Journal of Machine Learning Research 17 (2016) 1-58 Submitted 4/16; Revised 10/16; Published 12/16

On Bayes Risk Lower Bounds

Xi Chen XCHEN3@STERN.NYU.EDU
Stern School of Business
New York University
New York, NY 10012, USA

Adityanand Guntuboyina ADITYA@STAT.BERKELEY.EDU
Department of Statistics
University of California
Berkeley, CA 94720, USA

Yuchen Zhang ZHANGYUC@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Editor: Edo Airoldi

Abstract

This paper provides a general technique for lower bounding the Bayes risk of statistical estimation, applicable to arbitrary loss functions and arbitrary prior distributions. A lower bound on the Bayes risk not only serves as a lower bound on the minimax risk, but also characterizes the fundamental limit of any estimator given the prior knowledge. Our bounds are based on the notion of f-informativity (Csiszár, 1972), which is a function of the underlying class of probability measures and the prior. Application of our bounds requires upper bounds on the f-informativity, thus we derive new upper bounds on f-informativity which often lead to tight Bayes risk lower bounds. Our technique leads to generalizations of a variety of classical minimax bounds (e.g., generalized Fano's inequality). Our Bayes risk lower bounds can be directly applied to several concrete estimation problems, including Gaussian location models, generalized linear models, and principal component analysis for spiked covariance models. To further demonstrate the applications of our Bayes risk lower bounds to machine learning problems, we present two new theoretical results: (1) a precise characterization of the minimax risk of learning spherical Gaussian mixture models under the smoothed analysis framework, and (2) lower bounds for the Bayes risk under a natural prior for both the prediction and estimation errors for high-dimensional sparse linear regression under an improper learning setting.

Keywords: Bayes risk, Minimax risk, f-divergence, f-informativity, Fano's inequality, Smoothed analysis

1. Introduction

Consider a standard setting where we observe data points X taking values in a sample space $\mathcal{X}$. The distribution of X depends on an unknown parameter $\theta \in \Theta$ and is denoted by P_θ. The goal is to compute an estimate of θ based on the observed samples. Formally, we denote the estimator by $\mathfrak{d}(X)$, where $\mathfrak{d} : \mathcal{X} \to \Theta$ is a mapping from the sample space to the parameter space. The risk of the estimator is defined by $\mathbb{E}_\theta L(\theta, \mathfrak{d}(X))$ where L :

$\Theta \times \mathcal{A} \mapsto [0, \infty)$ is a non-negative loss function. This framework applies to a broad scope of machine learning problems. Taking sparse linear regression as a concrete example, the data X represents the design matrix and the response vector; the parameter space is the set of sparse vectors; the loss function can be chosen as a squared loss.

Given an estimation problem, we are interested in the lowest possible risk achievable by any estimator, which will be useful in justifying the potential of improving existing algorithms. The classical notion of optimality is formalized by the so-called *minimax risk*. More specifically, we assume that the statistician chooses an optimal estimator $\mathfrak{d}$, then the adversary chooses the worst parameter θ by knowing the choice of $\mathfrak{d}$. The minimax risk is defined as:

$$R_{\text{minimax}}(L; \Theta) := \inf_{\mathfrak{d}} \sup_{\theta \in \Theta} \mathbb{E}_\theta L(\theta, \mathfrak{d}(X)). \tag{1}$$

The minimax risk has been determined up to multiplicative constants for many important problems. Examples include sparse linear regression (Raskutti et al., 2011), classification (Yang, 1999), additive models over kernel classes (Raskutti et al., 2012), and crowdsourcing (Zhang et al., 2016).

The assumption that the adversary is capable of choosing a worst-case parameter is sometimes over-pessimistic. In practice, the parameter that incurs a worst-case risk may appear with very small probability. To capture the hardness of the problem with this prior knowledge, it is reasonable to assume that the true parameter is sampled from an underlying prior distribution w. In this case, we are interested in the *Bayes risk* of the problem. That is, the lowest possible risk when the true parameter is sampled from the prior distribution:

$$R_{\text{Bayes}}(w, L; \Theta) := \inf_{\mathfrak{d}} \int_{\Theta} \mathbb{E}_\theta L(\theta, \mathfrak{d}(X)) w(\mathrm{d}\theta). \tag{2}$$

If the prior distribution w is known to the learner, then the Bayes estimator attains the Bayes risk (Berger, 2013). But in general, the Bayes estimator is computationally hard to evaluate, and the Bayes risk has no closed-form expression. It is thus unclear what is the fundamental limit of estimators when the prior knowledge is available.

In this paper, we present a technique for establishing lower bounds on the Bayes risk for a general prior distribution w. When the lower bound matches the risk of any existing algorithm, it captures the convergence rate of the Bayes risk. The Bayes risk lower bounds are useful for three main reasons:

1. They provide an idea of the difficulty of the problem under a specific prior w.
2. They automatically provide lower bounds for the minimax risk and, because the minimax regret is always larger than or equal to the minimax risk (see, for example, Rakhlin et al. (2013)), they also yield lower bounds for the minimax regret.
3. As we will show, they have an important application in establishing the minimax lower bound under the *smoothed analysis framework*.

Throughout this paper, when the loss function L and the parameter space Θ are clear from the context, we simply denote the Bayes risk by $R_{\text{Bayes}}(w)$. When the prior w is also clear, the notation is further simplified to R.

1.1 Our Main Results

In order to give the reader a flavor of the kind of results proved in this paper, let us consider Fano's classical inequality (Han and Verdú, 1994; Cover and Thomas, 2006; Yu, 1997) which is one of the most widely used Bayes risk lower bounds in statistics and information theory. The standard version of Fano's inequality applies to the case when $\Theta = \mathcal{A} = \{1, \ldots, N\}$ for some positive integer N with the indicator loss $L(\theta, a) := \mathbb{I}\{\theta \neq a\}$ ($\mathbb{I}$ stands for the zero-one valued indicator function) and the prior w being the discrete uniform distribution on Θ. In this setting, Fano's inequality states that

$$R_{\text{Bayes}}(w) \geq 1 - \frac{I(w, \mathcal{P}) + \log 2}{\log N} \tag{3}$$

where $I(w, \mathcal{P})$ is the mutual information between the random variables $\theta \sim w$ and X with $X|\theta \sim P_\theta$ (note that this mutual information only depends on w and $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ which is why we denote it by $I(w, \mathcal{P})$). Fano's inequality implies that when $I(w; \mathcal{P})$ is large i.e., when the information that X has about θ is large, then the risk of estimation is small.

A natural question regarding Fano's inequality, which does not seem to have been asked until very recently, is the following: does there exist an analogue of (3) when w is not necessarily the uniform prior and/or when Θ and $\mathcal{A}$ are arbitrary sets, and/or when the loss function is not necessarily $\mathbb{I}\{\theta \neq a\}$? An interesting result in this direction is the following inequality which has been recently proved by Duchi and Wainwright (2013) who termed it the continuum Fano inequality. This inequality applies to the case when $\Theta = \mathcal{A}$ is a subset of Euclidean space with finite strictly positive Lebesgue measure, $L(\theta, a) = \mathbb{I}\{\|\theta - a\|_2 \geq \epsilon\}$ for a fixed $\epsilon > 0$ ($\|\cdot\|_2$ is the usual Euclidean metric) and the prior w being the uniform probability measure (i.e., normalized Lebesgue measure) on Θ. In this setting, Duchi and Wainwright (2013) proved that

$$R_{\text{Bayes}}(w) \geq 1 + \frac{I(w, \mathcal{P}) + \log 2}{\log \left(\sup_{a \in \mathcal{A}} w\{\theta \in \Theta : \|\theta - a\|_2 < \epsilon\}\right)}. \tag{4}$$

It turns out that there is a very clean connection between inequalities (3) and (4). Indeed, both these inequalities are special instances of the following inequality:

$$R_{\text{Bayes}}(w) \geq 1 + \frac{I(w, \mathcal{P}) + \log 2}{\log \left(\sup_{a \in \mathcal{A}} w\{\theta \in \Theta : L(\theta, a) = 0\}\right)} \tag{5}$$

Indeed, the term $w\{\theta \in \Theta : L(\theta, a) = 0\}$ equal to $1/N$ in the setting of (3) and it is equal to $w\{\theta \in \Theta : \|\theta - a\|_2 < \epsilon\}$ in the setting of (4).

Since both (3) and (4) are special instances of (5), one might reasonably conjecture that inequality (5) might hold more generally. In Section 3, we give an affirmative answer by proving that inequality (5) holds for any zero-one valued loss function L and any prior w. No assumptions on Θ, $\mathcal{A}$ and w are needed. We refer to this result as *generalized Fano's inequality*. Our proof of (5) is quite succinct and is based on the data processing inequality (Cover and Thomas, 2006; Liese, 2012) for Kullback-Leibler (KL) divergence. The use of the data processing inequality for proving Fano-type inequalities was introduced by Gushchin (2003).

The data processing inequality is not only available for the KL divergence. It can be generalized to any divergence belonging to a general family known as f-divergences (Csiszár, 1963; Ali and Silvey, 1966). This family includes the KL divergence, chi-squared divergence, squared Hellinger distance, total variation distance and power divergences as special cases. The usefulness of f-divergences in machine learning has been illustrated in Reid and Williamson (2011); Garcıa-Garcıa and Williamson (2012); Reid and Williamson (2009).

For every f-divergence, one can define a quantity called f-informativity (Csiszár, 1972) which plays the same role as the mutual information for KL divergence. The precise definitions of f-divergences and f-informativities are given in Section 2. Utilizing the data processing inequality for f-divergence, we prove general Bayes risk lower bounds which hold for every zero-one valued loss L and for arbitrary Θ, $\mathcal{A}$ and w (Theorem 2). The generalized Fano's inequality (5) is a special case by choosing the f-divergence to be KL. The proposed Bayes risk lower bounds can also be specialized to other f-divergences and have a variety of interesting connections to existing lower bounds in the literature such as Le Cam's inequality, Assouad's lemma (see Theorem 2.12 in Tsybakov (2010)), Birgé-Gushchin inequality (Gushchin, 2003; Birgé, 2005). These results are provided in Section 3.

In Section 4, we deal with nonnegative valued loss functions L which are not necessarily zero-one valued. Basically, we use the standard method of lower bounding the general loss function L by a zero-one valued function and then use our results from Section 3 for lower bounding the Bayes risk. This technique, in conjunction with the generalized Fano's inequality, gives the following lower bound (proved in Corollary 12)

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\{\theta : L(\theta, a) < t\} \leq \frac{1}{4} e^{-2I(w,\mathcal{P})} \right\}. \tag{6}$$

A special case of the above inequality has appeared previously in Zhang (2006, Theorem 6.1) (please refer to Remark 13 for a detailed explanation of the connection between inequality (6) and (Zhang, 2006, Theorem 6.1)).

