a generalization of the notion of mutual information which corresponds to \( f(x) = x \log x \). We briefly review the notions of \( f \)-divergence and \( f \)-informativity in Section 2. It turns out that for applications of Theorem 3.2, one needs to bound from above the \( f \)-informativity of \( \mathcal{P} \) and \( w \). Section 5 is dedicated to upper bounds for \( f \)-informativities for many convex functions \( f \) of interest. Tightest general upper bounds for the mutual information, to the best of our knowledge, are proved in Haussler and Opper (1997) and, in Section 5, we provide inequalities analogous to theirs for more general \( f \)-informativities.

In Section 6, we apply our results to yield lower bounds for Bayes risks in concrete problems. We consider a variety of estimation settings such as the gaussian location model, Bayesian Lasso (see, e.g., Park and Casella (2008)), Bayesian generalized linear model and principle component analysis for the spiked covariance model. We demonstrate that the Bayes risk in these problems is bounded from below by the minimum of the minimax rate and another term determined by the prior. Our lower bounds imply that when the prior information is weak, the Bayes risk matches the minimax rate; while when the prior information is strong, the Bayes risk can be much smaller than the minimax rate.

2 Preliminaries and Notation

The notions of \( f \)-divergence (Ali and Silvey, 1966; Csiszár, 1963) and \( f \)-informativity (Csiszár, 1972) will be heavily used in the sequel so we quickly review them here. Let \( \mathcal{C} \) denote the class of all convex functions \( f : (0, \infty) \to \mathbb{R} \) which satisfy \( f(1) = 0 \). Note that because of convexity, the limits \( f(0) := \lim_{x \downarrow 0} f(x) \) and \( f'(\infty) := \lim_{x \uparrow \infty} f(x)/x \) exist (even though they may be \( +\infty \)) for each \( f \in \mathcal{C} \). Each function \( f \in \mathcal{C} \) defines a divergence between probability measures which is referred to as \( f \)-divergence. For two probability measures \( P \) and \( Q \) on a sample space having densities \( p \) and \( q \) with respect to a common measure \( \mu \), the \( f \)-divergence \( D_f(P||Q) \) between \( P \) and \( Q \) is defined as follows:

\[
D_f(P||Q) := \int f \left( \frac{p}{q} \right) dQ + f'(\infty)P\{q = 0\}.
\]

Certain functions in \( \mathcal{C} \) correspond to divergences that are more commonly used in mathematical statistics. These are the power divergences which correspond to the functions \( f_\alpha, \alpha \in \mathbb{R} \) defined by

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

For power divergences, one has the identity

\[
D_{f_\alpha}(P||Q) = D_{f_{1-\alpha}}(P||Q) \quad \text{for all } \alpha \in \mathbb{R}.
\]

Popular examples of power divergences include:

1. Kullback-Leibler (KL) divergence: \( \alpha = 1 \), \( D_{f_1}(P||Q) = \int p \log(p/q) \) if \( P \) is absolutely continuous with respect to \( Q \) (and it is infinite if \( P \) is not absolutely continuous with respect to \( Q \)). We denote the KL divergence by simply \( D(P||Q) \) as opposed to \( D_{f_1}(P||Q) \).

2. Chi-squared divergence: \( \alpha = 2 \), \( D_{f_2}(P||Q) = \int (p^2/q) - 1 \) if \( P \) is absolutely continuous with respect to \( Q \) (and it is infinite if \( P \) is not absolutely continuous with respect to \( Q \)). We denote the chi-squared divergence by simply \( \chi^2(P||Q) \) as opposed to \( D_{f_2}(P||Q) \).
3. When $\alpha = 1/2$, one has $D_{f_{1/2}}(P\|Q) = 1 - \int \sqrt{\theta} \, d\theta$ which is the half of the square of the Hellinger distance. That is, $D_{f_{1/2}}(P\|Q) = H^2(P\|Q)/2$, where $H(P\|Q) = \left(\int (p - q)^2\right)^{1/2}$ is the Hellinger distance between $P$ and $Q$.

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

One of the most important properties of $f$-divergences is the “data processing inequality” (Csiszár, 1967) which states the following: let $\mathcal{X}$ and $\mathcal{Y}$ be two measurable spaces and let $\Gamma : \mathcal{X} \rightarrow \mathcal{Y}$ be measurable. Then for every $f \in \mathcal{C}$ and every pair of probability measures $P$ and $Q$ on $\mathcal{X}$, we have

$$D_f(PT^{-1}\|Q\Gamma^{-1}) \leq D_f(P\|Q).$$

We next define the notion of $f$-informativity. Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family of probability measures on a space $\mathcal{X}$ and $w$ be a probability measure on $\Theta$. For each $f \in \mathcal{C}$, the $f$-informativity, $I_f(w, \mathcal{P})$, is defined as

$$I_f(w, \mathcal{P}) = \inf_Q \int D_f(P_\theta\|Q)w(d\theta),$$

where the infimum is taken over all possible probability measures $Q$ on $\mathcal{X}$. Often the class of probability measures $\mathcal{P}$ will be clear from the context and we will denote $I_f(w, \mathcal{P})$ by simply $I_f(w)$. When $f(x) = x \log x$ so that the corresponding $f$-divergence is the KL divergence, the $f$-informativity equals the mutual information between the random variables $\theta$ and $X$ where $\theta \sim w$ and $X|\theta \sim P_\theta$ which is usually denoted by $I(\theta; X)$.

The following are additional notations and definitions that we use for the rest of the paper:

1. Recall, from (1) and (2), the definitions of minimax and Bayes risks. When the parameter space $\Theta$ is clear from the context, we drop the argument $\Theta$. Whenever the prior $w$ and the loss $L$ are also clear from the context, we denote the Bayes risk simply by $R$ and minimax risk by $R_{\min\max}$.

2. Throughout the paper, we assume that all probability measures $P_\theta$ and $Q$ have densities with respect to a common dominating measure $\mu$ on $\mathcal{X}$ and we denote these densities by $p_\theta$ and $q$ respectively.

3. Recall the definition of $B_t(a, L)$ from (4). We suppress the argument $L$ in $B_t(a, L)$ (i.e., denoted by $B_t(a)$) when $L$ is clear from the context.

4. We need certain notions of covering numbers. For a given $f$-divergence and a subset $S \subset \Theta$, let $M_f(\epsilon, S)$ denote any upper bound on the smallest number $M$ for which there exist probability measures $Q_1, \ldots, Q_M$ that form an $\epsilon^2$-cover of $\{P_\theta, \theta \in S\}$ under the $f$-divergence i.e.,

$$\sup_{w \in M} \min_{1 \leq j \leq M} D_f(P_\theta\|Q_j) \leq \epsilon^2.$$  

For notational simplicity, we write $M_{KL}(\epsilon, S)$ when $f(x) = x \log x$ and $M_{C}(\epsilon, S)$ when $f(x) = x^2 - 1$. Also, we write $M_{\alpha}(\epsilon, S)$ when $f = f_\alpha$ for other $\alpha \in \mathbb{R}$.

5. For a convex body $B$ in $d$-dimensional Euclidean space, let $\text{Vol}(B)$ be the volume of $B$ with respect to Lebesgue measure.

6. For a vector $x = (x_1, \ldots, x_d)$ and a real number $p \geq 1$, denote by $\|x\|_p$ the $\ell_p$-norm of $x$. In particular, $\|x\|_2$ will denote the Euclidean norm of $x$.

7. For a matrix $X$, let $\lambda_{\max}(X)$ and $\lambda_{\min}(X)$ be the largest and smallest eigenvalues of $X$ and $\|X\|_F$ be the Frobenius norm of $X$.

8. Throughout the paper, we use $I_d$ to denote the $d \times d$ identity matrix.

9. For any sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n \gtrless b_n$ if $a_n \geq c \, b_n$ holds for all $n$ and some absolute constant $c > 0$; $a_n \lesssim b_n$ if $a_n \leq c \, b_n$ holds for all $n$ and some $c > 0$; and $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold.

10. $I(A)$ denotes the indicator function which takes the value 1 when $A$ is true and 0 otherwise.

11. We use $C$, $c$, etc. to denote generic constants whose values might change from place to place.
3 Bayes Risk Lower Bounds

The goal of this section is to state and prove Theorem 3.2 which is the main result of this paper. Recall the standard decision theoretic setting and the definition of minimax risk described in the introduction. Also recall the definition of $f$-informativity $I_f(w, \mathcal{P})$ from (7). We need some more notations to state the theorem. For each $f \in \mathcal{C}$, let $\phi_f : [0, 1]^2 \rightarrow \mathbb{R}$ be the function given by

$$\phi_f(a, b) := bf\left(\frac{a}{b}\right) + (1 - b)f\left(\frac{1 - a}{1 - b}\right).$$

(9)

The convexity of $f$ implies some monotonicity properties of $\phi_f$. These are stated in the following lemma which is proved in the Appendix.

**Lemma 3.1.** For each $f \in \mathcal{C}$, for every fixed $b$, the map $g(a) : a \mapsto \phi_f(a, b)$ is non-increasing for $a \in [0, b]$ and $g(a)$ is convex in $a$. For every fixed $a$, the map $h(b) : b \mapsto \phi_f(a, b)$ is non-decreasing for $b \in [a, 1]$.

We further define $u_f : (0, \infty) \mapsto [1/2, 1]$ by

$$u_f(x) := \inf \left\{ 1/2 \leq b \leq 1 : \phi_f(1/2, b) > x \right\}$$

(10)

and if $\phi(1/2, b) \leq x$ for every $b \in [1/2, 1]$, then we take $u_f(x)$ to be 1.

We are now ready to state the main theorem of this paper.

**Theorem 3.2.** The following lower bound for the Bayes risk holds for every $w, L, \Theta$ and $\mathcal{P}$,

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{u \in A} w(B_t(a, L)) < T(w, \mathcal{P}) \right\},$$

(11)

where

$$T(w, \mathcal{P}) := 1 - \inf_{f \in \mathcal{C}} u_f(I_f(w, \mathcal{P})).$$

(12)

and $B_t(a, L)$ is defined in (4).

**Remark 3.1.** Theorem 3.2 can be restated for each $f \in \mathcal{C}$ as opposed to putting informativities for all $f \in \mathcal{C}$ together in an infimum. Moreover the $f$-informativity $I_f(w, \mathcal{P})$ can be replaced by any upper bound since $u_f(x)$ is non-decreasing in $x$. This yields the following equivalent version of Theorem 3.2: For every $f \in \mathcal{C}$ and $\tilde{I} \geq I_f(w, \mathcal{P})$, we have

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{u \in A} w(B_t(a, L)) < 1 - u_f(\tilde{I}) \right\}.$$ 

(13)

This form of Theorem 3.2 is more convenient to work with in applications. Note that because (13) holds for every $f \in \mathcal{C}$, one can choose $f \in \mathcal{C}$ for which the right hand side of (13) is large or for which getting an upper bound for $I_f(w, \mathcal{P})$ is manageable.

The first step in the proof of Theorem 3.2 is a reduction to the case of zero-one valued loss functions. For every $t > 0$, define the loss function $L_t(\theta, a) = I\{L(\theta, a) \geq t\}$. It is obvious that $L \geq tL_t$ and thus $R_{\text{Bayes}}(w, L) \geq tR_{\text{Bayes}}(w, L_t)$. Therefore, lower bounds for $R_{\text{Bayes}}(w, L_t)$ (note that $L_t$ is zero-one valued) automatically yield lower bounds for $R_{\text{Bayes}}(w, L)$. Theorem 3.3 below provides an implicit lower bound for $R_{\text{Bayes}}(w, L)$ when $L$ is zero-one valued. We first state and prove Theorem 3.3 after which we provide the proof of Theorem 3.2.

**Theorem 3.3.** Suppose that the loss function $L$ is zero-one valued. The following inequality holds with $R = R_{\text{Bayes}}(w, L; \Theta)$ for every $f \in \mathcal{C}$ and every probability measure $Q$ on $X$:

$$\int_{\Theta} D_f(P_\theta || Q) w(d\theta) \geq \phi_f(R, R_Q).$$

(14)

Here $R_Q$ is given by

$$R_Q := \int_X \int_{\Theta} L(\theta, \vartheta_w(x)) w(d\theta) Q(dx)$$

(15)

where $\vartheta_w$ denotes the Bayes decision rule for the prior $w$, i.e., $\vartheta_w$ minimizes $\int_{\Theta} E_{\vartheta} L(\theta, \vartheta(X)) w(d\theta)$ over all decision rules $\vartheta$. 


Remark 3.2. Theorem 3.3 is new although its special case corresponding to $\Theta = \mathcal{A} = \{1, \ldots, N\}$ and $L(\theta, a) := \mathbb{I}\{\theta \neq a\}$ is known and is due to Gushchin (2003) (it also appears in Guntuboyina (2011b)). Note that in this special case $R_Q = 1 - (1/N)$.

Remark 3.3. Because of the inequality $R \leq R_Q$ (proved in Lemma 7.1 in the Appendix) and Lemma 3.1, it should be clear that (14) provides an implicit lower bound for $R = R_{Bayes}(w; L; \Theta)$. This implicit lower bound can be converted into an explicit bound in the following way. First note that $r \mapsto \phi_f(r, R_Q)$ is convex (Lemma 3.1) and thus

$$
\phi_f(R, R_Q) \geq \phi_f(r, R_Q) + \phi_f'(r, R_Q)(R - r)
$$

for every $0 < r \leq R_Q$ where $\phi_f'(r, R_Q)$ denotes the left derivative of $x \mapsto \phi_f(x, R_Q)$ at $x = r$. The monotonicity of $\phi_f(r, R_Q)$ in $r$ gives $\phi_f'(r, R_Q) \leq 0$ and thus we get for every $0 < r \leq R_Q$,

$$
R_{Bayes}(w; L) \geq r + \frac{\int_0^1 D_f(P_0||Q)p(w)(d\theta) - \phi_f(r, R_Q)}{\phi_f'(r, R_Q)}.
$$

(16)

The proof of Theorem 3.3 is given next. It is based on the data processing inequality (6).