We also prove analogues of the above inequality for different f divergences. Specifically, using our f-divergence inequalities from Section 3, we prove, in Theorem 9, the following inequality which holds for every f divergence:

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\{\theta : L(\theta, a) < t\} < 1 - u_f(I_f(w, \mathcal{P})) \right\} \tag{7}$$

where $I_f(w, \mathcal{P})$ represents the f-informativity and $u_f(\cdot)$ is a non-decreasing $[0, 1]$-valued function that depends only on f. This function $u_f(\cdot)$ (see its definition from (31)) can be explicitly computed for many f-divergences of interest, which gives useful lower bounds in terms of f-informativity. For example, for the case of KL divergence and chi-squared divergence, inequality (7) gives the lower bound in (6) and the following inequality respectively,

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\{\theta : L(\theta, a) < t\} \leq \frac{1}{4(1 + I_{\chi^2}(w, \mathcal{P}))} \right\}. \tag{8}$$

where $I_{\chi^2}(w, \mathcal{P})$ is the chi-squared informativity.

Intuitively, inequality (7) shows that the Bayes risk is lower bounded by half of the largest possible t such that the maximum prior mass of any t-radius "ball" ($w\{\theta : L(\theta, a) < t\}$) is

less than some function of f-informativity. To apply (7), one needs to obtain upper bounds on the following two quantities:

1. The "small ball probability" $\sup_{a\in\mathcal{A}} w\{\theta : L(\theta, a) < t\}$, which does not depend of the family of probability measures $\mathcal{P}$.

2. The f-informativity $I_f(w, \mathcal{P})$, which does not depend on the loss function L.

We note that a nice feature of (7) is that L and $\mathcal{P}$ play separately roles. One may first obtain an upper bound I_f^{up} for the f-informativity $I_f(w, \mathcal{P})$, then choose t so that the small ball probability $w\{\theta : L(\theta, a) < t\}$ can be bounded from above by $1 - u_f(I_f^{\mathrm{up}})$. The Bayes risk will be bounded from below by $t/2$. It is noteworthy that the terminology "small ball probability" was used by Xu and Raginsky (2014) (this paper proved information-theoretic lower bounds on the minimum time in a distributed function computation problem).

We do not have a general guideline for bounding the small ball probability. It needs to be dealt with case by case based on the prior and the loss function. But for upper bounding the f-informativity, we offer a general recipe in Section 5 for a subclass of divergences of interest (power divergences for $\alpha \notin [0, 1)$), which covers the chi-squared divergence as one of the most important divergences in our applications. These bounds generalize results of Haussler and Opper (1997) and Yang and Barron (1999) for mutual information to f-informativities involving power divergences. As an illustration of our techniques (inequality (7) combined with the f-informativity upper bounds), we apply them to a concrete estimation problem in Section 5. We further apply our results to several popular machine learning and statistics problems (e.g., generalized linear model, spiked covariance model, and Gaussian model with general loss) in Appendix C.

In Section 6 and Section 7, we present non-trivial applications of our Bayes risk lower bounds to two learning problems: the first one is a unsupervised learning problem, while the second one is a supervised learning problem. Section 6 studies smoothed analysis for learning mixtures of spherical Gaussians with uniform weights. Although learning mixtures of Gaussians is a computationally hard problem, it has been shown recently by Hsu and Kakade (2013) that under the assumptions that the Gaussian means are linearly independent, it can be learnt in polynomial time by a spectral method. We perform a smoothed analysis on a variant of the algorithm (Hsu and Kakade, 2013), showing that the linear independence assumption can be replaced by perturbing the true parameters by a small random noise. The method described in Section 6 achieves a better convergence rate than the original algorithm of Hsu and Kakade (2013). Furthermore, we apply the Bayes risk lower bound techniques to show that the algorithm's convergence rate is unimprovable, even under smoothed analysis (i.e. when the true parameters are randomly perturbed). Section 6 highlights the usefulness of our techniques in proving lower bounds for smoothed analysis, which appears to be challenging using traditional techniques of the minimax theory.

In Section 7, we consider the high-dimensional sparse linear regression problem and we provide Bayes risk lower bounds for both prediction error and estimation error under a natural prior on the regression parameter belonging to the set of k-sparse vectors. Although lower bounds for sparse linear regression have been well-studied (see, e.g., Raskutti et al. (2011); Zhang et al. (2014) and references therein), these bounds only focus on the minimax or the worst-case scenario and thus are too pessimistic in practice. Indeed, the parameters

that usually attain these minimax lower bounds have zero probability under any continuous prior, so that their average effects might be negligible. The fundamental limits of sparse linear regression under a realistic prior is, to the best of the our knowledge, unknown. The developed tool of lower bounding Bayes risks can be directly applied to characterize these limits. Moreover, our Bayes risk lower bound is flexible in the sense that by tuning the variance of the prior of non-zero elements of θ, it provides a wide spectrum of lower bounds. For one particular choice of the variance, our Bayes risk lower bounds match the minimax risk lower bounds. This gives a natural *least favorable prior* for sparse linear regression, while the known least favorable prior in Raskutti et al. (2011) is a non-constructive discrete prior over a packing set of the parameter space that cannot be sampled from. We also work under the *improper learning* setting where we allow non-sparse estimators for the true regression vector (even though the true regression vector is assumed to be sparse).

1.2 Related Works

Before finishing this introduction section, we briefly describe related work on Bayes risk lower bounds. There are a few results dealing with special cases of 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 and Gajek (1990); Brown (1993); Gill and Levit (1995); Sato and Akahira (1996); Takada (1999). A few additional papers dealt with even more specialized problems e.g., Gaussian white noise model (Brown and Liu, 1993), scale models (Gajek and Kaluszka, 1994) and estimating Gaussian variance (Vidakovi and DasGupta, 1995). Most of these results are based on the van Trees inequality (see Gill and Levit (1995) and Theorem 2.13 in Tsybakov (2010)). Although the van Trees inequality usually leads to sharp constant in the Bayes risk lower bounds, it only applies to weighted quadratic loss functions (as its proof relies on Cauchy-Schwarz inequality) and requires the underlying Fisher information to be easily computable, which limits its applicability. There is also a vast body of literature on minimax lower bounds (see, e.g., Tsybakov (2010)) which can be viewed as Bayes risk lower bounds for certain priors. These priors are usually discrete and specially constructed so that the lower bounds do not apply to more general (continuous) priors. Another related area of work involves finding lower bounds on posterior contraction rates (see, e.g., Castillo (2008)).

1.3 Outline of the Paper

The rest of the paper is organized in the following way. In Section 2, we describe notations and review preliminaries such as f-divergences, f-informativity, data processing inequality, etc. Section 3 deals with inequalities for zero-one valued loss functions. These inequalities have many connections to existing lower bound techniques. Section 4 deals with nonnegative loss functions and we provide inequality (7) and its special cases. Section 5 presents upper bounds on the f-informativity for power divergences for $\alpha \notin [0, 1)$. Some examples are also given in this section. Section 6 studies smoothed analysis for learning mixtures of spherical Gaussians with uniform weights using our technique. We conclude the paper in Section 1.3. Due to space constraints, we have relegated some proofs and additional examples and results to the appendix.

2. Preliminaries and Notations

We first review the notions of f-divergence (Csiszár, 1963; Ali and Silvey, 1966) and f-informativity (Csiszár, 1972). Let $\mathcal{C}$ denote the class of all convex functions $f : (0, \infty) \to \mathbb{R}$ which satisfy $f(1) = 0$. 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 μ, the f-divergence $D_f(P||Q)$ between P and Q is defined as follows:

$$D_f(P||Q) := \int f\left(\frac{p}{q}\right) q \mathrm{d}\mu + f'(\infty) P\{q = 0\}. \tag{9}$$

We note that the convention $0 \cdot \infty = 0$ is adopted here so that $f'(\infty)P\{q = 0\} = 0$ when $f'(\infty) = \infty$ and $P\{q = 0\} = 0$. Note that $D_f(P\|Q) = +\infty$ when $f'(\infty) = +\infty$ and $P\{q = 0\} > 0$. Also note that $f(1) = 0$ implies that $D_f(P\|Q) = 0$ when $P = Q$.

Certain divergences are commonly used because they can be easily computed or bounded when P and Q are product measures. These divergences are the power divergences corresponding to the functions f_α defined by

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

Popular examples of power divergences include:

1) Kullback-Leibler (KL) divergence: $\alpha = 1$, $D_{f_1}(P||Q) = \int p \log(p/q) \mathrm{d}\mu$ if P is absolutely continuous with respect to Q (and it is infinite if P is not absolutely continuous with respect to Q). Following the conventional notation, we denote the KL divergence by $D(P||Q)$ (instead of $D_{f_1}(P||Q)$).

2) Chi-squared divergence: $\alpha = 2$, $D_{f_2}(P||Q) = \int (p^2/q) \mathrm{d}\mu - 1$ 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 $\chi^2(P||Q)$ following the conventional notation.

3) When $\alpha = 1/2$, one has $D_{f_{1/2}}(P||Q) = 1 - \int \sqrt{pq} \mathrm{d}\mu$ which is a half of the squared Hellinger distance. That is, $D_{f_{1/2}}(P||Q) = H^2(P||Q)/2$, where $H^2(P||Q) = \int (\sqrt{p} - \sqrt{q})^2 \mathrm{d}\mu$ is the squared Hellinger distance between P and Q.

The total variation distance $\|P - Q\|_{TV}$ is another 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 (1972) and Liese (2012, Theorem 3.1)) which states the following: let $\mathcal{X}$ and $\mathcal{Y}$ be two measurable spaces and let $\Gamma : \mathcal{X} \to \mathcal{Y}$ be a measurable function. For every $f \in \mathcal{C}$ and every pair of probability measures P and Q on $\mathcal{X}$, we have

$$D_f(P\Gamma^{-1}||Q\Gamma^{-1}) \leq D_f(P||Q), \tag{10}$$

where $P\Gamma^{-1}$ and $Q\Gamma^{-1}$ denote the *induced measures* of Γ on $\mathcal{Y}$, i.e., for any measurable set B on the space $\mathcal{Y}$, $P\Gamma^{-1}(B) := P(\Gamma^{-1}(B))$, $Q\Gamma^{-1}(B) := Q(\Gamma^{-1}(B))$ (see the definition of induced measure from Definition 2.2.1. in Athreya and Lahiri (2006)).

Next, we introduce the notion of f-informativity (Csiszár, 1972). 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 Θ. For each $f \in \mathcal{C}$, the f-informativity, $I_f(w, \mathcal{P})$, is defined as

$$I_f(w, \mathcal{P}) = \inf_Q \int D_f(P_\theta || Q) w(\mathrm{d}\theta), \tag{11}$$

where the infimum is taken over all possible probability measures Q on $\mathcal{X}$. When $f(x) = x \log x$ (so that the corresponding f-divergence is the KL divergence), the f-informativity is equal to the mutual information and is denoted by $I(w, \mathcal{P})$. We denote the informativity corresponding to the power divergence D_{f_α} by $I_{f_\alpha}(w, \mathcal{P})$. For the special case $\alpha = 2$, we use the more suggestive notation $I_{\chi^2}(w, \mathcal{P})$. The informativity corresponding to the total variation distance will be denoted by $I_{TV}(w, \mathcal{P})$.

Additional notations and definitions are described as follows. Recall the Bayes risk (2) and the minimax risk (1). When the loss function L and parameter space Θ are clear from the context, we drop the dependence on L and Θ. When the prior w is also clear from the context, we denote the Bayes risk by R and the minimax risk by R_{minimax}. We need certain notation for 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 ϵ^2-cover of $\{P_\theta, \theta \in S\}$ under the f-divergence i.e.,

$$\sup_{\theta \in S} \min_{1 \le j \le M} D_f(P_\theta || Q_j) \le \epsilon^2. \tag{12}$$

We write the covering number as $M_{KL}(\epsilon, S)$ when $f(x) = x \log x$ and $M_{\chi^2}(\epsilon, S)$ when $f(x) = x^2 - 1$. We write $M_\alpha(\epsilon, S)$ when $f = f_\alpha$ for other $\alpha \in \mathbb{R}$. We note that $\log M_f(\epsilon, S)$ is an upper bound on the metric entropy. The quantity $M_f(\epsilon, S)$ can be infinite if S is arbitrary. For a vector $x = (x_1, \ldots, x_d)$ and a real number $p \ge 1$, denote by $\|x\|_p$ the ℓ_p-norm of x. In particular, $\|x\|_2$ denotes the Euclidean norm of x. $\mathbb{I}(A)$ denotes the indicator function which takes value 1 when A is true and 0 otherwise. We use C, c, etc. to denote generic constants whose values might change from place to place.

3. Bayes Risk Lower Bounds for Zero-one Valued Loss Functions and Their Applications

In this section, we consider zero-one loss functions L and present a principled approach to derive Bayes risk lower bounds involving f-informativity for every $f \in \mathcal{C}$. Our results hold for any given prior w and zero-one loss L. By specializing the f-divergence to KL divergence, we obtain the generalized Fano's inequality (5). When specializing to other f-divergences, our bounds lead to some classical minimax bounds of Le Cam and Assouad (Assouad, 1983), more recent minimax results of Gushchin (2003); Birgé (2005) and also results in Tsybakov (2010, Chapter 2). Bayes risk lower bounds for general nonnegative loss functions will be presented in the next section.

We need additional notations to state the main results of this section. For each $f \in \mathcal{C}$, let $\phi_f : [0,1]^2 \to \mathbb{R}$ be the function defined in the following way: for $a, b \in [0,1]^2$, $\phi_f(a, b)$ is the

f-divergence between the two probability measures P and Q on $\{0,1\}$ given by $P\{1\} = a$ and $Q\{1\} = b$. By the definition (9), it is easy to see that $\phi_f(a,b)$ has the following expression (recall that $f'(\infty) := \lim_{x\uparrow\infty} f(x)/x$):

$$\phi_f(a,b) = \begin{cases} bf\left(\frac{a}{b}\right) + (1-b)f\left(\frac{1-a}{1-b}\right) & \text{for} \quad 0 < b < 1; \\ f(1-a) + af'(\infty) & \text{for} \quad b = 0; \\ f(a) + (1-a)f'(\infty) & \text{for} \quad b = 1. \end{cases} \tag{13}$$

The convexity of f implies monotonicity and convexity properties of ϕ_f, which is stated in the following lemma.

Lemma 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 and continuous in a. Further, for every fixed a, the map $h(b) : b \mapsto \phi_f(a,b)$ is non-decreasing for $b \in [a,1]$.*

We also define the quantity

$$R_0 := \inf_{a\in\mathcal{A}} \int_\Theta L(\theta,a) w(\mathrm{d}\theta), \tag{14}$$

where the decision a does not depend on data X. Note that R_0 represents the Bayes risk with respect to w in the "no data" problem i.e., when one only has information on Θ, $\mathcal{A}$, L and the prior w but not the data X. For simplicity, our notation for R_0 suppresses its dependence on w. Because the loss function is zero-one valued so that $L(\theta,a) = 1 - \mathbb{I}(L(\theta,a) = 0)$, the quantity R_0 has the following alternative expression:

$$R_0 = 1 - \sup_{a\in\mathcal{A}} w(B(a)), \tag{15}$$

where

$$B(a) := \{\theta \in \Theta : L(\theta,a) = 0\}, \tag{16}$$

and $w(B(a))$ is the prior mass of the "ball" $B(a)$. It will be important in the sequel to observe that the Bayes risk, $R_{\text{Bayes}}(w)$ is bounded from above by R_0. This is obvious because the risk with some data cannot be greater than the risk in the no data problem (which can be viewed as an application of the data processing inequality). Formally, if $\mathcal{D} = \{\mathfrak{d} : \exists a \in \mathcal{A} \text{ such that } \mathfrak{d}(x) = a \ \forall x \in \mathcal{X}\}$ is the class of the constant decision rules, then $R_0 = \inf_{\mathfrak{d}\in\mathcal{D}} \int_\Theta \mathbb{E}_\theta L(\theta, \mathfrak{d}(X)) w(\mathrm{d}\theta) \geq R_{\text{Bayes}}(w)$. Because $0 \leq R_{\text{Bayes}}(w) \leq R_0$, we have $R_{\text{Bayes}}(w) = 0$ when $R_0 = 0$. We shall therefore assume throughout this section that $R_0 > 0$.

The main result of this section is presented next. It provides an implicit lower bound for the Bayes risk in terms of R_0 and the f-informativity $I_f(w,\mathcal{P})$ for every $f \in \mathcal{C}$. The only assumption is that L is zero-one valued and we do not assume the existence of the Bayes decision rule.

Theorem 2 *Suppose that the loss function L is zero-one valued. For any $f \in \mathcal{C}$, we have*

$$I_f(w,\mathcal{P}) \geq \phi_f(R_{\text{Bayes}}(w), R_0) \tag{17}$$

where ϕ_f and R_0 are defined (13) *and* (14) *respectively.*

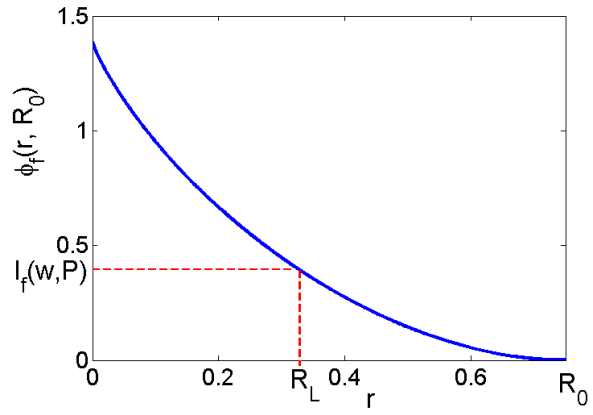


Figure 1: Illustration on why (17) leads to a lower bound on $R_{\text{Bayes}}(w)$. Recall that $R \le R_0$ and $r \mapsto \phi_f(r, R_0)$ is non-increasing in r for $r \in [0, R_0]$. Given $I_f(w, \mathcal{P})$ as an upper bound of $\phi_f(R_{\text{Bayes}}(w), R_0)$, we have $R_{\text{Bayes}}(w) \ge R_L = g^{-1}(I_f(w, \mathcal{P}))$ and thus R_L serves as a Bayes risk lower bound.

Before we prove Theorem 2, we first show that the inequality (17) indeed provides an implicit lower bound for the Bayes risk $R := R_{\text{Bayes}}(w)$ since $R \le R_0$ and $r \mapsto \phi_f(r, R_0)$ is non-increasing in r for $r \in [0, R_0]$ (Lemma 1). Therefore, let $g(r) := \phi_f(r, R_0)$. We have

$$R_{\text{Bayes}}(w) \ge g^{-1}(I_f(w, \mathcal{P})), \tag{18}$$

where $g^{-1}(x) := \inf\{0 \le r \le R_0, g(r) \le x\}$ is the generalized inverse function of the non-increasing $g(r)$. As an illustration, we plot $\phi_f(r, R_0)$ for $f(x) = x \log x$ and the corresponding Bayes risk lower bound $g^{-1}(I_f(w, \mathcal{P}))$ in Figure 1. The lower bound (18) can be immediately applied to obtain Bayes risk lower bounds when the f-divergence in (17) is chi-squared divergence, total variation distance, or Hellinger distance (see Corollary 7). However, for the KL divergence, there is no simple form of $g^{-1}(x)$. To obtain the corresponding Bayes risk lower bound, we can invert (17) by utilizing the convexity of $g(r)$, which will give a generalized Fano's inequality (see Corollary 5). In particular, since $r \mapsto \phi_f(r, R_0)$ is convex (see Lemma 1),

$$\phi_f(R, R_0) \ge \phi_f(r, R_0) + \phi_f'(r-, R_0)(R - r) \quad \text{for every } 0 < r \le R_0$$

where $\phi_f'(r-, R_0)$ denotes the left derivative of $x \mapsto \phi_f(x, R_0)$ at $x = r$. The monotonicity of $\phi_f(r, R_0)$ in r (Lemma 1) gives $\phi_f'(r-, R_0) \le 0$ and we thus have,

$$R \ge r + \frac{\phi_f(R, R_0) - \phi_f(r, R_0)}{\phi_f'(r-, R_0)} \quad \text{for every } 0 < r \le R_0.$$

Inequality (17) $I_f(w, \mathcal{P}) \ge \phi_f(R, R_0)$ can now be used to deduce that (note that $\phi_f'(r-, R_0) \le 0$)

$$R \ge r + \frac{I_f(w, \mathcal{P}) - \phi_f(r, R_0)}{\phi_f'(r-, R_0)} \quad \text{for every } 0 < r \le R_0. \tag{19}$$

The inequalities (18) and (19) provide general approaches to convert (17) to an explicit lower bound on R.