Proof of Theorem 3.3. Let $P$ denote the joint distribution of $\theta$ and $X$ (i.e., $\theta \sim w$ and $X|\theta \sim P_\theta$) and note that the Bayes risk, $R$, can be written in terms of the Bayes rule $\mathcal{d}_w$ as $R = E_\mathcal{d}_w L(\theta, \mathcal{d}_w(x))$. Also let $Q$ denote the joint distribution of $\theta$ and $X$ under which they are independently distributed according to $\theta \sim w$ and $X \sim Q$ respectively and thus $R_Q = E_{\mathcal{d}_w} L(\theta, \mathcal{d}_w(X))$.

Because the loss function is zero-one valued, the function $\Gamma(\theta, x) := L(\theta, \mathcal{d}_w(x))$ maps $\Theta \times X$ into $\{0, 1\}$. Our strategy is to fix $f \in \mathcal{C}$ and apply the data processing inequality (6) to the probability measures $P, Q$ and the mapping $\Gamma$. This gives

$$
D_f(P||Q) \geq D_f(P \Gamma^{-1}||Q \Gamma^{-1}).
$$

(17)

It is easy to see that $D_f(P||Q) = \int_\Theta D_f(P_\theta||Q)p(w)(d\theta)$. Also because $L$ is zero-one valued, we have

$$
\Gamma^{-1}\{1\} = \int \Gamma dP = E_{\mathcal{d}_w} L(\theta, \mathcal{d}_w(X)) = R.
$$

Similarly $Q \Gamma^{-1}\{1\} = R_Q$. Inequality (17) therefore gives (14) completing the proof. \[\square\]

Remark 3.4. The quantity $R_Q$ appearing in Theorem 3.3 can be controlled via simpler quantities as argued below. The resulting inequality will be useful in the sequel. Note that the loss function $L$ in Theorem 3.3 is zero-one valued. For every $a \in \mathcal{A}$, let

$$
B(a) := \{\theta \in \Theta : L(\theta, a) = 0\}.
$$

(18)

In other words, $B(a)$ denotes the “neighborhood” of the action $a$ consisting of all parameter values $\theta$ where the loss function equals 0. There is a slight abuse of notation here, $B(a) = B_1(a, L)$ for the zero-one valued $L$ and any $t \in [0, 1]$, where $B_t(a, L)$ is defined in (4). It is then easy to see that

$$
R_Q = 1 - E_Q w(B(\mathcal{d}_w(X))).
$$

(19)

where $E_Q$ denotes expectation taken under $X \sim Q$ and, as a result, we have

$$
1 - \max_{a \in \mathcal{A}} w(B(a)) \leq R_Q \leq 1 - \min_{a \in \mathcal{A}} w(B(a)).
$$

(20)

It should be noted that the bounds (20) for $R_Q$ depend neither on the Bayes estimator $\mathcal{d}_w(X)$ nor on the probability measure $Q$. Using such a lower bound on $R_Q$ enables us to replace the left hand side of (14) with the $f$-informativity by taking the infimum over $Q$. In particular, the monotonicity of $\phi_f(R, R_Q)$ in $R_Q$ (Lemma 3.1) implies that

$$
\phi_f(R, R_Q) \geq \phi_f(R, 1 - \max_{a \in \mathcal{A}} w(B(a))),
$$

provided that $1 - \max_{a \in \mathcal{A}} w(B(a)) \geq R$. Since $\phi_f(R, 1 - \max_{a \in \mathcal{A}} w(B(a)))$ is independent of the probability measure $Q$, one can take the infimum over $Q$ on the left hand side of (14), which leads to

$$
I_f(w, P) \geq \phi_f(R, 1 - \max_{a \in \mathcal{A}} w(B(a))).
$$
Remark 3.5. In the proof of Theorem 3.3, if we apply the data processing inequality to \( \Gamma(\theta, x) := L(\theta, \vartheta(x)) \) for an arbitrary decision rule \( \vartheta \) as opposed to the Bayes rule \( \vartheta_w \), we would obtain the following result:

\[
\int_{\Theta} D_f(P_{\theta}||Q)w(d\theta) \geq \phi_f(R^b, R_Q^b)
\]

where \( R^b := \int_{\Theta} \mathbb{E}_w L(\theta, \vartheta(x))w(d\theta) \) and \( R_Q^b := \int_{\Theta} \int_{\mathcal{X}} L(\theta, \vartheta(x))w(d\theta)Q(dx) \) (note that both \( R^b \) and \( R_Q^b \) also depend on the prior \( w \) but this is suppressed in the notation). Similar to (19), one has \( R_Q^b = 1 - \mathbb{E}_Q w(B(\vartheta(X))) \) and thus one has

\[
1 - \max_{a \in \mathcal{A}} w(B(a)) \leq R_Q^b \leq 1 - \min_{a \in \mathcal{A}} w(B(a)).
\]

Inequality (21) can be useful in deriving minimax lower bounds. For example, we can choose the decision rule \( \vartheta \) to be the minimax decision rule for the problem so that \( R^b \leq R_{\text{minimax}} \). If the distribution \( Q \) is then chosen so that \( R_{\text{minimax}} \leq R_Q^b \), then, by Lemma 3.1, the right hand side of (21) can be lower bounded by replacing \( R^b \) with \( R_{\text{minimax}} \) which yields

\[
\int_{\Theta} D_f(P_{\theta}||Q)w(d\theta) \geq \phi_f(R_{\text{minimax}}, R_Q^b).
\]

This can be converted into an explicit lower bound (by a argument similar to that outlined in Remark 3.3) for \( R_{\text{minimax}} \). We will describe two applications of this inequality in Section 4.

We are now ready to give the proof of our main result, Theorem 3.2.

Proof of Theorem 3.2. We will prove inequality (13). As described in Remark 3.1, inequality (13) is equivalent to Theorem 3.2. Fix \( f \in \mathcal{C} \), a prior \( w \) and suppose that \( \bar{I} \geq I_f(w, \mathcal{P}) \). Let \( t > 0 \) be such that

\[
\max_{a \in \mathcal{A}} w(B_t(a, L)) < 1 - u_f(\bar{I}).
\]

We prove below that \( R_{\text{Bayes}}(w, L) \geq t/2 \) and this would complete the proof. Let \( L_t \) denote the zero-one valued loss function \( L_t(\theta, a) := 1 \{ L(\theta, a) \geq t \} \). It is obvious that \( L \geq tL_t \) and hence the proof will be complete if we establish that \( R_{\text{Bayes}}(w, L_t) \geq 1/2 \). Let \( R = R_{\text{Bayes}}(w, L_t) \) for simplicity of notation.

Let \( Q \) be the probability measure which achieves the infimum in the definition (7) of \( I_f(w, \mathcal{P}) \).

Also, let

\[
R_Q := \int_{\Theta} \int_{\mathcal{X}} L_t(\theta, \vartheta_w(x))w(d\theta)Q(dx)
\]

where \( \vartheta_w \) is the Bayes decision rule for the loss function \( L_t \) under the prior \( w \). Theorem 3.3 then gives

\[
\bar{I} \geq I_f(w, \mathcal{P}) \geq \phi_f(R, R_Q).
\]

By (20) and (24), we have

\[
R_Q \geq 1 - \max_{a \in \mathcal{A}} w \{ \theta \in \Theta : L_t(\theta, a) = 0 \} = 1 - \max_{a \in \mathcal{A}} w \{ B_t(a, L) > u_f(\bar{I}) \}.
\]

This inequality and the definition (10) of \( u_f(\cdot) \) together imply that there exists \( 1/2 \leq b^* < R_Q \) such that \( \phi_f(1/2, b^*) > \bar{I} \). Now, according to Lemma 3.1, \( b \mapsto \phi_f(1/2, b) \) is non-decreasing, which yields \( \bar{I} < \phi_f(1/2, R_Q) \). Consequently, from (25), \( \phi_f(R, R_Q) \leq \bar{I} < \phi_f(1/2, R_Q) \). Further, by Lemma 3.1, \( a \mapsto \phi_f(a, R_Q) \) is non-increasing in \( a \). We thus have \( R \geq 1/2 \). The proof is complete.

Remark 3.6. We note that by the proof of Theorem 3.2, the constant \( 1/2 \) on the right hand side of (11) and in the definition of \( u_f(\cdot) \) can be replaced by any \( c \in (0, 1] \). In particular, we can have a sharper lower bound,

\[
R_{\text{Bayes}}(w, L) \geq \sup_{c \in (0, 1]} \left( c \sup \left\{ t > 0 : \max_{a \in \mathcal{A}} w \{ B_t(a, L) < 1 - u_{f,c}(\bar{I}) \} \right\} \right),
\]

where \( u_{f,c}(x) = \inf \{ c \leq b \leq 1 : \phi_f(c, b) \geq x \} \). Since the constant is not our main concern, the lower bound in (11) is usually sufficient to provide Bayes risk lower bounds with correct dependence on model and prior parameters.
Below we simplify inequality (13) for certain special functions \( f \in \mathcal{C} \).

**Example 3.4** (KL divergence). Suppose \( f(x) = x \log x \) so that \( D_f(P||Q) = D(P||Q) \) equals the KL divergence. Then the function \( u_f(x) \) in (10) has the expression for all \( x > 0 \),

\[
u_f(x) = \inf \{ 1/2 \leq b \leq 1 : b(1-b) < e^{-2x}/4 \} = \frac{1}{2} + \frac{1}{2} \sqrt{1 - e^{-2x}}.
\]

The elementary inequality \( \sqrt{1 - a} \leq 1 - a/2 \) gives for all \( x > 0 \),

\[
u_f(x) \leq 1 - \frac{1}{4} e^{-2x}.
\]

Inequality (13) reduces to

\[
R_{Bayes}(w,L) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{a \in A} w(B_t(a,L)) < \frac{1}{4} e^{-t} \right\}.
\]

**Example 3.5** (Chi-squared divergence). Suppose \( f(x) = x^2 - 1 \) so that \( D_f(P||Q) = \chi^2(P||Q) \). Then it is straightforward to see that

\[
u_f(x) = \inf \{ 1/2 \leq b \leq 1 : \frac{(1-2b)^2}{4b(1-b)} > x \} = \frac{1}{2} + \frac{1}{2} \sqrt{\frac{x}{1+x}}.
\]

Using the elementary inequality for all \( x > 0 \),

\[
\sqrt{\frac{x}{1+x}} \leq 1 - \frac{1}{2(1+x)}.
\]

we get from Theorem 3.2 that

\[
R_{Bayes}(w,L) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{a \in A} w(B_t(a,L)) < \frac{1}{4} \left( \frac{1}{1+t} \right) \right\}.
\]

**Example 3.6** (Total variation distance). Suppose \( f(x) = |x-1/2| \) so that \( D_f(P||Q) = \|P-Q\|_{TV} \) is the total variation distance between \( P \) and \( Q \). Then

\[
u_f(x) = \inf \{ 1/2 \leq b \leq 1 : |1-2b| > 2x \} = \frac{1}{2} + x
\]

which gives

\[
R_{Bayes}(w,L) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{a \in A} w(B_t(a,L)) < \frac{1}{2} - t \right\}.
\]

**Example 3.7** (Hellinger distance). Suppose \( f(x) = 1 - \sqrt{x} \) so that \( D_f(P||Q) = H^2(P||Q)/2 \). Then,

\[
u_f(x) = \inf \{ 1/2 \leq b \leq 1 : 1 - \sqrt{b/2} - \sqrt{(1-b)/2} > x \}.
\]

Since \( 0 \leq 1 - \sqrt{b/2} - \sqrt{(1-b)/2} \leq 1 - 1/\sqrt{2} \) for \( 1/2 \leq b \leq 1 \), \( u_f(x) = 1 \) when \( x \geq 1 - 1/\sqrt{2} \), which will not provide an interesting bound. On the other hand, when \( x < 1 - 1/\sqrt{2} \), we have

\[
u_f(x) = \frac{1}{2} + (1-x) \sqrt{x(2-x)},
\]

which gives that when \( \hat{t} < 1 - 1/\sqrt{2} \),

\[
R_{Bayes}(w,L) \geq \frac{1}{2} \sup \left\{ t > 0 : \max_{a \in A} w(B_t(a,L)) < \frac{1}{2} - \hat{t} \left( \hat{t} \left( 2 - \hat{t} \right) \right) \right\}.
\]

### 4 Special Instances of Theorem 3.3

Theorem 3.3, in the implicit form (14) and the explicit form (16), and its extension (21) have interesting connections to a number of existing bounds in the literature on minimaxity. These connections are described in this section. Except in the discussion on Assouad’s inequality in Subsection 4.3, the loss function \( L \) in this entire section is zero-one valued.
4.1 Generalized Fano inequalities

The Kullback-Leibler (KL) divergence is the most commonly used $f$-divergence in statistics and information theory. Below we examine the implications of Theorem 3.3 for the special case of the KL divergence. We show that the resulting inequalities generalize the classical Fano inequality (Cover and Thomas, 2006).

Theorem 3.3 with $f(x) = x \log x$ gives

$$
\int_{\Theta} D(P_\theta || Q) w(d\theta) \geq R \log \frac{R}{R_Q} + (1 - R) \log \frac{1 - R}{1 - R_Q}.
$$

The infimum of the left hand side above over all probability measures $Q$ is simply the mutual information between the random variables $\theta$ and $X$ under the distribution $\theta \sim w$ and $X|\theta \sim P_\theta$. Below we denote this by $I(\theta; X)$. The above inequality can be converted into an explicit bound for $R$ using (16). Indeed, taking $r = R_Q/(1 + R_Q)$ in (16), we obtain

$$
R \geq 1 + \int_{\Theta} D(P_\theta || Q) w(d\theta) + \log(1 + R_Q).
$$

We can simplify this inequality because $R_Q \leq 1$ (which means $\log(1 + R_Q) \leq \log 2$; also note that $\log(1 - R_Q) \leq 0$):

$$
R \geq 1 + \int_{\Theta} D(P_\theta || Q) w(d\theta) + \log 2.
$$

Using (20) in the denominator above and optimizing the resulting inequality over $Q$, we obtain

$$
R \geq 1 + \frac{\inf_Q \int_{\Theta} D(P_\theta || Q) w(d\theta) + \log 2}{\log(\max_{a \in A} w(B(a)))} = 1 + \frac{I(\theta; X) + \log 2}{\log(\max_{a \in A} w(B(a)))},
$$

where $B(a) := \{ \theta \in \Theta : L(\theta, a) = 0 \}$ is defined in (18). The right hand side of (28) also provides a lower bound on $R_{\text{minimax}}$ because $R_{\text{minimax}} \geq R$. The following special cases of (28) have appeared in the literature:

1. Suppose $\Theta (= A)$ is a finite set of cardinality $|\Theta| = N$. Also let $L(\theta, a) = I(\theta \neq a)$ so that $B(a) = \{ a \}$ for every $a$. If $w$ is the uniform probability measure on $\Theta$, then $w(B(a)) = w\{ a \} = 1/N$. Inequality (28) therefore gives

$$
R \geq 1 - \frac{I(\theta; X) + \log 2}{\log N}.
$$

This is the usual Fano’s inequality (Cover and Thomas, 2006) which is often used for obtaining minimax lower bounds for statistical estimation problems.

2. Suppose $\Theta (= A)$ is a finite set of cardinality $|\Theta| = N$ and $L$ is an arbitrary zero-one valued loss. Let $w$ be the uniform probability measure on $\Theta$ so that $w(B(a)) = |B(a)|/N$. Inequality (28) then gives

$$
R \geq 1 + \frac{I(\theta; X) + \log 2}{\log(\max_{a \in A} |B(a)|/N)}.
$$

This inequality has been proved in Duchi and Wainwright (2013, Corollary 1) where it is termed distance-based Fano’s inequality. Duchi and Wainwright (2013) make the assumption $N \geq \min_{a \in A} |B(a)| > \max_{a \in A} |B(a)|$ but our result does not require this assumption.

3. Suppose $\Theta (= A)$ is a subset of $\mathbb{R}^d$ with $0 < \nu(\Theta) < \infty$ where $\nu$ is Lebesgue measure. Also let $w$ denote the uniform probability measure over $\Theta$ so that $w(S) = \nu(\Theta \cap S)/\nu(S)$ for all $S$. Let $L$ be an arbitrary zero-one valued loss function on $\Theta \times \Theta$. Then (28) gives

$$
R \geq 1 + \frac{I(\theta; X) + \log 2}{\log(\max_{a \in A} \nu(B(a))/\nu(\Theta))}.
$$

This inequality recently appeared in Duchi and Wainwright (2013, Proposition 2) who termed it the continuum Fano inequality. The proof given in Duchi and Wainwright (2013) is rather complicated and is based on the inequality (29) and a discretization-approximation argument. It may be noted that our proof of (30) is no more difficult than that of (29) and is obtained as a simple corollary of Theorem 3.3.

9
Let us now look at inequality (21) for the special case when \( f(x) = x \log x \). Here we get

\[
\int_{\Theta} D_f(P_0 \| Q) w(d\theta) \geq R^0 \log \frac{R^0}{R_Q^0} + (1 - R^0) \log \frac{1 - R^0}{1 - R_Q^0}.
\]

We can rewrite this as

\[
\int_{\Theta} D_f(P_0 \| Q) w(d\theta) \geq -H(R^0) - R^0 \log R_Q^0 - (1 - R^0) \log (1 - R_Q^0)
\]

where \( H(x) := -x \log x - (1 - x) \log (1 - x) \). Using the bounds in (22) in the right hand side above, we deduce

\[
\int_{\Theta} D_f(P_0 \| Q) w(d\theta) \geq -H(R^0) - R^0 \log (1 - w_{\min}) - (1 - R^0) \log w_{\max}.
\]

where \( w_{\min} := \min_{a \in A} w(B(a)) \) and \( w_{\max} := \max_{a \in A} w(B(a)) \) for notational simplicity. Taking the infimum in the left hand side above over all probability measures \( Q \), we obtain

\[
I(\theta; X) \geq -H(R^0) - R^0 \log (1 - w_{\min}) - (1 - R^0) \log (w_{\max}).
\]

Provided \( w_{\min} + w_{\max} < 1 \), one can rewrite the above inequality as

\[
R^0 \geq \frac{-I(\theta; X) - H(R^0) - \log w_{\max}}{\log [(1 - w_{\min})/w_{\max}]}.
\] (31)

In a very different notation and with a very different proof, this inequality has recently appeared as the main result in (Braun and Pokutta, 2014, Proposition 2.2).

### 4.2 Birgé-Gushchin inequalities

Suppose \( \Theta = A = \{\theta_0, \theta_1, \ldots, \theta_N\} \) be a finite set of cardinality \( |\Theta| = N + 1 \) and consider the zero-one valued loss function, \( L(\theta, a) = \mathbb{I}(\theta \neq a) \). Let

\[
\psi_N(x) := x \log \left( \frac{N x}{N - x} \right) + (1 - x) \log \left( \frac{N(1 - x)}{x} \right).
\]

Birégé (2005) proved the following inequality which, in some ways, presents an improvement of the classical Fano inequality:

\[
\psi_N(R_{\text{minimax}}) \leq \min_{0 \leq j \leq N} \frac{1}{N} \sum_{i \neq j} D(P_{\theta_i} \| P_{\theta_j}),
\]

where \( D \) denotes the KL divergence. Gushchin (2003) extended Birégé’s inequality to arbitrary \( f \)-divergences by proving the following inequality:

\[
\psi_{N,f}(R_{\text{minimax}}) \leq \min_{0 \leq j \leq N} \frac{1}{N} \sum_{i \neq j} D_f(P_{\theta_i} \| P_{\theta_j}),
\] (32)

where

\[
\psi_{N,f}(x) := \frac{N - x}{N} f \left( \frac{N x}{N - x} \right) + \frac{x}{N} f \left( \frac{N(1 - x)}{x} \right).
\]

We will show below that (32) is a simple corollary of (21). For this, we need to prove that

\[
\sum_{i \neq j} D_f(P_{\theta_i} \| P_{\theta_j}) \geq \psi_{N,f}(R_{\text{minimax}}) \quad \text{for every } j \in \{0, \ldots, N\}.
\]

Without loss of generality, we assume that \( j = 0 \). We apply (21) with the uniform distribution on \( \Theta \setminus \{\theta_0\} = \{\theta_1, \ldots, \theta_N\} \) as \( w \), \( Q = P_{\theta_0} \) and the minimax rule for the problem as \( \psi \). Because \( \psi \) is the minimax rule, \( R^0 \leq R_{\text{minimax}} \).

Also

\[
R_Q^0 = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\theta_i} L(\theta_i, \psi(X)) = \frac{1}{N} \mathbb{E}_{\theta_0} \sum_{i=1}^N \mathbb{I} \{ \theta_i \neq \psi(X) \}.
\]

It is easy to check that \( \sum_{i=1}^N \mathbb{I} \{ \theta_i \neq \psi(X) \} = N - \{ \theta_0 \neq \psi(X) \} \). We thus have \( R_Q^0 = 1 - \mathbb{E}_{\theta_0} L(\theta_0, \psi(X))/N \). Because \( \psi \) is minimax, \( \mathbb{E}_{\theta_0} L(\theta_0, \psi(X)) \leq R_{\text{minimax}} \) and thus

\[
R_Q^0 \geq 1 - R_{\text{minimax}}/N.
\] (33)
On the other hand, it can be shown that $R_Q^2 \leq N/(N+1)$ (this can be seen, for example, by considering the randomized decision rule which takes each $\theta \in \Theta$ with probability $1/(N+1)$). We thus have, from (33), that $R_Q^2 \geq 1 - R_{\text{minimax}}/N \geq R_{\text{minimax}}$. We can thus apply (23) to obtain

$$\frac{1}{N} \sum_{i=1}^{N} D_f(P_{\theta_i}||P_{\theta_0}) \geq \phi_f(R_{\text{minimax}}, 1 - R_{\text{minimax}}/N).$$

The right hand side above equals $\psi_{N,f}(R_{\text{minimax}})$ which completes the proof of (32).

### 4.3 Inequalities involving total variation

The total variation distance appears in many classical minimax bounds including the well-known inequalities due to Le Cam (1973) and Assouad (1983). We will explain here how to derive these inequalities from Theorem 3.3. Let us first apply Theorem 3.3 with $f(x) = |x-1|/2$ so that $D_f(P||Q)$ equals the total variation distance $\|P - Q\|_{TV}$. In this case, Theorem 3.3 gives

$$\int_{\Theta} \|P_{\theta} - Q\|_{TV} w(d\theta) \geq \frac{R_Q}{2} \left( \frac{R}{R_Q} - 1 \right) + \frac{1 - R_Q}{2} \left( \frac{1 - R}{1 - R_Q} - 1 \right) = R_Q - R,$$

where the last equality uses the fact that $R \leq R_Q$ (see Lemma 7.1). Inverting the above inequality, we get

$$R \geq R_Q - \int_{\Theta} \|P_{\theta} - Q\|_{TV} w(d\theta).$$

Using (20) and optimizing the resulting inequality over $Q$, we further obtain,

$$R \geq \left( 1 - \max_{a \in A} w(B(a)) \right) - \inf_{Q} \int_{\Theta} \|P_{\theta} - Q\|_{TV} w(d\theta).$$

The simplest version of Le Cam’s inequality, the so-called two-point argument, is an easy corollary of (34) (or (35)). Indeed applying (34) to $\Theta = A = \{\theta_0, \theta_1\}$, $L(\theta, a) = \mathbb{I}\{\theta \neq a\}$ and $w(0) = w(1) = 1/2$ (these imply that $R_Q = 1/2$), we obtain

$$\frac{1}{2} (\|P_{\theta_0} - Q\|_{TV} + \|P_{\theta_1} - Q\|_{TV}) \geq 1/2 - R.$$

Taking $Q = (P_{\theta_0} + P_{\theta_1})/2$, we obtain Le Cam’s inequality:

$$R_{\text{minimax}} \geq \frac{1}{2} (1 - \|P_{\theta_0} - P_{\theta_1}\|_{TV}).$$

The more involved Le Cam’s inequality considers $\Theta = A = \Theta_0 \cup \Theta_1$ for two disjoint subsets $\Theta_0$ and $\Theta_1$ and loss function $L(\theta, a) = \mathbb{I}\{\theta \in \Theta_1, a \in \Theta_2\} + \mathbb{I}\{\theta \in \Theta_2, a \in \Theta_1\}$. The inequality states that for every pair of probability measures $w_0$ and $w_1$ concentrated on $\Theta_0$ and $\Theta_1$ respectively,

$$R_{\text{minimax}} \geq \frac{1}{2} \left( 1 - \|m_0 - m_1\|_{TV} \right)$$

where $m_0$ and $m_1$ are marginal densities given by $m_\tau(x) = \int p_\theta(x) w_\tau(d\theta)$ for $\tau = 0, 1$. To prove (37), consider the prior $w = (w_1 + w_2)/2$. Under this prior, the problem is easily converted to the previous binary testing problem. In particular, the data generating process under the prior $w$ can be viewed as first sampling $\tau \sim \text{Uniform} \{0, 1\}$ and then $X \sim m_\tau$. The decision $a \in A$ can be converted into the binary decision $\hat{\tau} = \mathbb{I}(a \in \Theta_1)$ and the loss function $L(\tau, \hat{\tau}) = \mathbb{I}(\tau \neq \hat{\tau})$. The Bayes risk under the prior $w$ can be re-written as,

$$R_{\text{Bayes}}(w, L) = \frac{1}{2} \inf_{\hat{\tau}} \sum_{\tau=0,1} \int_X \mathbb{I}(\tau \neq \hat{\tau}(x)) m_\tau(x) \mu(dx),$$

which has the same form as the Bayes risk in the earlier binary testing problem. Applying the same argument as for proving (36), we obtain the lower bound on the Bayes risk in (38), $R_{\text{Bayes}}(w, L) \geq \frac{1}{2} \left( 1 - \|m_0 - m_1\|_{TV} \right)$, which further implies (37).
Another classical minimax inequality involving the total variation distance is Assouad’s inequality (Assouad, 1983) which states that if $\Theta = A = \{0, 1\}^d$ and the loss function $L$ is defined by the Hamming distance, i.e., $L(\theta, a) = \sum_{i=1}^d I(\theta_i \neq a_i)$, then
\[
R_{\text{minimax}} \geq \frac{d}{2} \min_{L(\theta, a)=1} (1 - \|P_\theta - P_\theta'\|_{TV}).
\] (39)
This inequality is also a consequence of (34): let $w$ be the uniform probability measure on $\Theta$ and $L_1(\theta, a) = I(\theta_i \neq a_i)$. Under $w$, the marginal distribution of the first coordinate is $w_1(0) = w_1(1) = 1/2$. Let $m_\tau(x) := \sum_{\theta, \theta_i=\tau} p_\theta(x)/2^{d-1}$ for $\tau \in \{0, 1\}$ be the corresponding marginal density of $X$ and let $Q(x) = \frac{1}{2}(m_0(x) + m_1(x))$. Applying the same argument as for proving (36), we obtain that the minimax risk for the zero-one valued loss function $L_1(\theta, a)$ is bounded below by $\frac{1}{2}(1 - \|m_0 - m_1\|_{TV}) \geq \frac{1}{2} \min_{L(\theta, a')=1} (1 - \|P_\theta - P_\theta'\|_{TV})$. Repeating this argument for $L_i(\theta, a) := I(\theta_i \neq a_i)$ for $i = 2, \ldots, d$ and adding up the resulting bounds, we obtain (39).

### 4.4 Inequalities involving the chi-squared divergence

The chi-squared divergence (which corresponds to $f(x) = x^2 - 1$ and is denoted by $\chi^2(\|\cdot\|)$) is also a very popular $f$-divergence. Its use is facilitated by ease of computation for product measures (similar to the KL divergence) and also for mixtures of measures (unlike the KL divergence). Below we describe implications of Theorem 3.3 for the chi-squared divergence.

Theorem 3.3 for $f(x) = x^2 - 1$ gives
\[
\inf_Q \int_{\Theta} \chi^2(P_\theta\|Q)w(d\theta) \geq \frac{(R_Q - R)^2}{R_Q(1 - R_Q)}
\]
Because $R \leq R_Q$ (Lemma 7.1), we can invert the above to obtain,
\[
R \geq R_Q - \sqrt{R_Q(1 - R_Q)} \sqrt{\int_{\Theta} \inf_Q \chi^2(P_\theta\|Q)w(d\theta)}.
\]
The bounds (20) can now be employed so that
\[
R \geq \left(1 - \max_{a \in A} w(B(a))\right) - \sqrt{\left(1 - \min_{a \in A} w(B(a))\right) \max_{a \in A} w(B(a))} \sqrt{\int_{\Theta} \chi^2(P_\theta\|Q)w(d\theta)}
\] (40)
Specialized to the three situations as in Subsection 4.1, the above inequality yields distance-based and continuum Bayes risk bounds involving the chi-squared divergence. Inequality (21) applied to $f(x) = x^2 - 1$ can also be simplified to an inequality of the form (31). We omit the details.