Theorem 2 is new, but its special case $\Theta = \mathcal{A} = \{1, \ldots, N\}$, $L(\theta, a) := \mathbb{I}\{\theta \neq a\}$ and the uniform prior w is known (see Gushchin (2003) and Guntuboyina (2011b)). In such a discrete setting, $w(B(a)) = 1/N$ for any $a \in \mathcal{A}$ and thus $R_0 = 1 - 1/N$. The proof of Theorem 2 heavily relies on the following lemma, which is a consequence of the data processing inequality for f-divergences (see (10) in Section 2).

Lemma 3 *Suppose that the loss function L is zero-one valued. For every $f \in \mathcal{C}$, every probability measure Q on $\mathcal{X}$ and every decision rule $\mathfrak{d}$, we have*

$$\int_\Theta D_f(P_\theta||Q)w(\mathrm{d}\theta) \geq \phi_f(R^{\mathfrak{d}}, R^{\mathfrak{d}}_Q) \tag{20}$$

where

$$R^{\mathfrak{d}} := \int_\Theta \mathbb{E}_\theta L(\theta, \mathfrak{d}(X))w(\mathrm{d}\theta) \quad , \quad R^{\mathfrak{d}}_Q := \int_{\mathcal{X}} \int_\Theta L(\theta, \mathfrak{d}(x))w(\mathrm{d}\theta)Q(\mathrm{d}x). \tag{21}$$

We note that Lemma 3 is of independent interest, which can be applied to establish minimax lower bound as shown in the following remark.

Proof [Proof of Lemma 3]

Let $\mathbb{P}$ denote the joint distribution of θ and X under the prior w i.e., $\theta \sim w$ and $X|\theta \sim P_\theta$. For any decision rule $\mathfrak{d}$, $R^{\mathfrak{d}}$ in (21) can be written as $R^{\mathfrak{d}} = \mathbb{E}_{\mathbb{P}} L(\theta, \mathfrak{d}(X))$. Let $\mathbb{Q}$ denote the joint distribution of θ and X under which they are independently distributed according to $\theta \sim w$ and $X \sim Q$ respectively. The quantity $R^{\mathfrak{d}}_Q$ in (21) can then be written as $R^{\mathfrak{d}}_Q = \mathbb{E}_{\mathbb{Q}} L\left(\theta, \mathfrak{d}(X)\right)$.

Because the loss function is zero-one valued, the function $\Gamma(\theta, x) := L(\theta, \mathfrak{d}(x))$ maps $\Theta \times \mathcal{X}$ into $\{0, 1\}$. Our strategy is to fix $f \in \mathcal{C}$ and apply the data processing inequality (10) to the probability measures $\mathbb{P}, \mathbb{Q}$ and the mapping Γ. This gives

$$D_f(\mathbb{P}||\mathbb{Q}) \geq D_f(\mathbb{P}\Gamma^{-1}||\mathbb{Q}\Gamma^{-1}), \tag{22}$$

where $\mathbb{P}\Gamma^{-1}$ and $\mathbb{Q}\Gamma^{-1}$ are induced measures on the space $\{0, 1\}$ of Γ. In other words, since L is zero-one valued, both $\mathbb{P}\Gamma^{-1}$ and $\mathbb{Q}\Gamma^{-1}$ are two-point distributions on $\{0, 1\}$ with

$$\mathbb{P}\Gamma^{-1}\{1\} = \int \Gamma d\mathbb{P} = \mathbb{E}_{\mathbb{P}} L(\theta, \mathfrak{d}(X)) = R^{\mathfrak{d}}, \qquad \mathbb{Q}\Gamma^{-1}\{1\} = \int \Gamma d\mathbb{Q} = R^{\mathfrak{d}}_Q.$$

By the definition of the function $\phi_f(\cdot, \cdot)$, it follows that $D_f(\mathbb{P}\Gamma^{-1}||\mathbb{Q}\Gamma^{-1}) = \phi_f(R^{\mathfrak{d}}, R^{\mathfrak{d}}_Q)$. It is also easy to see $D_f(\mathbb{P}||\mathbb{Q}) = \int_\Theta D_f(P_\theta||Q)w(\mathrm{d}\theta)$. Combining this equation with inequality (22) establishes inequality (20). ■

With Lemma 3 in place, we are ready to prove Theorem 2.

Proof [Proof of Theorem 2]

We write R as a shorthand notation of $R_{\mathrm{Bayes}}(w)$. By the definition (11) of $I_f(w, \mathcal{P})$, it suffices to prove that

$$\int D_f(P_\theta \| Q)w(d\theta) \geq \phi_f(R, R_0) \tag{23}$$

for every probability measure Q.

Notice that $R \leq R_0$. If $R = R_0$, then the right hand side of (17) is zero and hence the inequality immediately holds. Assume that $R < R_0$. Let $\epsilon > 0$ be small enough so that $R + \epsilon < R_0$. Let $\mathfrak{d}$ denote any decision rule for which $R \leq R^{\mathfrak{d}} < R + \epsilon$ and note that such a rule exists since $R = \inf_{\mathfrak{d}} R^{\mathfrak{d}}$. It is easy to see that

$$R_Q^{\mathfrak{d}} = \int_{\mathcal{X}} \int_{\Theta} L(\theta, \mathfrak{d}(x)) w(\mathrm{d}\theta) Q(dx) \geq \int_{\mathcal{X}} \left(\inf_{a \in \mathcal{A}} \int_{\Theta} L(\theta, a) w(\mathrm{d}\theta) \right) Q(dx) = R_0.$$

We thus have $R \leq R^{\mathfrak{d}} < R + \epsilon < R_0 \leq R_Q^{\mathfrak{d}}$. By Lemma 3, we have

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

Because $x \mapsto \phi_f(x, R_Q^{\mathfrak{d}})$ is non-increasing on $x \in [0, R_Q^{\mathfrak{d}}]$, we have

$$\phi_f(R^{\mathfrak{d}}, R_Q^{\mathfrak{d}}) \geq \phi_f(R + \epsilon, R_Q^{\mathfrak{d}}).$$

Because $x \mapsto \phi_f(R + \epsilon, x)$ is non-decreasing on $x \in [R + \epsilon, 1]$, we have

$$\phi_f(R + \epsilon, R_Q^{\mathfrak{d}}) \geq \phi_f(R + \epsilon, R_0).$$

Combining the above three inequalities, we have

$$\int_{\Theta} D_f(P_\theta \| Q) w(d\theta) \geq \phi_f(R^{\mathfrak{d}}, R_Q^{\mathfrak{d}}) \geq \phi_f(R + \epsilon, R_Q^{\mathfrak{d}}) \geq \phi_f(R + \epsilon, R_0).$$

The proof of (23) completes by letting $\epsilon \downarrow 0$ and using the continuity of $\phi_f(\cdot, R_0)$ (continuity was noted in Lemma 1). This completes the proof of Theorem 2. ■

Remark 4 *Lemma 3 can also be used to derive minimax lower bounds in a different way. For example, when the minimax decision rule $\mathfrak{d}$ exists (e.g., for finite space Θ and $\mathcal{A}$ (Ferguson, 1967)), we have $R^{\mathfrak{d}} \leq R_{\text{minimax}}$. If the probability measure Q is chosen so that $R_{\text{minimax}} \leq R_Q^{\mathfrak{d}}$, then, by Lemma 1, the right hand side of* (17) *can be lower bounded by replacing $R^{\mathfrak{d}}$ with R_{minimax} which yields*

$$\int_{\Theta} D_f(P_\theta || Q) w(\mathrm{d}\theta) \geq \phi_f(R_{\text{minimax}}, R_Q^{\mathfrak{d}}). \tag{24}$$

Similarly, this inequality can be converted to an explicit lower bound on minimax risk. We will show an application of this inequality in deriving Birgé-Gushchin inequality (Gushchin, 2003; Birgé, 2005) in Section 3.3.

3.1 Generalized Fano's Inequality

In the next result, we derive the generalized Fano's ienquality (5) using Theorem 2. The inequality proved here is in fact slightly stronger than (5); see Remark 6 for the clarification.

Corollary 5 (Generalized Fano's inequality) *For any given prior w and zero-one loss L, we have*

$$R_{\text{Bayes}}(w, L; \Theta) \geq 1 + \frac{I(w, \mathcal{P}) + \log(1 + R_0)}{\log\left(\sup_{a \in \mathcal{A}} w(B(a))\right)}, \tag{25}$$

where $B(a)$ is defined in (16).

Proof [Proof of Corollary 5]

We simply apply (19) to $f(x) = x \log x$ and $r = R_0/(1 + R_0)$, it can then be checked that

$$\phi_f(r, R_0) = -\log(1 + R_0) - \frac{1}{1 + R_0} \log(1 - R_0), \quad \phi_f'(r-, R_0) = \log(1 - R_0),$$

Inequality (19) then gives

$$R \geq 1 + \frac{I(w, \mathcal{P}) + \log(1 + R_0)}{\log(1 - R_0)}$$

which proves (25). ■

Remark 6 *This inequality is slightly stronger than* (5) *because $R_0 \leq 1$ (thus $\log(1 + R_0) \leq \log 2$). For example, when $\Theta = \mathcal{A} = \{0, 1\}$, $L(\theta, a) := \mathbb{I}\{\theta \neq a\}$ and $w\{0\} = w\{1\} = 1/2$, the inequality* (5) *leads to a trivial bound since the right hand side of* (5) *is negative. However, since $R_0 = 1/2$, the inequality* (25) *still provides a useful lower bound when $I(w, \mathcal{P})$ is strictly smaller than $\log 2 - \log(3/2)$.*

As mentioned in the introduction, the classical Fano inequality (3) and the recent continuum Fano inequality (4) are both special cases (restricted to uniform priors) of Corollary 5. The proof of (4) given in Duchi and Wainwright (2013) is rather complicated with a stronger assumption and a discretization-approximation argument. Our proof based on Theorem 2 is much simpler. Lemma 3 also has its independent interest. Using Lemma 3, we are able to recover another recently proposed variant of Fano's inequality in Braun and Pokutta (2014, Proposition 2.2). Details of this argument are provided in Appendix A.2.

3.2 Specialization of Theorem 2 to Different f-divergences and Their Applications

In addition to the generalized Fano's inequality, Theorem 2 allows us to derive a class of lower bounds on Bayes risk for zero-one losses by plugging other f-divergences. In the next corollary, we consider some widely used f-divergences and provide the corresponding Bayes risk lower bounds by inverting (17) in Theorem 2.

Corollary 7 *Let L be zero-one valued, w be any prior on Θ and $R = R_{\text{Bayes}}(w, L, \Theta)$. We then have the following inequalities*

(i) Chi-squared divergence:

$$R \geq R_0 - \sqrt{R_0(1 - R_0) I_{\chi^2}(w, \mathcal{P})}. \tag{26}$$

(ii) Total variation distance:

$$R \geq R_0 - I_{TV}(w, \mathcal{P}). \tag{27}$$

(iii) Hellinger distance:

$$R \geq R_0 - (2R_0 - 1)\frac{h^2}{2} - \sqrt{R_0(1-R_0)h^2(2-h^2)}. \tag{28}$$

provided $h^2 \leq 2R_0$. *Here* $h^2 = \int_\Theta \int_\Theta H^2(P_\theta \| P_{\theta'}) w(\mathrm{d}\theta) w(\mathrm{d}\theta')$.

See Appendix A.3 for the proof of the corollary. The special case of Corollary 7 for $\Theta = \mathcal{A} = \{1, \ldots, N\}$, $L(\theta, a) = \mathbb{I}\{\theta \neq a\}$ and w being the uniform prior has been discovered previously in Guntuboyina (2011b). It is clear from Corollary 7 that the choice of f-divergence will affect the tightness of the lower bound for R. In Appendix A.5, we provide a qualitative comparison of the lower bounds (25), (26) and (28). In particular, we show that in the discrete setting with $\Theta = \mathcal{A} = \{1, \ldots, N\}$, the lower bounds induced by the KL divergence and the chi-squared divergence are much stronger than the bounds given by the Hellinger distance. Therefore, in most applications in this paper, we shall only use the bounds involving the KL divergence and the chi-squared divergence.

Corollary 7 can be used to recover classical inequalities of Le Cam (for two point hypotheses) and Assouad (Theorem 2.12 in Tsybakov (2010) with both total variation distance and Hellinger distance) and Theorem 2.15 in Tsybakov (2010) that involves fuzzy hypotheses. The details are presented in Appendix A.4.

3.3 Birgé-Gushchin's Inequality

In this section, we expand (24) in Remark 4 to obtain a minimax risk lower bound due to Gushchin (2003) and Birgé (2005), which presents an improvement of the classical Fano's inequality when specializing to KL divergence.

Proposition 8 *(Gushchin, 2003; Birgé, 2005) Consider the finite parameter and action space* $\Theta = \mathcal{A} = \{\theta_0, \theta_1, \ldots, \theta_N\}$ *and the zero-one valued indicator loss* $L(\theta, a) = \mathbb{I}\{\theta \neq a\}$, *for any* f*-divergence,*

$$\phi_f(R_{\text{minimax}}, 1 - R_{\text{minimax}}/N) \leq \min_{0 \leq j \leq N} \frac{1}{N} \sum_{i: i \neq j} D_f\left(P_{\theta_i} || P_{\theta_j}\right). \tag{29}$$

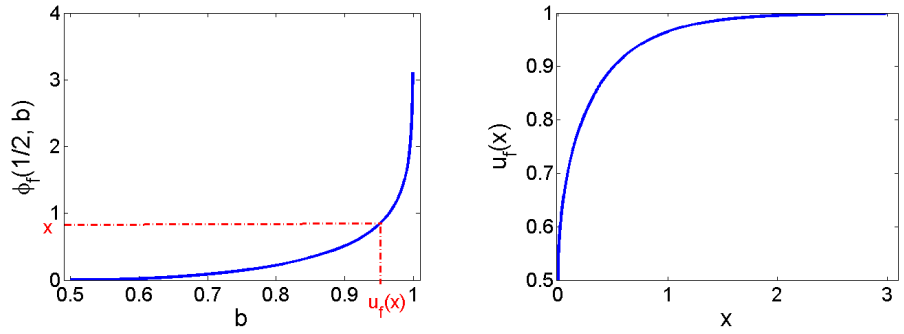
Proof [Proof of Proposition 8]

To prove Proposition 8, it is enough to prove that $\frac{1}{N}\sum_{i:i\neq j} D_f(P_{\theta_i}||P_{\theta_j}) \geq \phi_f(R_{\text{minimax}}, 1 - R_{\text{minimax}}/N)$ for every $j \in \{0, \ldots, N\}$. Without loss of generality, we assume that $j = 0$. We apply (20) 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 $\mathfrak{d}$. Because $\mathfrak{d}$ is the minimax rule, $R^{\mathfrak{d}} \leq R_{\text{minimax}}$. Also

$$R_Q^{\mathfrak{d}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\theta_0} L(\theta_i, \mathfrak{d}(X)) = \frac{1}{N}\mathbb{E}_{\theta_0} \sum_{i=1}^N \mathbb{I}\{\theta_i \neq \mathfrak{d}(X)\}.$$

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

$$R_Q^{\mathfrak{d}} \geq 1 - R_{\text{minimax}}/N. \tag{30}$$



(a) $\phi_f(1/2, b)$ (b) $u_f(x)$

Figure 2: Illustration of $\phi_f(1/2, b)$ and $u_f(x)$ for $f(x) = x \log x$.

On the other hand, we have $R_{\text{minimax}} \leq N/(N+1)$. To see this, note that the minimax risk is upper bounded by the maximum risk of a random decision rule, which chooses among the $N+1$ hypotheses uniformly at random. For this random decision rule, its risk is $\frac{N}{N+1}$ no matter what the true hypothesis is. Thus, $\frac{N}{N+1}$ is an upper bound on the minimax risk. We thus have, from (30), that $R_Q^{\mathfrak{d}} \geq 1 - R_{\text{minimax}}/N \geq R_{\text{minimax}}$. We can thus apply (24) to obtain

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

which completes the proof Proposition 8. ■

4. Bayes Risk Lower Bounds for Nonnegative Loss Functions

In the previous section, we discussed Bayes risk lower bounds for zero-one valued loss functions. We deal with general nonnegative loss functions in this section. The main result of this section, Theorem 9, provides lower bounds for $R_{\text{Bayes}}(w, L; \Theta)$ for any given loss L and prior w. To state this result, we need the following notion. Fix $f \in \mathcal{C}$ and recall the definition of ϕ_f in (13). We define $u_f : [0, \infty) \mapsto [1/2, 1]$ by

$$u_f(x) := \inf \{1/2 \leq b \leq 1 : \phi_f(1/2, b) > x\} \tag{31}$$

and if $\phi_f(1/2, b) \leq x$ for every $b \in [1/2, 1]$, then we take $u_f(x)$ to be 1. By Lemma 1, it is easy to see that $u_f(x)$ is a non-decreasing function of x. For example, for KL-divergence with $f(x) = x \log x$, we have $\phi_f(1/2, b) = \frac{1}{2} \log \frac{1}{4b(1-b)}$ and $u_f(x) = \frac{1}{2} + \frac{1}{2}\sqrt{1 - e^{-2x}}$ (see Figure 2). We are now ready to state the main theorem of this paper.

Theorem 9 *For every $\Theta, \mathcal{A}, L, w$ and $f \in \mathcal{C}$, we have*

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w(B_t(a, L)) < 1 - u_f(I_f(w, \mathcal{P})) \right\}, \tag{32}$$

where

$$B_t(a, L) := \{\theta \in \Theta : L(\theta, a) < t\} \quad \text{for } a \in \mathcal{A} \text{ and } t > 0. \tag{33}$$

Proof [Proof of Theorem 9]

Fix $\Theta, \mathcal{A}, L, w$ and f. Let $I := I_f(w, \mathcal{P})$ be a shorthand notation. Suppose $t > 0$ is such that

$$\sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < 1 - u_f(I). \tag{34}$$

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

Because L_t is a zero-one valued loss function, Theorem 2 gives

$$I \geq \phi_f(R, R_0) \quad \text{where } R_0 = 1 - \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right). \tag{35}$$

By (34), it then follows that $R_0 > u_f(I)$. By definition of $u_f(\cdot)$, it is clear that there exists $b^* \in [1/2, R_0)$ such that $\phi(1/2, b^*) > I$ (this in particular implies that $R_0 \geq 1/2$). Lemma 1 implies that $b \mapsto \phi_f(1/2, b)$ is non-decreasing for $b \in [1/2, 1]$, which yields $\phi_f(1/2, b^*) \leq \phi_f(1/2, R_0)$. The above two inequalities imply $I < \phi_f(1/2, R_0)$. Combining this inequality with (35), we have

$$\phi_f(1/2, R_0) > I \geq \phi_f(R, R_0).$$

Lemma 1 shows that $a \mapsto \phi_f(a, R_0)$ is non-increasing for $a \in [0, R_0]$. Thus, we have $R \geq 1/2$. ■

We further note that because $u_f(x)$ is non-decreasing in x, one can replace $I_f(w, \mathcal{P})$ in (32) by any upper bound I_f^{up} i.e., for any $I_f^{\text{up}} \geq I_f(w, \mathcal{P})$, we have

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup\left\{t > 0 : \sup_{a \in \mathcal{A}} w(B_t(a, L)) < 1 - u_f(I_f^{\text{up}})\right\}. \tag{36}$$

This is useful since $I_f(w, \mathcal{P})$ is often difficult to calculate exactly. When $f(x) = x \log x$, Haussler and Opper (1997) provided a useful upper bound on the mutual information $I(w, \mathcal{P})$. We describe this result in Section 5 where we also extend it to power divergences f_α for $\alpha \notin [0, 1]$ (which covers the case of chi-squared divergence).