### 4.5 Inequalities involving the Hellinger distance

Suppose $f(x) = 1 - \sqrt{x}$ so that $D_f(P||Q) = H^2(P||Q)/2$, where $H(P||Q)$ is the Hellinger distance between $P$ and $Q$. Theorem 3.3 gives
\[
\int_{\Theta} D_f(P_\theta||Q)w(d\theta) = 1 - \int_X \sqrt{q} \int_{\Theta} \sqrt{p_\theta}w(d\theta)d\mu \geq 1 - \sqrt{RR_Q} - \sqrt{(1 - R)(1 - R_Q)}
\] (41)
Since (41) holds for any probability measure $Q$, we find $Q^*$ with the density $q^*$ which minimizes the left hand side of (41). In particular, let $u := \int_{\Theta} \sqrt{p_\theta}w(d\theta)$, using the Cauchy-Schwartz inequality,
\[
\int_{\Theta} D_f(P_\theta||Q)w(d\theta) = 1 - \int_X \sqrt{qu^2} d\mu \geq 1 - \int_X u^2 d\mu,
\]
where the equality holds when $q$ is proportional to $u^2$. In other words, $\int_{\Theta} D_f(P_\theta||Q)w(d\theta)$ is minimized when choosing $q^*$ proportional to $u^2$. Plugging the probability measure $Q^*$ with the density $q^*$ into (41), we obtain that,
\[
\sqrt{RR_{Q^*}} + \sqrt{(1 - R)(1 - R_{Q^*})} \geq \int_X u^2 d\mu
\] (42)
To obtain a lower bound on $R$ from (42), we first provide the upper and lower bound for both the left hand side and right hand side of (42). For the right hand side, using Fubini’s theorem, we obtain that,

$$
\int_X u^2 \, d\mu = \int_\Theta \int_\Theta \int_X \sqrt{\rho_{\theta_0} \rho_{\theta'}} \, d\mu \, w(d\theta)w(d\theta') = 1 - \frac{1}{2}h^2 \tag{43}
$$

where $h^2 = \int_\Theta \int_\Theta H^2(P_\theta || P_{\theta'}) \, w(d\theta)w(d\theta')$.

If we assume that $h^2 \leq 2RQ^\ast$, we have $\int_X u^2 \geq 1 - RQ^\ast$. Therefore, the right hand side of the inequality (42) lies between $\sqrt{1 - RQ^\ast}$ and 1. On the other hand, it can be checked that, as a function in $R$, the left hand side of (42) is strictly increasing from $\sqrt{1 - RQ^\ast}$ (at $R = 0$) to 1 at $(R = RQ^\ast)$. Therefore, from (42), we know that $R \geq \tilde{R}$ where $\tilde{R} \in [0, RQ^\ast]$ is the solution to the equation obtained by replacing the inequality (42) with an equality. One can solve this equation and obtain two solutions. One of two solutions can be discarded by the fact that $R \leq RQ^\ast$. The other solution is given by:

$$
\tilde{R} = RQ^\ast - (2RQ^\ast - 1) \frac{h^2}{2} - \sqrt{RQ^\ast(1 - RQ^\ast)h^2(2 - h^2)}. \tag{44}
$$

Using (20), we obtain that,

$$
R \geq \left(1 - \max_{a \in A} w(B(a))\right) - \left(1 - 2 \min_{a \in A} w(B(a))\right) \frac{h^2}{2} - \sqrt{\left(1 - \min_{a \in A} w(B(a))\right) \max_{a \in A} w(B(a))} \sqrt{h^2(2 - h^2)} \tag{45}
$$

We note that the lower bound on $R$ in (45) only holds under the condition $h^2 \leq 2RQ^\ast$. When $h^2 > 2RQ^\ast$, which implies that $\int_X u^2 \, d\mu < \sqrt{1 - RQ^\ast}$, the inequality (45) holds for any $R \in [0, RQ^\ast]$ and thus cannot provide us a useful lower bound on $R$. As in the standard setting with the finite set $\Theta$, uniform prior and the loss $L(\theta, a) = \mathbb{I}(\theta \neq a)$, it is easy to check that the condition $h^2 \leq 2RQ^\ast$ indeed holds. In particular, suppose $\Theta = A$ be a finite set of cardinality $|\Theta| = N$. Also let $L(\theta, a) = \mathbb{I}(\theta \neq a)$ and $w$ be the uniform probability measure on $\Theta$ so that $w(B(a)) = w\{a\} = 1/N$. According to (19), $RQ^\ast = 1 - \frac{1}{N}$ and we have that,

$$
h^2 = \frac{1}{N^2} \sum_{\theta \neq \theta'} H^2(P_\theta || P_{\theta'}) \leq 2 \frac{N(N - 1)}{N^2} = 2RQ^\ast. \tag{46}
$$

Inequality (45) therefore gives,

$$
R \geq 1 - \frac{1}{N} - N - \frac{2h^2}{2} - \sqrt{\frac{N - 1}{N}} \sqrt{h^2(2 - h^2)},
$$

which recovers the result in Example II.6 in Guntuboyina (2011b).

## 5 Upper Bounds on $f$-informativity

For applications of Theorem 3.2, we need good upper bounds on the $f$-informativity $I_f(w; P)$. For the case of the KL divergence ($f(x) = x \log x$) when the $f$-informativity equals the mutual information, Haussler and Opper (1997) proved very useful upper bounds. Our goal in this section is to extend the upper bounds of Haussler and Opper (1997) to $f$-informativities for $f_\alpha \in C$ for every $\alpha \notin [0, 1]$ (recall the definition of $f_\alpha$ from Section 2). Our method does not seem to apply the case when $\alpha \in (0, 1)$. But the case when $\alpha \notin [0, 1]$ seems to suffice for the applications (note that the chi-squared divergence corresponds to $\alpha = 2$).

Let us begin by reviewing the inequality of Haussler and Opper (1997). Let $P$ and $\{\Theta, \Theta \in \Xi\}$ be probability measures on a space $X$ and let $\rho$ and $\{\rho_\theta, \theta \in \Xi\}$ denote their densities with respect to a common positive measure $\lambda$. Let $\nu$ be an arbitrary probability measure on $\Xi$ and $\bar{Q}$ be the probability measure on $X$ with density $\bar{Q} = \int_\Xi \rho_\theta \nu(d\theta)$. Haussler and Opper (1997) proved the following inequality

$$
D(P || \bar{Q}) \leq -\log \left( \int_\Xi \exp \left( -D(P || \rho_\theta) \right) \nu(d\theta) \right), \tag{47}
$$
Now given a class of probability measures \( \{P_\theta, \theta \in \Theta \} \), applying the above inequality for each \( P_\theta \) and integrating the resulting inequalities with respect to a probability measure \( w \) on \( \Theta \), (Haussler and Opper, 1997, Theorem 2) obtained the following mutual information upper bound:
\[
\inf_Q \int_\Theta D_f(P_\theta||Q)w(d\theta) \leq - \int_\Theta w(d\theta) \log \left( \int_\Xi \exp(-D(P_\theta||Q_\theta)) \nu(d\theta) \right). \tag{48}
\]

In the special case when \( \Xi = \{1, \ldots, M\} \) and \( \nu \) is the uniform probability measure on \( \Xi \), we have \( \bar{Q} = (Q_1 + \ldots + Q_M)/M \). Inequality (47) then becomes,
\[
D(P||\bar{Q}) \leq - \log \left( \frac{1}{M} \sum_{j=1}^M \exp(-D(P||Q_j)) \right).
\]
Because \( \sum_{j=1}^M \exp(-D(P||Q_j)) \geq \exp(-\min_j D(P||Q_j)) \), we obtain
\[
D(P||\bar{Q}) \leq \log M + \min_{1 \leq j \leq M} D(P||Q_j).
\]
Further inequality (48) can be simplified to
\[
\inf_Q \int_\Theta D_f(P_\theta||Q)w(d\theta) \leq \log M + \int_\Theta \min_{1 \leq j \leq M} D(P_\theta||Q_j)w(d\theta). \tag{49}
\]

This inequality gives an upper bound for \( f \)-informativity in terms of KL covering numbers. Recall the definition of \( M_{KL}(\epsilon, \Theta) \) from (8). Applying (49) for any fixed \( \epsilon > 0 \) and \( \{Q_1, \ldots, Q_M\} \) chosen to be an \( \epsilon^2 \)-cover, we obtain
\[
\inf_Q \int_\Theta D_f(P_\theta||Q)w(d\theta) \leq \inf_{\epsilon > 0} (\log M_{KL}(\epsilon, \Theta) + \epsilon^2). \tag{50}
\]

We note that in the case when \( w \) is the uniform probability measure on a finite subset of \( \Theta \), the above inequality was proved by Yang and Barron (1999, Page 1571).

The next theorem extends inequalities (47) and (48) to \( f_\alpha \) with \( \alpha \notin [0,1] \).

**Theorem 5.1.** Fix \( \alpha \notin [0,1] \) and let \( f_\alpha \in \mathcal{C} \) be as defined in Section 2. Under the setting of inequalities (47) and (48), we have
\[
D_{f_\alpha}(P||\bar{Q}) \leq \left[ \int_\Xi (D_{f_\alpha}(P_\theta||Q_\theta) + 1)^{1/(1-\alpha)} \nu(d\theta) \right]^{1-\alpha} - 1. \tag{51}
\]
and
\[
I_{f_\alpha}(w, P) \leq \int_\Xi \left[ \int_\Xi (D_{f_\alpha}(P_\theta||Q_\theta) + 1)^{1/(1-\alpha)} \nu(d\theta) \right]^{1-\alpha} w(d\theta) - 1. \tag{52}
\]

In the proof of Theorem 5.1, the following technical lemma will be crucially used and its proof is provided in the Appendix.

**Lemma 5.2.** Let \( \mu \) be a probability measure on the space \( T \) and let \( S := \{ u : T \to \mathbb{R}_+ : u \in L_p^p(T) \} \). Then the map \( f : S \to \mathbb{R} \) defined by \( f(u) := (\int_T u(t)^p \mu(dt))^{1/p} \) is concave in \( u \) when \( p < 1 \).

Note that the discrete version of Lemma 5.2 states that \( f(u) = \left( \frac{1}{M} \sum_{i=1}^M u_i^p \right)^{1/p} \) is concave function in \( u \in \mathbb{R}_{>0}^M \) when \( p < 1 \).

**Proof of Theorem 5.1.** By the identity that \( D_{f_\alpha}(P||Q) = D_{f_{1-\alpha}}(Q||P) \) in (5), we have
\[
D_{f_\alpha}(P||\bar{Q}) = D_{f_{1-\alpha}}(\bar{Q}||P) = \int_X p \left( \int_\Xi \frac{q_\theta}{p} \nu(d\theta) \right)^{1-\alpha} - 1
\]
\[
= \int_X p \left( \int_\Xi \left( \frac{q_\theta}{p} \right)^{1/(1-\alpha)} \nu(d\theta) \right)^{1-\alpha} - 1
\]
Let $u(\vartheta, x) = \left(\frac{d}{d\vartheta}\right)^{1-\alpha}$. Since $\frac{1}{\alpha} < 1$ when $\alpha \notin [0, 1]$, Lemma 5.2 implies that $u(\vartheta, x) \mapsto \left(\int_{\Xi} u(\vartheta, x)^{(1-\alpha)/\alpha} \nu(d\theta)\right)^{1-\alpha}$ is concave in $u$. Applying Jensen’s inequality,

$$D_{f_\alpha}(P||Q) \leq \left(\int_{\Xi} \left[\int_{\mathcal{X}} \frac{q_\vartheta}{p_\vartheta} \nu(d\theta)\right]^{1/(1-\alpha)} \nu(d\theta)\right)^{1-\alpha} - 1 = \left(\int_{\Xi} [D_{f_\alpha}(Q_\vartheta||P)]^{1/(1-\alpha)} \nu(d\theta)\right)^{1-\alpha} - 1.$$

This completes the proof of (51) because $D_{f_{1-\alpha}}(Q_\vartheta||P) = D_{f_\alpha}(P||Q)$. The proof of (52) follows by applying (51) for $P = P_\vartheta$ and then integrating the resulting bound with respect to $w(d\vartheta)$. \qed

We demonstrate the effectiveness of the bound (51) by means of the following example.

**Example 5.3.** In this example, we show the tightness of the upper bound in (51) in terms of chi-squared divergence ($\alpha = 2$). In particular, let the distribution $P$ be the $n$-fold product of $N(0,1)$ and $Q_\vartheta$ be the $n$-fold product of $N(\vartheta, 1)$ where $\vartheta \sim N(0,1)$. It is straightforward to show that the marginal distribution $\bar{Q}$ is a $n$-dimensional Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $I_n + \mathbf{1}_n\mathbf{1}_n^T$, where $\mathbf{1}_n$ denotes the $n$-dimensional all one vector and $I_n$ the $n \times n$ identity matrix.

Since $\chi^2(P||Q_\vartheta) = \exp(n\vartheta^2) - 1$, the right hand side of (51) equals to $\sqrt{2n+1} - 1$. The term $\chi^2(P||\bar{Q})$ on the left hand side of (51) is difficult to evaluate. However, we can lower bound $\chi^2(P||\bar{Q})$ using the following standard inequality $\exp(D(P||\bar{Q})) - 1 \leq \chi^2(P||\bar{Q})$ (see Lemma 2.7 in Tsybakov (2010)). By the closed-form expression for KL divergence between two multivariate Gaussian distributions, we have $D(P||\bar{Q}) = \frac{1}{2}(\log(n+1) - n/(n+1))$ and thus

$$e^{-1/2}\sqrt{n+1} - 1 \leq \exp(D(P||\bar{Q})) - 1 \leq \chi^2(P||\bar{Q})$$

As we can see, the upper bound $\sqrt{2n+1} - 1$ in (51) is quite tight and $\chi^2(P||\bar{Q})$ is on the order of $\sqrt{n}$.