Remark 10 *From the proof of Theorem 9, it can be observed that the constant* $1/2$ *in the right hand side of* (32) *and in the definition of* $u_f(\cdot)$ *can be replaced by any* $c \in (0, 1]$. *This gives the sharper lower bound:*

$$R_{\text{Bayes}}(w, L; \Theta) \geq \sup_{c \in (0,1]} \left(c \sup\left\{t > 0 : \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < 1 - u_{f,c}(I_f(w, \mathcal{P}))\right\}\right),$$

where $u_{f,c}(x) = \inf\{c \leq b \leq 1 : \phi_f(c, b) \geq x\}$. *Since obtaining exact constants is not our main concern, the inequality* (32) *is usually sufficient to provide Bayes risk lower bounds with correct dependence on the model and prior.*

Remark 11 *We note that the lower bound presented in Theorem 9 might not be tight for some special priors, e.g., when the prior w has extremely large density in some small region of the parameter space. We call such regions with unbounded density as* spikes *in the prior distribution. As a concrete example, let $\Theta = \mathcal{A}$ be a subset of a finite dimensional Euclidean space containing the origin with L being the Euclidean distance and let w denote the mixture of the uniform priors over the balls $B_1(0, L)$ and $B_\epsilon(0, L)$ for some very small $0 < \epsilon \ll 1$. In this case, the mixture component $B_\epsilon(0, L)$ is a spike. If ϵ is very small, then the term $\sup_{a\in\mathcal{A}} w(B_t(a, L))$ might be too big for Theorem 9 to establish a tight lower bound.*

Even in such extreme cases, the tight lower bound can be salvaged by partitioning the parameter space Θ into finite or countably many disjoint subsets $\Theta_i, i \geq 0$ and to apply Theorem 9 to w restricted to each Θ_i. To illustrate this technique, suppose that w has a Lebesgue density φ that is bounded from above. Let $\varphi_{\max}$ denote the supremum of φ. We partition the parameter space Θ into disjoint subsets $\Theta_0, \Theta_1, \ldots$ with

$$\Theta_i := \{\theta \in \Theta : 2^{-(i+1)}\varphi_{\max} < \varphi(\theta) \leq 2^{-i}\varphi_{\max}\}. \tag{37}$$

Then, we apply Theorem 9 to w restricted to each Θ_i. More specifically, let w_i denote the probability measure w restricted to Θ_i i.e., $w_i(S) := w(S \cap \Theta_i)/w(\Theta_i)$ for any measurable set $S \subseteq \Theta_i$. we have

$$R_{\text{Bayes}}(w, L; \Theta) \geq \sum_i w(\Theta_i) R_{\text{Bayes}}(w_i, L; \Theta_i), \tag{38}$$

where $R_{\text{Bayes}}(w_i, L; \Theta_i) = \inf_{\mathfrak{d}} \int_{\Theta_i} \mathbb{E}_\theta L(\theta, \mathfrak{d}(X)) w_i(\mathrm{d}\theta)$. To see this, for any decision rule $\mathfrak{d}$, we have $R^{\mathfrak{d}}(w, L; \Theta) = \sum_{i=1}^{\infty} w(\Theta_i) R^{\mathfrak{d}}(w_i, L; \Theta_i)$; then take infimum over all possible $\mathfrak{d}$ on both sides,

$$\begin{aligned} R_{\text{Bayes}}(w, L; \Theta) &= \inf_{\mathfrak{d}} R^{\mathfrak{d}}(w, L; \Theta) \\ &\geq \sum_{i=1}^{\infty} w(\Theta_i) \inf_{\mathfrak{d}} R^{\mathfrak{d}}(w_i, L; \Theta_i) = \sum_{i=1}^{\infty} w(\Theta_i) R_{\text{Bayes}}(w_i, L; \Theta_i) \end{aligned}$$

One can lower bound each Bayes risk $R_{\text{Bayes}}(w_i, L; \Theta_i)$ for all i using Theorem 9. Since the density of w_i differs by a factor at most 2, the spiking prior problem will no longer exist while applying Theorem 9 for w_i. We also note that another useful application of such a partitioning technique is presented in Corollary 17.

Now take the concrete example of the mixture of the uniform priors over $B_1 := B_1(0, L)$ and $B_\epsilon := B_\epsilon(0, L)$. It is clear from (37) that $\Theta_0 = B_\epsilon$ and $\Theta_k = B_1 \backslash B_\epsilon$ for some $k > 0$ and the rest of Θ_i's are empty sets. Applying (38), we have

$$\begin{aligned} R_{\text{Bayes}}(w, L; \Theta) \geq& w(B_\epsilon) R_{\text{Bayes}}(w_1, L; B_\epsilon) + w(B_1 \backslash B_\epsilon) R_{\text{Bayes}}(w_2, L; B_1 \backslash B_\epsilon) \\ \geq& w(B_1 \backslash B_\epsilon) R_{\text{Bayes}}(w_2, L; B_1 \backslash B_\epsilon) \end{aligned}$$

Note that $w(B_1 \backslash B_\epsilon)$ is lower bounded by a universal constant. Then we can lower bound $R_{\text{Bayes}}(w_2, L; B_1 \backslash B_\epsilon)$ using Theorem 9 and obtain a tight lower bound up to a constant factor that is independent of ϵ (see an example of deriving Bayes risk lower bound for estimating the mean of a Gaussian model with uniform prior on a ball in Section 5).

For specific $f \in \mathcal{C}$, the right hand side of (36) can be explicitly evaluated as shown in the next corollary.

Corollary 12 *Fix $\Theta, \mathcal{A}, L, w$ and $\mathcal{P}$. The Bayes risk $R_{\text{Bayes}}(w, L; \Theta)$ satisfies each of the following inequalities (the quantity I_f^{up} represents an upper bound on the corresponding f-informativity):*

(i) KL divergence:

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < \frac{1}{4} e^{-2I_f^{\text{up}}} \right\}. \tag{39}$$

(ii) Chi-squared divergence:

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < \frac{1}{4\left(1 + I_f^{\text{up}}\right)} \right\}. \tag{40}$$

(iii) Total variation distance:

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < \frac{1}{2} - I_f^{\text{up}} \right\}. \tag{41}$$

(iv) Hellinger distance: If $I_f^{\text{up}} < 1 - 1/\sqrt{2}$, then we have

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < \frac{1}{2} - \left(1 - I_f^{\text{up}}\right) \sqrt{I_f^{\text{up}}\left(2 - I_f^{\text{up}}\right)} \right\}. \tag{42}$$

Proof [Proof of Corollary 12]

Inequality (39) involving 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 (31) has the expression for all $x > 0$,

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

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 (32) reduces to the desired inequality (39):

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2} \sup \left\{ t > 0 : \sup_{a \in \mathcal{A}} w\left(B_t(a, L)\right) < \frac{1}{4} e^{-2I_f^{\text{up}}} \right\}.$$

The proof of the Bayes risk lower bounds for the other three f-divergences are similar and thus we only present the form of $u_f(x)$. Inequality (40) involves chi-squared divergence with $f(x) = x^2 - 1$. Therefore, we have for all $x > 0$,

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

Inequality (41) involves total variation distance with $f(x) = |x-1|/2$. Then

$$u_f(x) = \inf\{1/2 \le b \le 1 : |1-2b| > 2x\} = \frac{1}{2} + x.$$

Inequality (42) involves Hellinger divergence with $f(x) = 1 - \sqrt{x}$ and thus

$$\begin{aligned} u_f(x) &= \inf\left\{1/2 \le b \le 1 : 1 - \sqrt{b/2} - \sqrt{(1-b)/2} > x\right\} \\ &= \begin{cases} 1 & \text{if } x \ge 1 - 1/\sqrt{2} \\ \frac{1}{2} + (1-x)\sqrt{x(2-x)} & \text{if } x < 1 - 1/\sqrt{2}. \end{cases} \end{aligned}$$

■

Remark 13 *A special case of Corollary 12(i) appeared as Zhang (2006, Theorem 6.1). To see that Zhang (2006, Theorem 6.1) is indeed a special case of* (39)*, note first that* (39) *is equivalent to*

$$R_{Bayes}(w, L; \Theta) \ge \frac{1}{2} \sup\left\{t > 0 : \inf_{a \in \mathcal{A}} \frac{1}{w(B_t(a, L))} > 2I^{up} + \log 4\right\}. \tag{43}$$

Here I^{up} is any upper bound on the mutual information. One such upper bound on the mutual information is

$$I^{up} = \int_\Theta \int_\Theta D(P_\theta \| P_\xi) w(d\xi) w(d\theta) \tag{44}$$

That I^{up} is an upper bound on the mutual information can be seen for example by using concavity of the logarithm (46) *when the family $\{Q_\xi, \xi \in \Xi\}$ is chosen to be the same as $\{P_\theta, \theta \in \Theta\}$. Using* (44) *in* (43)*, we obtain*

$$R_{Bayes}(w, L; \Theta) \ge \frac{1}{2} \sup\left\{t > 0 : \inf_{a \in \mathcal{A}} \frac{1}{w(B_t(a, L))} > 2\int_\Theta \int_\Theta D(P_\theta \| P_\xi) w(d\xi) w(d\theta) + \log 4\right\}.$$

If we now specialize to the setting when the probability measures $\{P_\theta, \theta \in \Theta\}$ are all n-fold product measures i.e., when each P_θ is of the form $\mathfrak{P}_\theta^n$ for some class of probabilities $\{\mathfrak{P}_\theta, \theta \in \Theta\}$, then the inequality becomes

$$R_{Bayes}(w, L; \Theta) \ge \frac{1}{2} \sup\left\{t > 0 : \inf_{a \in \mathcal{A}} \frac{1}{w(B_t(a, L))} > 2n\int_\Theta \int_\Theta D(\mathfrak{P}_\theta \| \mathfrak{P}_\xi) w(d\xi) w(d\theta) + \log 4\right\}.$$

This inequality is precisely Zhang (2006, Theorem 6.1).

5. Upper Bounds on f-informativity and Examples

Application of Theorem 9 requires upper bounds on the f-informativity $I_f(w; \mathcal{P})$. This is the subject of this section. We focus on the power divergence f_α for $\alpha \ge 1$ which includes the KL divergence and chi-squared divergence as special cases. Recall that in the comment/paragraph below Corollary 7 (see also Section A.5 in the appendix), we provided

motivation for restricting our attention to such divergences as opposed to e.g., Hellinger distance.

We assume that there is a measure μ on $\mathcal{X}$ that dominates P_θ for every $\theta \in \Theta$. None of our results depend on the choice of the dominating measure μ.

When the f-informativity is the mutual information, Haussler and Opper (1997) have proved useful upper bounds which we briefly review here. Let P and $\{Q_\xi, \xi \in \Xi\}$ be probability measures on $\mathcal{X}$ having densities p and $\{q_\xi, \xi \in \Xi\}$ respectively with respect to μ. Let ν be an arbitrary probability measure on Ξ and $\bar{Q}$ be the probability measure on $\mathcal{X}$ having density $\bar{q} = \int_\Xi q_\xi \nu(\mathrm{d}\xi)$ with respect to μ. Haussler and Opper (1997) proved the following inequality

$$D\left(P||\bar{Q}\right) \leq -\log\left(\int_\Xi \exp\left(-D(P||Q_\xi)\right)\nu(\mathrm{d}\xi)\right). \tag{45}$$

Now given a class of probability measures $\{P_\theta, \theta \in \Theta\}$, applying the above inequality for each P_θ and integrating the resulting inequalities with respect to a probability measure w on Θ, Haussler and Opper (1997, Theorem 2) obtained the following mutual information upper bound:

$$I(w, \mathcal{P}) \leq -\int_\Theta \log\left(\int_\Xi \exp\left(-D(P_\theta||Q_\xi)\right)\nu(\mathrm{d}\xi)\right) w(\mathrm{d}\theta). \tag{46}$$

In the special case when $\Xi = \{1, \ldots, M\}$ and ν is the uniform probability measure on Ξ, we have $\bar{Q} = \left(Q_1 + \ldots + Q_M\right)/M$ and inequality (45) then becomes $D(P||\bar{Q}) \leq -\log\left(\frac{1}{M}\sum_{j=1}^M \exp\left(-D(P||Q_j)\right)\right)$. Because $\sum_{j=1}^M \exp(-D(P\|Q_j)) \geq \exp\left(-\min_j D(P\|Q_j)\right)$, we obtain

$$D(P\|\bar{Q}) \leq \log M + \min_{1\leq j\leq M} D(P\|Q_j).$$

Inequality (46) can be further simplified to

$$I(w, \mathcal{P}) \leq \log M + \int_\Theta \min_{1\leq j\leq M} D(P_\theta||Q_j) w(\mathrm{d}\theta). \tag{47}$$

This inequality can be used to give an upper bound for f-informativity in terms of the KL covering numbers. Recall the definition of $M_{KL}(\epsilon, \Theta)$ from (12). Applying (47) to any fixed $\epsilon > 0$ and choosing $\{Q_1, \ldots, Q_M\}$ to be an ϵ^2-covering, we have

$$I(w, \mathcal{P}) \leq \inf_{\epsilon>0}\left(\log M_{KL}(\epsilon, \Theta) + \epsilon^2\right). \tag{48}$$

When w is the uniform prior on a finite subset of Θ, the above inequality has been proved by Yang and Barron (1999, Page 1571). If $M_{KL}(\epsilon, \Theta)$ is infinity for all ϵ, then (48) gives ∞ as the upper bound on $I(w, \mathcal{P})$ and thus (39) will lead to a trivial lower bound 0 for R_{Bayes}. In such a case, one may find a subset $\tilde{\Theta} \subset \Theta$ for which $M_{KL}(\epsilon, \tilde{\Theta})$ is bounded and contains most prior mass. If $\tilde{w}$ denotes the prior w restricted in $\tilde{\Theta}$, then it is easy to see that $R_{\text{Bayes}}(w, L; \Theta) \geq w(\tilde{\Theta}) R_{\text{Bayes}}(\tilde{w}, L; \tilde{\Theta})$. Then we can use (39) and (48) to lower bound $R_{\text{Bayes}}(\tilde{w}, L; \tilde{\Theta})$.

In the next theorem, we extend inequalities (45) and (46) to power divergences corresponding to f_α for $\alpha \notin [0, 1]$. We also note that in Appendix B.2, we demonstrate the tightness of the bound (49) in Theorem 14 by a simple example.

Theorem 14 *Fix $\alpha \notin [0,1]$ and let $f_\alpha \in \mathcal{C}$ be as defined in Section 2. Under the setting of inequalities* (45) *and* (46)*, we have*

$$D_{f_\alpha}(P||\bar{Q}) \leq \left[\int_\Xi \left(D_{f_\alpha}(P||Q_\xi)+1\right)^{1/(1-\alpha)} \nu(\mathrm{d}\xi)\right]^{1-\alpha} - 1. \tag{49}$$

and

$$I_{f_\alpha}(w,\mathcal{P}) \leq \int_\Theta \left[\int_\Xi \left(D_{f_\alpha}(P_\theta||Q_\xi)+1\right)^{1/(1-\alpha)} \nu(\mathrm{d}\xi)\right]^{1-\alpha} w(\mathrm{d}\theta) - 1. \tag{50}$$

To prove Theorem 14, the following lemma is critical (the proof of this lemma in given in Appendix B.1).

Lemma 15 *Fix $r < 1$. Let μ be a probability measure on the space T and let $S := \{u : T \to \mathbb{R}_+ : u \in L^r_\mu(T)\}$. Then the map $f : S \to \mathbb{R}$ defined by $f(u) := \left(\int_T u(t)^r \mu(\mathrm{d}t)\right)^{1/r}$ is concave in u.*

Note that the discrete version of Lemma 15 states that $f(u) = \left(\sum_{i=1}^M u_i^r/M\right)^{1/r}$ is a concave function of $u \in \mathbb{R}_+^M$ when $r < 1$.

In fact, since we will apply this lemma to prove Theorem 14 with $r = \frac{1}{1-\alpha}$, the condition $r < 1$ in Lemma 15 translates into $\alpha \notin [0,1]$ in Theorem 14. We are now ready to prove Theorem 14.

Proof [Proof of Theorem 14]

By the identity that $D_{f_\alpha}(P||Q) = D_{f_{1-\alpha}}(Q||P)$, we have

$$\begin{aligned} D_{f_\alpha}(P||\bar{Q}) = D_{f_{1-\alpha}}(\bar{Q}||P) &= \int_{\mathcal{X}} p \left(\int_\Xi \frac{q_\xi}{p} \nu(\mathrm{d}\xi)\mathrm{d}\mu\right)^{1-\alpha} - 1 \\ &= \int_{\mathcal{X}} p \left(\int_\Xi \left[\left(\frac{q_\xi}{p}\right)^{1-\alpha}\right]^{1/(1-\alpha)} \nu(\mathrm{d}\xi)\mathrm{d}\mu\right)^{1-\alpha} - 1 \end{aligned}$$

Let $u(\xi,x) = \left(\frac{q_\xi}{p}\right)^{1-\alpha}$. Since $\frac{1}{1-\alpha} < 1$ when $\alpha \notin [0,1]$, Lemma 15 implies that $u(\xi,x) \mapsto \left(\int_\Xi u(\xi,x)^{1/(1-\alpha)} \nu(\mathrm{d}\xi)\right)^{1-\alpha}$ is concave in u. Applying Jensen's inequality,

$$\begin{aligned} D_{f_\alpha}(P||\bar{Q}) &\leq \left(\int_\Xi \left[\int_{\mathcal{X}} p \left(\frac{q_\xi}{p}\right)^{1-\alpha} \mathrm{d}\mu\right]^{1/(1-\alpha)} \nu(\mathrm{d}\xi)\right)^{1-\alpha} - 1 \\ &= \left(\int_\Xi \left[D_{f_{1-\alpha}}(Q_\xi||P)\right]^{1/(1-\alpha)} \nu(\mathrm{d}\xi)\right)^{1-\alpha} - 1. \end{aligned}$$

This completes the proof of (49) because $D_{f_{1-\alpha}}(Q_\xi||P) = D_{f_\alpha}(P||Q_\xi)$. The proof of (50) follows by applying (49) for $P = P_\theta$ and then integrating the resulting bound with respect to $w(\mathrm{d}\theta)$. ■

For $\alpha > 1$, one can deduce an upper bound analogous to (48) for the f_α-informativity which is described in the next corollary. Recall the notion of the covering numbers $M_\alpha(\epsilon,\Theta)$ from Section 2.

Corollary 16 *For every $\alpha > 1$, we have*

$$I_{f_\alpha}(w, \mathcal{P}) \leq \inf_{\epsilon > 0} (1 + \epsilon^2) M_\alpha(\epsilon, \Theta)^{\alpha - 1} - 1. \tag{51}$$

In particular, when D_{f_α} is the chi-square divergence, Corollary 16 implies

$$I_{\chi^2}(w, \mathcal{P}) \leq \inf_{\epsilon > 0} (1 + \epsilon^2) M_{\chi^2}(\epsilon, \Theta) - 1. \tag{52}$$

Note that Corollary 16 gives trivial bound when $M_\alpha(\epsilon, \Theta)$ equals ∞ for all $\epsilon > 0$. This can be handled in a way similar to that outlined in the discussion after (48).

Proof [Proof of Corollary 16]

Let $Q_1, \ldots, Q_M$ be probability measures on $\mathcal{X}$ and fix $\theta \in \Theta$. Inequality (49) applied to $P = P_\theta$, $\Xi := \{1, \ldots, M\}$ and the uniform probability measure on Ξ as ν gives

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

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

$$\begin{aligned} \sum_{j=1}^{M} (1 + D_{f_\alpha}(P_\theta \| Q_j))^{1/(1-\alpha)} \geq & \quad \max_{1 \leq j \leq M} (1 + D_{f_\alpha}(P_\theta \| Q_j))^{1/(1-\alpha)} \\ = & \quad (1 + \min_{1 \leq j \leq M} D_{f_\alpha}(P_\theta \| Q_j))^{1/(1-\alpha)}. \end{aligned}$$

This gives

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

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

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

The proof is complete by integrating the above inequality with respect to $w(\mathrm{d}\theta)$.

■

We now turn to applications of the Bayes risk lower bounds in Corollary 12 and the informativity upper bounds in this section. We present a toy example here and postpone more complicated examples (e.g., generalized linear model, spiked covariance model, Gaussian model with general prior and loss) to Appendix C.

Example 1 (Gaussian model with uniform priors on large balls) *Fix $d \geq 1$. Suppose $\Theta = \mathcal{A} \subseteq \mathbb{R}^d$ and let $L(\theta, a) := \|\theta - a\|_2^2$. For each $\theta \in \mathbb{R}^d$, let P_θ denote the Gaussian distribution with mean θ and covariance matrix $\sigma^2 I_{d \times d}$ ($\sigma^2 > 0$ is a constant). Let w be the uniform distribution on the closed ball of radius Γ centered at the origin. Let $\Gamma \geq \sigma\sqrt{d}$.*

We will show below how to obtain the tight Bayes risk lower bound using Corollary 12 along with the f-informativity upper bound in Corollary 16.

We can assume that Θ (and $\mathcal{A}$) is the closed ball of radius Γ centered at the origin as w puts zero probability outside this ball. We use the inequality (40) *induced by the chi-squared divergence. To establish the lower bound, we need to upper bound $\sup_{a\in\mathcal{A}} w(B_t(a,L))$ and the chi-squared informativity. The former can be easily controlled because $\sup_{a\in\mathcal{A}} w(B_t(a,L)) \le \left(\sqrt{t}/\Gamma\right)^d$. For the latter, we use* (52)*, which requires an upper bound on $M_{\chi^2}(\epsilon,\Theta)$. Note that $\chi^2(P_\theta\|P_{\theta'}) = \exp\left(\|\theta-\theta'\|_2/\sigma^2\right)-1$ for $\theta,\theta'\in\Theta$. As a consequence, $\chi^2(P_\theta\|P_{\theta'}) \le \epsilon^2$ if and only if $\|\theta-\theta\|_2 \le \epsilon' := \sigma\sqrt{\log(1+\epsilon^2)}$. Therefore, by a standard volumetric argument, we have*

$$M_{\chi^2}(\epsilon,\Theta) \le \left(\frac{\Gamma+\epsilon'/2}{\epsilon'/2}\right)^d \le \left(\frac{3\Gamma}{\epsilon'}\right)^d = \left(\frac{3\Gamma}{\sigma\sqrt{\log(1+\epsilon^2)}}\right)^d$$

provided $\epsilon' \le \Gamma$. In particular, if we take $\epsilon := \sqrt{e^d-1}$, then $\epsilon' = \sigma\sqrt{d} \le \Gamma$, we will obtain $M_{\chi^2}(\epsilon,\Theta) \le (3\Gamma/(\sigma\sqrt{d}))^d$. Inequality (52) *then gives $I_{\chi^2}(w,\mathcal{P}) \le \left(\frac{3e\Gamma}{\sigma\sqrt{d}}\right)^d - 1$. Let I_f^{up} be the right hand side. If we choose $t = cd\sigma^2$ for a sufficiently small constant $c>0$, then we have $\sup_{a\in\mathcal{A}} w(B_t(a,L)) < \frac{1}{4}(1+I_f^{\mathrm{up}})^{-1}$. Inequality* (40) *then gives*

$$R_{\mathrm{Bayes}}(w,L;\Theta) \ge cd\sigma^2. \tag{53}$$

This lower bound is tight due to the trivial upper bound $R_{\mathrm{Bayes}}(w,L;\Theta) \le d\min(\sigma^2,\Gamma^2)$ since $R_{\mathrm{Bayes}}(w,L;\Theta)$ is smaller than the risk of the constant estimator 0 as well as the trivial estimator of the observation itself.