For $\alpha > 1$, one can obtain upper bounds analogous to (50) for the $f_\alpha$-informativity. These are described below; recall the notation $M_\alpha(\varepsilon, S)$ from (8).

**Corollary 5.4.** For every $\alpha > 1$, we have

$$I_{f_\alpha}(w, P) \leq \inf_{\varepsilon > 0} (1 + \varepsilon^2) M_\alpha(\varepsilon, \Theta)^{\alpha-1} - 1. \quad (53)$$

**Proof.** Let $Q_1, \ldots, Q_M$ be probability measures on $\mathcal{X}$ and fix $\vartheta \in \Theta$. Inequality (51) applied to $P = P_\vartheta$, $\Xi := \{1, \ldots, M\}$ and the uniform probability measure on $\Xi$ as $\nu$ gives

$$D_{f_\alpha}(P_\vartheta||\bar{Q}) \leq M^{\alpha-1} \left[\sum_{j=1}^{M} (1 + D_{f_\alpha}(P_\vartheta||Q_j))^{1/(1-\alpha)}\right]^{1-\alpha} - 1.$$

We now use (note that $\alpha > 1$)

$$\sum_{j=1}^{M} (1 + D_{f_\alpha}(P_\vartheta||Q_j))^{1/(1-\alpha)} \geq \max_{1 \leq j \leq M} (1 + D_{f_\alpha}(P_\vartheta||Q_j))^{1/(1-\alpha)}$$

$$= (1 + \min_{1 \leq j \leq M} D_{f_\alpha}(P_\vartheta||Q_j))^{1/(1-\alpha)}.$$

This gives

$$D_{f_\alpha}(P_\vartheta||\bar{Q}) \leq M^{\alpha-1} \left(1 + \min_{1 \leq j \leq M} D_{f_\alpha}(P_\vartheta||Q_j)\right)^{1/(1-\alpha)} - 1.$$

We now fix $\varepsilon > 0$ and apply the above with $\{Q_1, \ldots, Q_M\}$ taken to be an $\varepsilon^2$-cover of $\Theta$ under the $f_\alpha$-divergence. We then obtain

$$D_{f_\alpha}(P_\vartheta||\bar{Q}) \leq \inf_{\varepsilon > 0} (1 + \varepsilon^2) M_\alpha(\varepsilon, \Theta)^{\alpha-1} - 1.$$

The proof is complete by integrating the above inequality with respect to $w(d\vartheta)$. \qed

Corollary 5.4 will be mostly used with $\alpha = 2$ i.e., when $D_{f_\alpha}$ is the chi-squared divergence $\chi^2$. 

15
6 Examples

In this section, we apply the bounds developed so far to derive Bayes risk lower bounds in actual problems. We consider four different estimation settings: the gaussian location model (Examples 6.1, 6.2, 6.3 and 6.5), Bayesian lasso (Example 6.6), generalized linear models (Example 6.7) and a spiked covariance model (Example 6.8). We also present general lower bound for finite dimensional estimation problems as a corollary of Theorem 3.2 with \( f(x) = x \log x \) (Corollary 6.4).

Example 6.1 (Gaussian Model). Fix \( d \geq 1 \). Suppose \( \Theta = A = \mathbb{R}^d \) and let \( L(\theta, a) := \| \theta - a \|_2^2 \) where \( \| \cdot \|_2 \) is the usual Euclidean norm on \( \mathbb{R}^d \). For each \( \theta \in \mathbb{R}^d \), let \( P_\theta \) denote the Gaussian distribution with \( \theta \) and covariance matrix \( \sigma^2 I_d \) (\( \sigma^2 > 0 \) is a constant). We shall then show below that for every prior \( w \) on \( \mathbb{R}^d \) having a Lebesgue density that is bounded from above by \( W > 0 \), we have

\[
R_{\text{Bayes}}(w, L; \Theta) \geq \frac{d \sigma^4 W^{-2/d}}{(\sigma^2 + V)^2},
\]

(54)

where \( V := \max_{1 \leq i \leq d} \int \theta_i^2 w(d\theta) \).

To prove (54), we shall apply (13) with \( f(x) = x \log x \), i.e., we apply (26). The resulting \( f \)-informativity (a.k.a mutual information) can be bounded in the following way. Because \( I_f(w, P) \leq \int D(P||Q)w(d\theta) \) for every \( Q \), taking \( Q \) to be the Gaussian distribution with mean zero and covariance matrix \( (\sigma^2 + V)I_d \), we obtain

\[
I_f(w, P) \leq \int_{\Theta} D(N(\theta, \sigma^2 I_d) || N(0, (\sigma^2 + V)I_d))w(d\theta).
\]

Using the standard formula for the KL divergence between two Gaussians, we deduce that

\[
I_f(w, P) \leq \frac{1}{2} \int_{\Theta} \left( \sum_{i=1}^{d} \frac{\theta_i^2 - V}{\sigma^2 + V} + d \log \frac{\sigma^2 + V}{\sigma^2} \right) w(d\theta)
\]

which implies, because \( \int \theta_i^2 w(d\theta) \leq V \), that

\[
I_f(w, P) \leq \frac{d}{2} \log \frac{\sigma^2 + V}{\sigma^2}.
\]

(55)

Let \( \hat{I} \) denote the right hand side above. To apply (26), we also need an upper bound on \( \max_{a \in A} w(B_t(a, L)) \). Because of the assumption that the Lebesgue density of \( w \) is bounded from above by \( W \), we get

\[
\max_{a \in A} w(B_t(a, L)) \leq W^{t^{d/2}} \text{Vol}(B)
\]

(56)

where \( B \) is the Euclidean ball with unit radius. Thus the choice

\[
t = c W^{-2/d} \text{Vol}(B)^{-2/d} \frac{\sigma^4}{(\sigma^2 + V)^2},
\]

for a small enough universal positive constant \( c \), ensures \( \max_{a \in A} w(B_t(a)) < \frac{1}{2} e^{-2t} \) (recall that \( \hat{I} \) is the right hand side of (55)). Consequently, inequality (26) implies that \( R_{\text{Bayes}} \geq \hat{I}/2 \). The proof of (54) is now completed using the standard fact: \( \text{Vol}(B)^{1/d} \leq d^{-1/2} \).

To investigate whether (54) presents a tight Bayes risk lower bound, it is helpful to note the trivial upper bound:

\[
R_{\text{Bayes}}(w, L; \Theta) \leq d \min(\sigma^2, V).
\]

(57)

This inequality follows from the fact that \( R_{\text{Bayes}}(w, L; \Theta) \) is smaller than the risk of the constant estimator 0 as well as that of the trivial estimator \( X \). Below we isolate three special priors \( w \) for which the lower bound (54) can be explicitly evaluated and compared with (57).

1. Suppose \( w \) is the uniform distribution on the closed ball of radius \( \Gamma \) centered at the origin. Then \( W = \Gamma^{-d}/\text{Vol}(B) \). Because \( \text{Vol}(B)^{1/d} \leq d^{-1/2} \), we have \( W^{-2/d} \geq \Gamma^{2}/d \). It is also easy to check that \( V \leq \Gamma^{2}/d \). Inequality (54) therefore gives

\[
R_{\text{Bayes}}(w, L; \Theta) \geq \frac{d^2 \sigma^4 \Gamma^{2}}{(d \sigma^2 + \Gamma^{2})^2}.
\]

(58)
Also note that (57) becomes \( R_{\text{Bayes}}(w, L; \Theta) \leq \min(d\sigma^2, \Gamma^2) \). Now when \( \Gamma \leq \sigma \sqrt{d} \), then (58) gives \( R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{16} \sigma^2 \) which is tight in light of the upper bound. On the other hand, (58) is not tight when \( \Gamma \) becomes large because then the right hand side of (58) goes to zero. In Example 6.3, we describe how to derive tight lower bounds even when \( \Gamma \) becomes large.

2. Suppose \( w \) is the uniform distribution on the hypercube \([-\Gamma, \Gamma]^d\). Then it is easy to see that \( W = (2\Gamma)^{-d} \) and \( V \leq \Gamma^2 \). Inequality (54) thus gives

\[
R_{\text{Bayes}}(w, L; \Theta) \geq \frac{d\sigma^4 \Gamma^2}{(\sigma^2 + \Gamma^2)^2}.
\]

The upper bound (57) becomes \( R_{\text{Bayes}}(w, L; \Theta) \leq \min(d\sigma^2, d\Gamma^2) \). Similar to the previous case, the lower bound given by (54) is tight when \( \Gamma \leq \sigma \) but becomes loose when \( \Gamma \) becomes large.

3. Suppose \( w \sim N(0, \tau^2 I_d) \), i.e., the Gaussian distribution with mean zero and covariance matrix \( \tau^2 I_d \). Then \( W = (2\tau^2)^{-d/2} \) and \( V = \tau^2 \) so that (54) becomes

\[
R_{\text{Bayes}}(w, L; \Theta) \geq \frac{d\sigma^4 \tau^2}{(\sigma^2 + \tau^2)^2}.
\]

(59)

In this case, the Bayes risk can be evaluated explicitly. Indeed, it is easy to show that

\[
R_{\text{Bayes}}(w, L; \Theta) = \frac{d\sigma^2 \tau^2}{\sigma^2 + \tau^2}.
\]

(60)

Comparing (59) and (60), we see that (59) is tight when \( \tau^2 \leq \sigma^2 \) but becomes loose for large \( \tau^2 \).

**Example 6.2** (Gaussian Model with general loss). Consider the same setup as in the previous example but now allow the loss function to be \( L(\theta, a) = \|\theta - a\|^2 \) for an arbitrary norm \( \|\cdot\| \) (not necessarily the Euclidean norm) on \( \mathbb{R}^d \). In this case, we obtain the following Bayes risk lower bound:

\[
R_{\text{Bayes}}(w, L; \Theta) \geq \frac{\sigma^2 W^{-d/2} d^2}{(\sigma^2 + V)^2 (E\|Z\|)^2}.
\]

(61)

where \( Z \) is a standard Gaussian vector and \( \|\cdot\|_* \) is the dual norm corresponding to \( \|\cdot\| \) defined by \( \|x\|_* := \sup \{ \langle x, y \rangle : \|y\| \leq 1 \} \). The quantities \( W \) and \( V \) are as defined in Example 6.1.

The proof of (61) is largely similar to that of (54). We use (26) along with (55) for controlling \( I_f(w, \tilde{P}) \). To control \( \max_{a \in A} w(B_t(a, L)) \), we again use the fact that the Lebesgue density of \( w \) is bounded from above by \( W \) to obtain

\[
\max_{a \in A} w(B_t(a, L)) \leq W \text{Vol}\left\{ \theta \in \mathbb{R}^d : \|\theta\| < \sqrt{t} \right\}.
\]

(62)

To deal with the volume term above, we use Urysohn’s inequality to obtain an upper bound in terms of the volume of the unit Euclidean unit ball \( B \). The original reference for Urysohn’s inequality is Urysohn (1924) but it has been recently used in a statistical context by Ma and Wu (2013). Urysohn’s inequality gives

\[
\left( \frac{\text{Vol}\left\{ \theta \in \mathbb{R}^d : \|\theta\| < \sqrt{t} \right\}}{\text{Vol}(B)} \right)^{\frac{1}{d}} \leq \frac{\sqrt{t}}{\sqrt{d}} E\|Z\|_* \quad \text{with} \quad Z \sim N(0, I_d).
\]

(63)

Inequalities (62) and (63) together give

\[
\max_{a \in A} w(B_t(a, L)) \leq W t^{d/2} \text{Vol}(B) \left( \frac{E\|Z\|_*}{\sqrt{d}} \right)^d.
\]

The choice

\[
t = c \text{Vol}(B)^{-2/d} W^{-2/d} \sigma^4 \frac{d}{(\sigma^2 + V)^2 (E\|Z\|_*)^2}
\]

for a small enough universal positive constant \( c \) ensures \( \max_{a \in A} w(B_t(a)) \leq \frac{1}{c} e^{-2t} (I \text{ is the right hand side of (55))}. \) The proof of (61) is then completed by noting that \( \text{Vol}(B) \frac{1}{d} \gg d^{-1/2} \).
Example 6.3 (Gaussian model with uniform priors on large balls). Consider the same setting of Example 6.1 and let \( w \) be the uniform distribution on the closed ball of radius \( \Gamma \) centered at the origin. Let \( \Gamma \geq \sigma \sqrt{d} \). As discussed in Example 6.1, the bound given by (54) is not always tight in this setting. We show below how to derive the optimal Bayes risk lower bound on using chi-squared divergence.

We assume here that \( \Theta \) (and \( A \)) equal the closed ball of radius \( \Gamma \) centered at the origin because \( w \) puts zero probability outside this ball. We apply (13) with \( f(x) = x^2 - 1 \), i.e., we use (27). For this, we need to bound \( \max_{a \in A} w(B_2(a, L)) \) and the \( f \)-informativity corresponding to the chi-squared divergence. The former can be easily controlled because

\[
\max_{a \in A} w(B_2(a, L)) \leq \left( \sqrt[4]{\Gamma} \right)^d.
\]

For the latter, we use (53) in Corollary 5.4 with \( \alpha = 2 \). This entails getting an upper bound for \( M_C(\epsilon, \Theta) \). Note that \( \chi^2(P_\theta || P_{\theta'}) = \exp\left( \|\theta - \theta'^{\prime}\|^2/\sigma^2 \right) - 1 \) for \( \theta, \theta' \in \Theta \). As a consequence, \( \chi^2(P_\theta || P_{\theta'}) \leq \epsilon^2 \) if and only if \( \|\theta - \theta'^{\prime}\| \leq \epsilon':=\sigma \sqrt{\log(1+\epsilon^2)} \). Therefore, by a standard volumetric argument, we get

\[
M_C(\epsilon, \Theta) \leq \left( \frac{3\Gamma}{\epsilon'} \right)^d \leq \left( \frac{3\Gamma}{\sigma \sqrt{\log(1+\epsilon^2)}} \right)^d
\]

provided \( \epsilon' \leq \Gamma \). In particular, if we take \( \epsilon := \sqrt{\epsilon'} - 1 \), then \( \epsilon' = \sigma \sqrt{d} \leq \Gamma \), we would get \( M_C(\epsilon, \Theta) \leq (3\Gamma/(\sigma \sqrt{d}))^d \). Inequality (53) for \( \alpha = 2 \) then gives

\[
I_f(w, \mathcal{P}) \leq \left( \frac{3\epsilon \Gamma}{\sigma \sqrt{d}} \right)^d - 1
\]

Let \( \tilde{I} \) be the right hand side above. It is then easy to see that \( \max_{a \in A} w(B_2(a, L)) < \frac{1}{4}(1 + \tilde{I})^{-1} \) provided \( t = c d \sigma^2 \) for a small enough positive constant \( c \). Inequality (27) then gives

\[
R_{\text{Bayes}}(w, L; \Theta) \geq d \sigma^2.
\]

(64)

This lower bound is clearly tight because of (57).

This example also gives us a chance to contrast the bounds given by (13) for different \( f \in \mathcal{C} \). We argue below that it is not possible to derive (64) by applying (13) to \( f(x) = x \log x \), i.e., (26) along with inequality (50) for controlling the informativity \( I_f(w, \mathcal{P}) \). In other words, the same strategy which works for \( f(x) = x^2 - 1 \) does not work for \( f(x) = x \log x \). To see this, first note that \( D(P_\theta || P_{\theta'}) = \|\theta - \theta'^{\prime}\|^2/\sigma^2 \) for \( \theta, \theta' \in \Theta \). As a result, \( D(P_\theta || P_{\theta'}) \leq \epsilon^2 \) if and only if \( \|\theta - \theta'^{\prime}\| \leq \sqrt{2} \sigma \). The same volumetric argument again gives

\[
M_{KL}(\epsilon, \Theta) \leq \left( \frac{3\Gamma}{\sqrt{2} \sigma} \right)^d \quad \text{provided } \sqrt{2} \sigma \leq \Gamma.
\]

The bound (50) says that the mutual information \( I_f(w, \mathcal{P}) \) (note that \( f(x) = x \log x \)) is bounded by

\[
I_f(w, \mathcal{P}) \leq \inf_{0 < \epsilon' \leq \sqrt{2} / \sqrt{1+4\epsilon^2} \sigma} \left( d \log \left( \frac{3\Gamma}{\sqrt{2} \sigma} \right) + \epsilon^2 \right).
\]

The infimum above can be easily calculated by calculus which results in

\[
I_f(w, \mathcal{P}) \leq d \log \left( \frac{3\Gamma}{\sqrt{2} \sigma} \right) + \frac{d}{2}.
\]

Let \( \tilde{I} \) be the right hand side above. The maximum \( t > 0 \) for which \( (\sqrt{t}/\Gamma)^d \) is smaller than \( \frac{1}{4} \exp\left( -2 \tilde{I} \right) \) is easily seen to be of the order \( d^2 \sigma^4 / \Gamma^2 \). This means that inequality (26) can only result in the weaker lower bound of \( d^2 \sigma^4 / \Gamma^2 \) for \( R_{\text{Bayes}}(w, L; \Theta) \). As we have seen in Example 6.1, this lower bound is suboptimal when \( \Gamma \) becomes large. This should be contrasted with the optimal bound (64).

As we observed in the previous example, direct application of Theorem 3.2 with \( f(x) = x \log x \) and the upper bound on mutual information in (50) does not always produce optimal bounds. It turns out that if we partition the parameter space \( \Theta \) into small hypercubes and if we separately apply Theorem 3.2 with \( f(x) = x \log x \) to \( w \) partitioned on each individual hypercube, then the optimal rate can be recovered. We describe below this method in a general setting and then apply it to three specific examples.
Corollary 6.4. Let $\Theta = A$ be a subset of $\mathbb{R}^d$. Suppose that the prior $w$ has a Lebesgue density $f_w$ that is positive over $\Theta$. For each $\theta \in \Theta$ and $\delta > 0$, let

$$r_\delta(\theta) := \sup \left\{ \frac{\int f_w(\theta)}{f_w(\theta')} : \theta \in \Theta \text{ and } \|\theta - \theta\|_2 \leq \sqrt{d}\delta \text{ for } i = 1, 2 \right\}.$$

Suppose also the existence of $A > 0$ such that $D(P_\theta || P_{\theta'}) \leq 1\|\theta - \theta\|^2/2$ for all $\theta, \theta' \in \Theta$ and the existence of $V > 0$ (which may depend on $d$) and $p > 0$ such that $\max_{a \in A} \text{Vol}(B(a, L)) \leq V^{d/p}$ for every $t > 0$. Then

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup_{0 < \delta \leq A^{-1/2}} \left[ e^{-2p\delta^2} (8V)^{-p/d} \int_{\Theta} \frac{1}{r_\delta(\theta)} \ w(d\theta) \right]. \quad (65)$$

Proof. Fix $0 < \delta \leq A^{-1/2}$. Partition the entire parameter space $\Theta$ into small hypercubes each with side length $\delta$. For each such hypercube $S$ and let $\pi_S$ denote the probability measure $w$ conditioned to be in $S$ i.e., $\pi_S(C) := w(C)/w(S)$ for measurable set $C \subseteq S$.

For every decision rule $d(X)$, clearly

$$\int_{\Theta} \mathbb{E}_d L(\theta, d(X)) w(d\theta) = \sum_S w(S) \int_S \mathbb{E}_d L(\theta, d(X)) d\pi_S(\theta)$$

where the sum above is over all hypercubes $S$ in the partition. This implies therefore that

$$R_{\text{Bayes}}(w, L; \Theta) \geq \sum_S w(S) R_{\text{Bayes}}(\pi_S, L; S).$$

The proof will therefore be completed if we show that

$$R_{\text{Bayes}}(\pi_S, L; S) \geq \frac{1}{2} \sup_{0 < \delta \leq A^{-1/2}} \left[ e^{-2p\delta^2} (8V)^{-p/d} \int_S \frac{1}{r_\delta(\theta)} \ \pi_S(d\theta) \right] \quad (66)$$

for every fixed hypercube $S$. So let us fix $S$ and, for notational simplicity, let $\pi := \pi_S$. We will use (26) to prove a lower bound on $R_{\text{Bayes}}(\pi_S, L; S)$. Note first that

$$\inf_{Q} \int_S D(P_\theta || Q\pi(d\theta)) \leq \int_S D(P_\theta || P_{\theta'}) \pi(d\theta) \pi(d\theta') \leq A \max_{\theta_1, \theta_2 \in S} \|\theta - \theta\|^2 \leq Ad\delta^2 := \bar{I}. \quad (67)$$

Also, letting $f_{\pi}^{\max}$ and $f_{\pi}^{\min}$ be the maximum and minimum values of $f_w$ in $S$, we have

$$\max_{a \in S} \pi(B(a, L)) \leq f_{\pi}^{\max} \text{Vol}(B(a, L)) \leq f_{\pi}^{\max} \bar{I} \leq \frac{V^{d/p}}{f_{\pi}^{\min}}.$$

Let $\tilde{\theta}$ be an arbitrary point in the set $S$. Since $S$ has diameter $\sqrt{d}\delta$, the set $\{\theta : \|\theta - \tilde{\theta}\|_2 \leq \sqrt{d}\delta\}$ contains $S$. We obtain from the definition of $r_\delta(\theta)$ that $f_{\pi}^{\max} / f_{\pi}^{\min} \leq r_\delta(\tilde{\theta})$ so that

$$\max_{a \in S} \pi(B(a, L)) \leq r_\delta(\tilde{\theta}) V^{-d/p} \delta^d.$$

Thus, by (67), the choice

$$t = e^{-2p\delta^2} \delta^p \left( \frac{1}{8Vr_\delta(\theta)} \right)^{p/d},$$

leads to $\max_{a \in S} \pi(B(a, L)) < \frac{1}{4} e^{-t}. $ Employing (26), we deduce

$$R_{\text{Bayes}}(\pi, L; S) \geq \frac{1}{2} e^{-2p\delta^2} \delta^p \left( \frac{1}{8Vr_\delta(\theta)} \right)^{p/d} \geq \frac{1}{2} e^{-2p\delta^2} \delta^p \left( \frac{1}{8Vr_\delta(\theta)} \right)^{p/d}$$

where we used the fact that $\delta^2 \leq 1/A$. Because $\tilde{\theta} \in S$ is arbitrary, we can write

$$R_{\text{Bayes}}(\pi, L; S) \geq \frac{1}{2} e^{-2p\delta^2} \delta^p \left( 8V \right)^{-p/d} \sup_{0 < \delta \leq A^{-1/2}} \left( \frac{1}{r_\delta(\tilde{\theta})} \right)^{p/d} \pi(d\theta).$$

This proves (66). \qed
We first demonstrate that this corollary yields the correct rate in Example 6.3.

**Example 6.5** (Gaussian model with uniform priors on large balls (continued)). Consider the same setting as in Example 6.3. Because $D(P_\theta || P_{\theta'}) = ||\theta - \theta'||^2/(2\sigma^2)$, we can take $A = (2\sigma^2)^{-1}$ in Corollary 6.4. Moreover, because $L(\theta, a) = ||\theta - a||^2_2$, it is easy to see that $\max_{\theta \in \Theta} \text{Vol}(B_t(a, L)) \leq t^{d/2} \text{Vol}(B)$ which means that we can take $p = 2$ and $V = \text{Vol}(B)$ in Corollary 6.4. Finally, because $w$ is the uniform prior, we have $r_3(\theta) = 1$ for all $\theta \in \Theta$. Inequality (65) therefore gives

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup_{0 < \delta \leq \sqrt{\sigma}} \left(e^{-4d^{-1/d} \text{Vol}(B)^{-2/d}}\right).$$

From here the optimal bound (64) follows because $\text{Vol}(B)^{1/d} \approx d^{-1/2}$.

To further illustrate Corollary 6.4, we apply it to establish Bayes risk lower bounds in the Bayesian Lasso and an example involving generalized linear models.

**Example 6.6** (Bayesian Lasso (Park and Casella, 2008)). Fix $d \geq 1$ and let $\Theta = A = \mathbb{R}^d$ with $L(\theta, a) = ||\theta - a||_2^p$ for a fixed $p > 0$. Also fix $n \geq 1$ and an $n \times d$ matrix $X$ with $\lambda_{\text{max}}$ being the maximum eigenvalue of $X^TX/n$.

For each $\theta \in \Theta$, let $P_\theta$ denote the Gaussian distribution with mean $\theta$ and covariance matrix $\sigma^2 I_n$ where $\sigma^2 > 0$ is a constant. Let $w$ be the $d$-dimensional Laplace distribution with parameter $b$, i.e., $w$ has Lebesgue density $f_w(\theta) = (2b)^d \exp(-||\theta||_1/b)$ where $||\theta||_1 := \sum_{j=1}^d |\theta_j|$. We will apply Corollary 6.4 to prove the following lower bound on the Bayes risk under $w$:

$$R_{\text{Bayes}}(w, L; \theta) \geq \left[d \min \left(\frac{\sigma^2}{n}, b^2\right)\right]^{p/2}. \quad (68)$$

Assume that $\lambda_{\text{max}}$ can be bounded from above by a constant independent of $d$ or $n$. The constant hidden in $\geq$ only depends on $p$. The proof of (68) is given below. The merits of this lower bound will be discussed after the proof.

In order to employ Corollary 6.4 for proving (68), we need to determine the three quantities $A$, $V$ and $r_3(\theta)$. Because, for all $\theta_1, \theta_2 \in \Theta$,

$$D(P_{\theta_1} || P_{\theta_2}) = \frac{||X\theta_1 - X\theta_2||_2^2}{2\sigma^2} \leq \frac{\lambda_{\text{max}} n}{2\sigma^2} ||\theta_1 - \theta_2||_2^2,$$

we can take $A = n\lambda_{\text{max}}/(2\sigma^2)$. It is also easy to see that $V$ can be taken to be $\text{Vol}(B)$ where $B$ is the unit Euclidean ball. It only remains to determine $r_3(\theta)$. By definition, for given $\theta$ and $\delta$,

$$r_3(\theta) = \sup \left\{ \exp \left(- \frac{||\theta_1||_1 - ||\theta_2||_1}{b}\right); ||\theta_1 - \theta||_2 \leq \sqrt{d}\delta \text{ for } i = 1, 2 \right\}.$$

Clearly, whenever $||\theta_1 - \theta||_2 \leq \sqrt{d}\delta$ for $i = 1, 2$,

$$||\theta_1||_1 - ||\theta_2||_1 \leq ||\theta_1 - \theta_2||_1 \leq ||\theta_1 - \theta||_2 + ||\theta_2 - \theta||_2 \leq 2d\delta.$$

As a result, $r_3(\theta)^{-p/d} \geq \exp(-2\delta p/b)$ and Corollary 6.4 allows us to deduce that

$$R_{\text{Bayes}}(w, L; \theta) \geq \frac{1}{2} e^{-2p\delta p/8V} e^{p/d} \exp \left(- \frac{2\delta p}{b}\right) \quad \text{for every } \delta^2 < 1/A.$$

We now make the choice

$$\delta^2 := \min(1/A, b^2) = \min \left(\frac{2\sigma^2}{n\lambda_{\text{max}}}, b^2\right).$$

This implies $\exp(-2\delta p/d) \geq \exp(-2p)$ and we obtain

$$R_{\text{Bayes}}(w, L; \theta) \geq \frac{1}{2} e^{-4p(8V)^{-p/d}} \left[\min \left(\frac{2\sigma^2}{n\lambda_{\text{max}}}, b\right)\right]^{p/2}.$$

Because $V^{1/d} \leq d^{-1/2}$, we deduce

$$R_{\text{Bayes}}(w, L; \Theta) \geq C \left[d \min \left(\frac{2\sigma^2}{n\lambda_{\text{max}}}, b^2\right)\right]^{p/2}$$
where the constant $C > 0$ depends only on $p$. This clearly implies (68).

Let us now briefly discuss the merits of the bound (68). For simplicity, we take $p = 2$ so that $L$ becomes the usual squared error loss and (68) becomes,

$$R_{\text{Bayes}}(w, L; \Theta) \gtrsim d \min \left( \frac{\sigma^2}{n}, b^2 \right).$$

(69)

It may be noted that the term $\sigma^2 d/n$ is the minimax risk of linear regression under the squared error loss. The parameter $b$ characterizes the strength of the prior information. In fact, since $2b^2$ is the variance of a Laplace distribution with parameter $b$, small values of $b$ provide strong prior information that each $\theta_j$ should be concentrated around 0. When $b$ is large, i.e., with less prior information, the lower bound of the Bayes risk in (69) is the same as the minimax risk up to a constant factor. On the other hand, when $b$ is small, i.e., with strong prior information, the lower bound of the Bayes risk becomes $d \cdot b^2$, which is smaller than the minimax risk.

We show below that the lower bound on the Bayes risk in (69) is tight when $b \leq \sigma/\sqrt{d}$ and when the rank of $X$ equals $\min(n, d)$. To see this, consider the estimator

$$\hat{\theta} = \begin{cases} \arg\min_{\theta \in \mathbb{R}^d} \|y - X\theta\|_2^2 & \text{if } b > \sigma/\sqrt{n}; \\ 0 & \text{if } b \leq \sigma/\sqrt{n}. \end{cases}$$

When $b \leq \sigma/\sqrt{n}$, we have $\hat{\theta} = 0$ whose Bayes risk is $\mathbb{E}[\|\theta - 0\|_2^2] = 2db^2$, which shows that (69) is tight. On the other hand, when $b > \sigma/\sqrt{n}$, by our assumption that $b \leq \sigma/\sqrt{d}$, we have $n > d$ and thus $\hat{\theta} = (X^T X)^{-1}X^T y$ which gives $R_{\text{Bayes}} \lesssim db^2/n$, which again proves that (69) is tight.

It may be noted that the condition $b \leq \sigma/\sqrt{d}$ is equivalent to requiring that the prior information is strong enough so that the variance for each $\theta_j$ is bounded above by $\sigma^2/d$. Under the Laplace prior, because $\mathbb{E}[|\theta|_1] = db$, such an assumption also implies that $\theta$ is “sparse” in the sense that the expected $L_1$-norm of $\theta$ is sufficiently small (observe that since we are working in the Bayesian setting, the vector $\theta$, which comes from a Laplace prior, will not contain exactly zero elements with probability 1). The lower bound (69) is therefore tight in this sparse setting.

All our examples so far have included the Gaussian observation model. In the next example, we consider generalized linear models.

**Example 6.7** (Generalized Linear Model). Fix $d \geq 1$ and let $\Theta = \mathcal{A} = \mathbb{R}^d$ with $L(\theta, a) = \|\theta - a\|_p^p$ for a fixed $p > 0$. Also fix $n \geq 1$ and an $n \times d$ matrix $X$ whose rows are written as $x_1^T, \ldots, x_n^T$. As in the last example, $\lambda_{\text{max}}$ denotes the maximum eigenvalue of $X^TX/n$.

For $\theta \in \Theta$, let $P_\theta$ denote the joint distribution of independent random variables $Y_1, \ldots, Y_n$ where $Y_i$ has the density

$$\exp \left[ \frac{y\beta_i - b(\beta_i)}{a(\phi)} + c(y, \phi) \right] \quad \text{for } y \in \mathbb{R}$$

(70)

with $\beta_i = x_i^T \theta$ for $i = 1, \ldots, n$. The parameter $\phi$ is taken to be a constant and the functions $a(\cdot), c(\cdot, \cdot)$ and $b(\cdot)$ are assumed to be known. We assume the existence of a constant $K > 0$ such that $b''(\beta) \leq K$ for all $\beta$ where $b''(\cdot)$ is the second derivative of $b(\cdot)$. This assumption indeed holds for many generalized linear models (e.g., binomial, Gaussian). Also, for some models that violates this assumption (e.g., poisson), the established lower bound still holds by restricting the prior to a compact subset of $\Theta$. We will elaborate this in more details at the end of the example.

Let $w$ denote the Gaussian prior distribution with mean zero and covariance matrix $\tau^2 I_d$. We shall prove here that

$$R_{\text{Bayes}}(w, L; \Theta) \geq C \left[ d \min \left( \frac{a(\phi)}{nK}, \tau^2 \right) \right]^{p/2}$$

(71)

for a constant $C$ that depends only on $p$. This bound can be understood and evaluated in a way that is similar to the bound (68) of the previous example.

The proof of (68) will involve Corollary 6.4 for which we need to determine $A, V$ and $r_3(\theta)$. As before, it is easy to check that $V = \text{Vol}(B)$. To determine $A$, fix a pair $\theta_1, \theta_2$ and, letting $\beta_i^{(j)} = x_i^T \theta_j$ for $j = 1, 2$ and $i = 1, \ldots, n$, observe that

$$D(P_{\theta_1} || P_{\theta_2}) = \frac{1}{a(\phi)} \sum_{i=1}^n \left( b'(\beta_i^{(1)}) \left( \beta_i^{(1)} - \beta_i^{(2)} \right) - \left( b(\beta_i^{(1)}) - b(\beta_i^{(2)}) \right) \right)$$

21
By the second order Taylor expansion of $b(\beta_i^{(2)})$ at the point $\beta_i^{(1)}$, we obtain

$$D(P_{\theta_1}||P_{\theta_2}) = \frac{1}{a(\phi)} \sum_{i=1}^n \frac{b''(\tilde{\beta}_i)}{2} (\beta_i^{(1)} - \beta_i^{(2)})^2$$

where $\tilde{\beta}_i$ lies between $\min(\beta_i^{(1)}, \beta_i^{(2)})$ and $\max(\beta_i^{(1)}, \beta_i^{(2)})$. Now because of our assumption that $b''(\cdot)$ is bounded from above by $K$, we get

$$D(P_{\theta_1}||P_{\theta_2}) \leq \frac{K}{2a(\phi)} ||\beta^{(1)} - \beta^{(2)}||^2 = \frac{K}{2a(\phi)} (\theta_1 - \theta_2)^T X^T X (\theta_1 - \theta_2) \leq nK\lambda_{\max}/2a(\phi) \|\theta_1 - \theta_2\|^2.$$  

We can thus take $A = nK\lambda_{\max}/(2a(\phi))$ in Corollary 6.4. Next we control $r_\delta(\theta)$. For given $\theta$ and $\delta$,

$$r_\delta(\theta) = \sup \left\{ \exp \left( -\frac{1}{2\tau^2} (||\theta^1||^2 - ||\theta^2||^2) \right) : ||\theta^1 - \theta||_2 \leq \sqrt{d}\delta \right\}.$$  

For $\theta_1, \theta_2$ with $||\theta_1 - \theta||_2 \leq \sqrt{d}\delta$, $i = 1, 2$, we have

$$||\theta_1||^2 - ||\theta_2||^2 = ||\theta_1 - \theta||^2_2 + 2\theta^T (\theta_1 - \theta) - ||\theta_2 - \theta||^2_2 + 2\theta^T (\theta_2 - \theta) \leq 2||\theta||_2 (||\theta_1 - \theta||_2 + ||\theta_2 - \theta||_2) \leq 2d\delta^2 + 4\sqrt{d}\delta||\theta||_2.$$  

As a result $r_\delta(\theta) \leq \exp(-p\delta^2/(2\tau^2)) \exp(-2p\delta||\theta||_2/(\tau^2\sqrt{d}))$ and hence

$$\int_{\Theta} \left( \frac{1}{r_\delta(\theta)} \right)^{p/d} w(d\theta) \geq \exp \left( - \frac{p\delta^2}{2\tau^2} \right) \int_{\Theta} \exp \left( - \frac{2p\delta||\theta||_2}{\tau\sqrt{d}} \right) w(d\theta) \geq \exp \left( - \frac{p\delta^2}{2\tau^2} - \frac{4p\delta}{\tau} \right) \int_{\Theta} \mathbf{1} \left\{ ||\theta||_2 < 2\tau\sqrt{d} \right\} w(d\theta).$$

By Chebyshev’s inequality, we have

$$\int_{\Theta} \mathbf{1} \left\{ ||\theta||_2 \geq 2\tau\sqrt{d} \right\} w(d\theta) \leq \frac{1}{4\tau^2 d} \int_{\Theta} ||\theta||^2_2 w(d\theta) = \frac{1}{4}.$$  

Consequently,

$$\int_{\Theta} \left( \frac{1}{r_\delta(\theta)} \right)^{p/d} w(d\theta) \geq \frac{3}{4} \exp \left( - \left( \frac{p\delta^2}{2\tau^2} - \frac{4p\delta}{\tau} \right) \right).$$  

(73)

Corollary 6.4 therefore gives

$$R_{Bayes}(w, L, \Theta) \geq \frac{3}{8} e^{-2p(8V)^{-p/d}} \delta^p \exp \left( - \frac{p\delta^2}{2\tau^2} - \frac{4p\delta}{\tau} \right) \text{ whenever } \delta^2 \leq 1/A.$$  

We make the choice

$$\delta^2 := \min \left( \frac{1}{A}, \frac{\tau^2}{\tau^2} \right) = \min \left( \frac{2a(\phi)}{nK\lambda_{\max}}, \frac{\tau^2}{\tau^2} \right)$$

which implies that the exponential term in the right hand side of (73) is bounded from below by $\exp(-9p/2)$. We thus have

$$R_{Bayes}(w, L, \Theta) \geq \frac{3}{8} e^{-13p/2(8V)^{-p/d}} \left[ \min \left( \frac{2a(\phi)}{nK\lambda_{\max}}, \frac{\tau^2}{\tau^2} \right) \right]^{p/2}.$$  

The inequality (71) now follows because $V^{1/d} \leq d^{-1/2}$. Finally, let us check the assumption that $b''(\beta) \leq K$ under some widely used densities of $Y_i$ in (70). For Gaussian distribution in (70), we have $b(\beta) = \frac{\beta^2}{2}$ so that $b''(\beta) = 1$ for $\beta \in \mathbb{R}$. For binomial distribution, $b(\beta) = \log(1 + \exp(\beta))$ and $b''(\beta) = \frac{\exp(\beta)}{(1 + \exp(\beta))^2} \leq \frac{1}{2}$ for all $\beta \in \mathbb{R}$. However, for Poisson distribution, $b(\beta) = \exp(\beta)$ and thus $b''(\beta) = \exp(\beta)$ is unbounded on $\mathbb{R}$. To address this issue, we restrict the prior to the subset $\Theta = \{ \theta \in \Theta : ||\theta||_2 \leq 2\tau\sqrt{d} \}$ and define the re-scaled prior.
distribution $\pi$ on $\Theta$ as $\pi(S) = w(S)/w(\Theta)$ for any measurable set $S \subset \Theta$. Let $B = \max_{i=1,\ldots,n} \|x_i\|_2$. For any $\beta = x_i^T \theta$ for some $i = 1, \ldots, n$ and $\theta \in \Theta$, we have $b''(\beta) \leq \exp(2\sqrt{\beta}B) = K$. We note that such a restriction of the parameter space will not affect the order of the Bayes risk lower bound. In particular, since now $b''(\beta) \leq K$ when $\theta \in \Theta$, applying the same argument, we obtain the lower bound on $R_{\text{Bayes}}(\pi, \theta, \Theta)$. By (72), we have $w(\Theta) \geq 3/A$ and the lower bound on $R_{\text{Bayes}}(w, L; \Theta)$ can be easily established by noticing that $R_{\text{Bayes}}(w, L; \Theta) \geq w(\Theta)R_{\text{Bayes}}(\pi, L; \Theta) \geq \frac{1}{4} R_{\text{Bayes}}(\pi, L; \Theta)$.

Example 6.8 (Spiked Covariance Model). Fix $\theta = A = B$ where $B$ is the unit Euclidean closed ball of radius one and let $L(\theta, a) := \|\theta - a\|_p^p$ for a fixed $p > 0$. Also fix $n \geq d/2$. For $\theta \in \Theta$, let $P_{\theta}$ denote the joint distribution of independent and identically distributed observations $X_1, \ldots, X_n$ having the Gaussian distribution with zero mean and covariance matrix $\Sigma := I_d + \theta \theta^T$. This is the problem of estimating the principal component for a rank-one spiked covariance model. Let $w$ denote the uniform distribution on $B$. We shall prove below that

$$R_{\text{Bayes}}(w, L; \Theta) \geq C \left[ \min \left( \frac{1}{2}, \frac{d}{n} \right) \right]^{p/2} \tag{74}$$

where $C$ only depends on $p$. The proof is based on the application of (13) with $f(x) = x^2 - 1$, i.e., we use (27). For this, we need to bound the term $\max_{a \in A} w(B_1(a, L))$ and the $f$-informativity corresponding to the chi-squared divergence. It is easy to see that $\max_{a \in A} w(B_1(a, L)) \leq t^{d/p}$.

For the $f$-informativity, we will use the bound (53) with $\alpha = 2$ which requires bounding $M_C(\epsilon, \Theta)$. According to (Guntheringer, 2011a, Theorem 4.6.1), for two Gaussian distributions with mean zero and covariance matrices $\Sigma_1$ and $\Sigma_2$ such that $2\Sigma_1^{-1} - \Sigma_2^{-1}$ is positive definite and $\|\Sigma_1 - \Sigma_2\|^2_2 \leq \frac{1}{2} \lambda_{\min}(\Sigma_2)$, we have

$$\chi^2(\Sigma_1(0, \Sigma_1)||\Sigma_2(0, \Sigma_2)) \leq \exp \left( \frac{\|\Sigma_1 - \Sigma_2\|^2_2}{\lambda_{\min}(\Sigma_2)} \right) - 1. \tag{75}$$

Here $\| \cdot \|_F$ denotes the Frobenius norm defined as $\|A\|_F^2 := \sum_{i,j} a_{ij}^2$ where $A = (a_{ij})$ and $\lambda_{\min}$ denotes the smallest eigenvalue.

Using this result, we get that for $\theta_1, \theta_2 \in \Theta$ (note that $\lambda_{\min}(\Sigma_0) = 1$ for all $\theta$),

$$\chi^2(P_{\theta_1}||P_{\theta_2}) \leq \exp \left( \min n \|\theta_1 - \theta_2\|^2_2 \right) - 1, \tag{76}$$

provided

$$2\Sigma_1^{-1} - \Sigma_2^{-1} \text{ is positive definite and } \|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|^2_F \leq 1/2. \tag{77}$$

In the sequel, whenever we employ (76), the conditions (77) hold. But, for ease of presentation, instead of verifying (77) for every application of (76), we will simply assume (76) and verify the necessary conditions at the end of the proof. Assuming (76), we see that $\chi^2(P_{\theta_1}||P_{\theta_2}) \leq \epsilon^2$ provided $\|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|^2_F \leq \log(1 + \epsilon^2)/n$. Now for $\theta_1, \theta_2 \in \Theta$

$$\|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|^2_F = \|\theta_1 \theta_1^T - \theta_2 \theta_2^T\|^2_F = \|\theta_1 \theta_1^T - \theta_1 \theta_2^T + \theta_1 \theta_2^T - \theta_2 \theta_2^T\|^2_F \leq 2 \left( \|\theta_1\|^2_2 + \|\theta_2\|^2_2 \right) \|\theta_1 - \theta_2\|^2_2 \leq 4 \|\theta_1 - \theta_2\|^2_2.$$

It follows therefore that the $\epsilon^2$-covering number in the chi-squared divergence can be bounded from above by the $\sqrt{\log(1 + \epsilon^2)/(2\sqrt{n})}$-covering number of $B$ under the usual Euclidean norm. Consequently

$$M_C(\epsilon, \Theta) \leq \left( \frac{36n}{\log(1 + \epsilon^2)} \right)^{d/2} \text{ provided } \log(1 + \epsilon^2) \leq 4n.$$

We now set $\epsilon$ to satisfy $\log(1 + \epsilon^2) = \min(n/2, d)$ so that Corollary 5.4 gives

$$I_f(w, P) \leq M_C(\epsilon)(1 + \epsilon^2) - 1 \leq \exp \left( \min \left( \frac{n}{2}, d \right) \right) \left[ \frac{36n}{d} \right]^{d/2} - 1 := \breve{I}.$$

It follows that $\max_{a \in A} w(B_1(a, L)) < \frac{1}{4}(1 + \breve{I})^{-1}$ provided $t = (4(1 + \breve{I}))^{-p/d}$. Inequality (27) then proves

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \left( 4(1 + \breve{I}) \right)^{-p/d} \geq \frac{1}{2}(24c)^{-p} \left[ \min \left( \frac{1}{2}, \frac{d}{n} \right) \right]^{p/2}.$$
which implies (74).

It remains to justify the conditions (77) when we used (76). It should be clear that for this, we only need to verify (77) when

\[ \| \Sigma_{\theta_1} - \Sigma_{\theta_2} \|^2_F \leq \frac{\log(1 + \epsilon^2)}{n} = \min \left( \frac{1}{2} \frac{d}{n} \right). \]  

(78)

We only need to check that \( 2 \Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \) is positive definite under the above condition. For this, observe that by Weyl’s inequality,

\[ \lambda_{\min} \left( 2 \Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \right) \geq \lambda_{\min} \left( 2 \Sigma_{\theta_1}^{-1} \right) - \lambda_{\max} \left( \Sigma_{\theta_2}^{-1} \right) = \frac{2}{1 + \| \theta_1 \|^2} - 1 \geq 0. \]

This implies that \( 2 \Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \) is positive semi-definite and \( \| \theta_1 \|_2 = 1 \) is a necessary condition for \( \lambda_{\min} \left( 2 \Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \right) = 0. \) Under the condition that \( \| \theta_1 \|_2 = 1, \) by Sherman-Morrison formula,

\[ 2 \Sigma_{\theta_2}^{-1} - \Sigma_{\theta_1}^{-1} = I_d - \theta_1 \theta_1^T + \frac{\theta_2 \theta_2^T}{1 + \theta_2^T \theta_2}. \]

It is then easy to check that \( \lambda_{\min} \left( 2 \Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \right) = 0 \) only if \( \theta_2 \) is orthogonal to \( \theta_1. \) However, when \( \| \theta_1 \|_2 = 1 \) and \( \theta_2 \) is orthogonal to \( \theta_1, \) \( \| \Sigma_{\theta_1} - \Sigma_{\theta_2} \|^2_F = \| \theta_1 \|^2 + \| \theta_2 \|^2 > 1, \) which contradicts (78). Therefore \( 2 \Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \) is positive definite and this completes the proof of (74).

7 Conclusions

In this paper, we developed a unified approach to lower bound Bayes risk, which not only serves as a lower bound on minimax risk, but also characterizes the fundamental limitation of the decision problem under a given prior. Our lower bound involves the calculation of upper bounds on \( f \)-informativity. To achieve this goal, we developed upper bounds on \( f \)-informativity for a class of power divergence. Such upper bounds are useful as independent results themselves with potential applications to other information theoretic problems. Our technique also leads to extensions of several classical minimax results. We applied our technique to establish Bayes risk lower bounds for several important estimation problems and compared the obtained Bayes risk lower bounds to the corresponding minimax rates.

As future directions, the theory developed in this paper can potentially be extended to deal with nonparametric priors. It would also be meaningful to derive a general upper bound on \( f \)-informativity (as oppose to the upper bound for a class of power divergences established in Section 5); and it is also interesting to establish lower bounds on \( f \)-informativity which justifies the tightness of the upper bound. We leave them for future works.

Acknowledgements

The authors would like to thank Michael I. Jordan for helpful discussions.

References

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, 28(1):131–142, 1966.

P. Assouad. Deux remarques sur l’estimation. Comptes Rendus de L’Academie des Sciences de Paris, 296:1021–1024, 1983.

L. Birgé. A new lower bound for multiple hypothesis testing. IEEE Transactions on Information Theory, 51(4):1611–1615, 2005.

B. Z. Borovkov and A. U. Sakhanienko. On estimates of the expected quadratic risk. Probab. Math. Statist., 1:185–195, 1980.
G. Braun and S. Pokutta. A general Fano inequality. Preprint; available at http://www.pokutta.com/Homepage/Publications.html, 2014.

L. D. Brown. An information inequality for the Bayes risk under truncated squared error loss. Multivariate Analysis: Future Directions, pages 85–94, 1993.

L. D. Brown and L. Gajek. Information inequalities for the Bayes risk. The Annals of Stat., 18(4):1578–1594, 1990.

L. D. Brown and R. C. Liu. Bounds on the Bayes and minimax risk for signal parameter estimation. IEEE Trans. Inform. Theory, 39(4):1386–1394, 1993.

T. M. Cover and J. A. Thomas. Elements of Information Theory. Wiley, 2nd edition, 2006.

I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der erdodizität von markoffschen ketten. Publ. Math. Inst. Hungar. Acad. Sci., Series A, 8:84–108, 1963.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. Studia Sci. Math Hungar., 2:299–318, 1967.

I. Csiszár. A class of measures of informatitivy of observation channels. Periodica Mathematica Hungarica, 2 (1–4):191–213, 1972.

J. C. Duchi and M. J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. Technical report, UC Berkeley, 2013.

L. Gajek and M. Kaluszka. Lower bounds for the asymptotic Bayes risk in the scale model (with an application to the second-order minimax estimation). The Annals of Statistics, 22(4):1831–1839, 1994.

A. Guntuboyina. Minimax Lower Bounds. PhD thesis, Yale University, 2011a.

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

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

T. S. Han and S. Verdú. Generalizing the fano inequality. IEEE Trans. Inform. Theory, 40:1247–1251, 1994.

D. Haussler and M. Opper. Mutual information, metric entropy and cumulative relative entropy risk. The Annals of Statistics, 25(6):2451–2492, 12 1997.

Z. Ma and Y. Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. Arxiv:1306.3609v1, 2013.

T. Park and G. Casella. The Bayesian Lasso. Journal of the American Statistical Association, 103:681–686, 2008.

M. Sato and M. Akahira. An information inequalities for the Bayes risk. The Annals of Statistics, 24(5):2288–2295, 1996.

Y. Takada. Lower bounds on the Bayes risk for statistical precision problem. Communications in Statistics - Theory and Methods, 28:693–703, 1999.

A. B. Tsybakov. Introduction to Nonparametric Estimation. Springer, 2010.

P. S. Urysohn. Mean width and volume of convex bodives in n-dimensional space. Mat. Sbornik, 31:477–486, 1924.

25
Appendix

Proof of Lemma 3.1. We first fix $b$ and show that $g(a) : a \mapsto \phi_f(a, b)$ is a non-increasing for $a \in [0, b]$. Indeed, for any $a \in (0, b]$, we have,

$$g_L(a) = f_L(a, b) - f_R(\frac{1-a}{1-b}),$$

where $g_L$ and $f_L$ represent left derivatives and $f_R$ represents right derivative (note that $f_L$ and $f_R$ exist because of the convexity of $f$). Because $\frac{a}{b} \leq \frac{1-a}{1-b}$ for every $0 \leq a \leq b$ and $f$ is convex, we see that

$$g_L(a) \leq f_R(\frac{a}{b}) - f_R(\frac{1-a}{1-b}) \leq 0$$

for every $a \in (0, b]$ which implies that $g(a)$ is non-increasing on $[0, b]$. Also, note that the convexity of $f$ implies that $g$ is convex as well.

Next, we fix $a$ and show that $h(b) : b \mapsto \phi_f(a, b)$ is non-decreasing in $b \in [a, 1]$. For any $b \in [a, 1)$, we have,

$$h_R(b) = f(\frac{a}{b}) - f(\frac{1-a}{1-b}) + \frac{1-a}{1-b} f_R(\frac{1-a}{1-b}),$$

where $h_R$ represents the right derivative of $h$. By the convexity of $f$,

$$f(\frac{a}{b}) - f(\frac{1-a}{1-b}) \geq f_R(\frac{1-a}{1-b}) (\frac{a}{b} - \frac{1-a}{1-b}).$$

Combining (79) with (80), we obtain that,

$$h_R(b) \geq \frac{a}{b} (f_L(\frac{1-a}{1-b}) - f_R(\frac{a}{b})) \geq \frac{a}{b} (f_L(\frac{1-a}{1-b}) - f_L(\frac{a}{b}) \geq 0,$$

where the last inequality is because that $\frac{a}{b} \leq \frac{1-a}{1-b}$ for every $0 \leq a \leq b$ and $f$ is convex. The non-negativity of $h_R(b)$ implies that $h(b)$ is non-decreasing on $[a, 1]$.

Lemma 7.1. Recall the definition of $R_Q$ from (15) in Theorem 3.3. The inequality $R_{Bayes,w}(w, L) \leq R_Q$ holds for every probability measure $Q$ on $\mathcal{X}$.

Proof. Note that $R_Q$ is defined in terms of the Bayes decision rule $\mathfrak{d}_w$ under the prior $w$. Fix an arbitrary action $a^* \in \mathcal{A}$ and define the decision rule $\mathfrak{d}(x) := a^*$ for all $x \in \mathcal{X}$. Because $\mathfrak{d}_w$ is the Bayes rule for the prior $w$, its Bayes risk must be smaller than the Bayes risk of $\mathfrak{d}$. This gives the inequality

$$\int_{\mathcal{X}} E_{\mathfrak{d}} L(\theta, \mathfrak{d}_w(X))w(d\theta) \leq \int_{\mathcal{X}} L(\theta, a^*)w(d\theta).$$

The above is true for every $a^* \in \mathcal{A}$; so we fix $x \in \mathcal{X}$ and use the above with $a^* = \mathfrak{d}_w(x)$ to get

$$\int_{\mathcal{X}} E_{\mathfrak{d}} L(\theta, \mathfrak{d}_w(X))w(d\theta) \leq \int_{\mathcal{X}} L(\theta, \mathfrak{d}_w(x))w(d\theta).$$

Integrating both sides of the above with respect to $Q(dx)$, we obtain that $R_{Bayes,w}(w, L) \leq R_Q$. 

H. Van Trees. Detection, Estimation and Modulation Theory. Wiley, 1968.

B. Vidakovic and A. DasGupta. Lower bounds on Bayes risk for estimating a normal variable: With applications. The Canadian Journal of Statistics, 23(3):269–282, 1995.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. The Annals of Statistics, 27(5):1564–1599, 1999.

B. Yu. Assouad, fano, and le cam. In Festschrift for Lucien Le Cam, pages 423–435. Springer, 1997.
Proof of Lemma 5.2. Let $\phi(t) \equiv t^p$ with $\phi'(t) = pt^{p-1}$ and $\phi''(t) = p(p-1)t^{p-2}$ and $\varphi(t) = t^{1/p}$ with $\varphi'(t) = \frac{1}{p(t^{1-p})/p}$. Then

$$f(u) = \varphi \left( \int_T \phi(u(t)) \mu(dt) \right).$$

To prove the concavity of $f(u)$, considering the scalar function

$$h(s) = \varphi \left( \int_T \phi(u(t)) \mu(dt) \right),$$

for arbitrary $u, v \in L^p_\mu(T)$. We notice that concavity of $f$ is equivalent to concavity at zero for all functions of the form $h$, and we therefore only have to show that $h''(0) \leq 0$. Let $r(s) = \int_T \phi(u(t) + sv(t)) \mu(dt)$,

$$h'(s) = \varphi'(r(s)) \int_T \phi'(u(t) + sv(t)) v(t) \mu(dt)$$

$$h''(s) = \varphi''(r(s)) \left( \int_T \phi'(u(t) + sv(t)) v(t) \mu(dt) \right)^2$$

$$+ \varphi'(r(s)) \int_T \phi''(u(t) + sv(t)) v^2(t) \mu(dt)$$

By plugging in the definitions of $\phi(t), \varphi(t), r(s)$ and setting $s = 0$, we have

$$h''(0) = \frac{1 - p}{f(u)} \left( \left( f(u)^{1-p} \int_T u(t)^{p-1} v(t) \mu(dt) \right)^2 - f(u)^{2-p} \int_T u(t)^{p-2} v^2(t) \mu(dt) \right)$$

Applying the Cauchy-Schwarz inequality

$$\left( \int_T a(t) b(t) \mu(dt) \right)^2 \leq \left( \int_T a(t)^2 \mu(dt) \right) \left( \int_T b(t)^2 \mu(dt) \right)$$

with $a(t) = \left( \frac{f(u)}{u(t)} \right)^{-p/2}$ and $b(t) = v(t) \left( \frac{f(u)}{u(t)} \right)^{1-p/2}$ and noticing that $p < 1$, we have $h''(0) \leq 0$, which completes the proof. \qed