This example allows us to compare the bound given by Theorem 9 for different $f\in\mathcal{C}$. We argue below that using KL divergence and applying (39) *along with inequality* (48) *for controlling the mutual information will not yield a tight lower bound for this example. In other words, the same strategy that works for $f(x)=x^2-1$ does not work for $f(x)=x\log x$. To see this, notice that $D(P_\theta\|P_{\theta'}) = \|\theta-\theta'\|^2/\sigma^2$ for $\theta,\theta'\in\Theta$. As a result, $D(P_\theta\|P_{\theta'}) \le \epsilon^2$ if and only if $\|\theta-\theta'\| \le \sqrt{2}\epsilon\sigma$. The same volumetric argument again gives $M_{KL}(\epsilon,\Theta) \le \left(\frac{3\Gamma}{\sqrt{2}\epsilon\sigma}\right)^d$ provided $\sqrt{2}\epsilon\sigma \le \Gamma$. The bound* (48) *implies that the mutual information $I(w,\mathcal{P})$ is bounded by*

$$I(w,\mathcal{P}) \le \inf_{0<\epsilon\le\Gamma/(\sqrt{2}\epsilon\sigma)} \left(d\log\left(\frac{3\Gamma}{\sqrt{2}\epsilon\sigma}\right)+\epsilon^2\right) = d\log\left(\frac{3\Gamma}{\sigma\sqrt{d}}\right)+\frac{d}{2}.$$

Let I_f^{up} be the right hand side above. The maximum $t>0$ for which $(\sqrt{t}/\Gamma)^d < \frac{1}{4}\exp\left(-2I_f^{\mathrm{up}}\right)$ is on the order of $d^2\sigma^4/\Gamma^2$. This means that inequality (39) *implies a weaker lower bound $\Omega(d^2\sigma^4/\Gamma^2)$, which is suboptimal when $d\sigma^2$ is small or when Γ is large. This is in contrast with the optimal bound* (53)*.*

In the above example, a direct application of Theorem 9 with $f(x)=x\log x$ does not produce a tight lower bound. This is mainly because, when the prior is over a large parameter space (e.g., a ball of a constant radius), the upper bound of mutual information over the entire parameter space Θ in (48) could be too loose. This can be corrected by partitioning the

parameter space Θ into small hypercubes, and applying our bounds for the prior restricted to each hypercube separately so that the mutual information inside the partition can be appropriately upper bounded using (48). This is another illustration of the idea described in Remark 11. We first describe this method in a more general setting in the following corollary and then apply it to the setting of Example 1. We use the following notation. For measurable subsets S of a Euclidean space, $\mathrm{Vol}(S)$ denotes the volume (Lebesgue measure) of S.

Corollary 17 *Let $\Theta = \mathcal{A} \subseteq \mathbb{R}^d$. Suppose that the prior w has a Lebesgue density f_w that is positive over Θ. For each $\theta \in \Theta$ and $\delta > 0$, let*

$$r_\delta(\theta) := \sup\left\{\frac{f_w(\theta_1)}{f_w(\theta_2)} : \theta_i \in \Theta \text{ and } \|\theta_i - \theta\|_2 \le \sqrt{d}\delta \text{ for } i = 1, 2\right\}.$$

Suppose also the existence of $A > 0$ such that $D(P_{\theta_1}\|P_{\theta_2}) \le A\|\theta_1 - \theta_2\|_2^2$ for all $\theta_1, \theta_2 \in \Theta$ and the existence of $V > 0$ (which may depend on d) and $p > 0$ such that $\sup_{a\in\mathcal{A}} \mathrm{Vol}(B_t(a, L)) \le Vt^{d/p}$ for every $t > 0$. Then

$$R_{\mathrm{Bayes}}(w, L; \Theta) \ge \frac{1}{2} \sup_{0<\delta\le A^{-1/2}} \left[e^{-2p}\delta^p (8V)^{-p/d} \int_\Theta \left(\frac{1}{r_\delta(\theta)}\right)^{p/d} w(\mathrm{d}\theta)\right]. \tag{54}$$

The proof of Corollary is quite technically involved and thus is deferred to Appendix B.3.

We demonstrate below that this corollary yields the correct rate in Example 1. More examples (e.g., estimation problem in generalized linear model, spiked covariance model, and Gaussian model with a general loss) are given in Appendix C.

Example 2 (Gaussian model with uniform priors on large balls (continued)) *Consider the same setting as in Example 1. Because $D(P_\theta\|P_{\theta'}) = \|\theta - \theta'\|_2^2/(2\sigma^2)$, we can take $A = (2\sigma^2)^{-1}$ in Corollary 17. Moreover, because $L(\theta, a) = \|\theta - a\|_2^2$, it is easy to see that $\sup_{a\in\mathcal{A}} \mathrm{Vol}(B_t(a, L)) \le t^{d/2}\mathrm{Vol}(B)$ which means that we can take $p = 2$ and $V = \mathrm{Vol}(B)$ in Corollary 17 where B is the unit ball in $\mathbb{R}^d$. Finally, because w is the uniform prior, we have $r_\delta(\theta) = 1$ for all $\theta \in \Theta$. Corollary 17 therefore gives*

$$R_{\mathrm{Bayes}}(w, L; \Theta) \ge \frac{1}{2} \sup_{0<\delta\le\sqrt{2}\sigma} \left(e^{-4}8^{-2/d}\delta^2 \mathrm{Vol}(B)^{-2/d}\right).$$

This matches the tight lower bound (53) *by noting that $\mathrm{Vol}(B)^{1/d} \asymp d^{-1/2}$.*

6. Smoothed Analysis for Spherical Gaussian Mixture Models with Uniform Weights

Smoothed analysis is a useful technique for analyzing algorithms that fail in the worst case but succeed with high probability in the average case. For parameter estimation problems, smoothed analysis assumes that the parameter to be estimated is randomly perturbed

by a small noise, and the data is generated with respect to the perturbed parameter as well. Under this setting, if the set of "bad" parameters that fail the estimator has zero measure, then the estimator will succeed almost surely after the perturbation. Smoothed analysis has been successfully applied to analyze linear programming (Blum and Dunagan, 2002; Dunagan et al., 2011; Hsu and Kakade, 2013; Spielman and Teng, 2003), integer programming (Röglin and Vöcking, 2007), binary search trees (Manthey and Reischuk, 2007), and other combinatorial problems (Banderier et al., 2003). See the paper by Spielman and Teng (2003) for a survey of existing works.

In this section, we use smoothed analysis to study an important problem in statistical estimation: learning mixture of spherical Gaussians. The problem of computing the maximum log-likelihood estimator is NP-hard (Arora and Kannan, 2005). However, if the true parameters are perturbed by a random noise, then we demonstrate that a variant of the polynomial-time algorithm proposed by Hsu and Kakade (2013) succeeds in estimating the Gaussian means. We present an upper bound on the algorithm's mean-squared error using smoothed analysis, which achieves a better rate than the original algorithm of Hsu and Kakade (2013). Furthermore, we apply the Bayes risk lower bound developed in this paper to show that, the mean squared-error achieved by this algorithm is unimprovable, even under smoothed analysis. To the best of our knowledge, the lower bound cannot be established by traditional information-theoretic techniques for lower bounding minimax risks.

6.1 Learning Mixture of Gaussians

We study estimating the parameter of a Gaussian mixture model (GMM). The parameter of a GMM is a d-by-k matrix $\theta := (\theta_1, \ldots, \theta_k)$. Each $\theta_i \in \mathbb{R}^d$ represents the mean of the i-th mixture component. We assume that the number of components k is much less than the dimensionality d. Suppose that n i.i.d. instances $\{x_i\}_{i=1}^n$ are sampled from the GMM with each $x_i \in \mathbb{R}^d$. Equivalently, it is generated by the following procedure: First, an integer z_i is uniformly sampled from $\{1, \ldots, k\}$. This integer is called the *membership* of the i-th instance[1]. Then, the vector x_i is drawn from the spherical Gaussian distribution $N(\theta_{z_i}; I_{d\times d})$. The goal is to estimate the parameters θ.

Information theoretically, the GMM model is learnable if the Gaussian means are well seperated. Let D represent the minimum distance between two distinct component means. Vempala and Wang (2004) show that, as long as $D > C$ for C being a sufficiently large constant, the estimation error on θ scales as $\mathcal{O}(n^{-1/2})$. However, the algorithm achieving this rate has $\mathcal{O}(k^k)$ time complexity. When the mutual distance D is large enough, there are $\mathrm{poly}(n, d, k)$-time algorithms to estimate the model parameters. In particular, Dasgupta (1999) presents an algorithm for $D = \Omega(\sqrt{d})$. Arora and Kannan (2005) and Dasgupta and Schulman (2000) present algorithms for $D = \Omega(d^{1/4})$. Vempala and Wang (2004) reduce this distance lower bound to $\Omega(k^{1/4})$. However, designing $\mathrm{poly}(n, d, k)$-time algorithm for $\widetilde{\Omega}(1)$-separated GMMs is a long-standing open problem.

Hsu and Kakade (2013) proposed a method that does not need the well-separation condition. The only assumption is that $\{\theta_1, \ldots, \theta_k\}$ are linearly independent. Let $\sigma_{\min} > 0$ be the smallest singular value of the matrix θ. Their algorithm runs in $\mathrm{poly}(n, d, k)$-time

1. For simplicity, we focus on the case when all mixture components have equal weights, but our argument can be easily generalized to the case of non-uniform weights.

and achieves the following bound for estimator $\widehat{\theta}$:

$$\|\widehat{\theta} - \theta\|_F^2 = \mathcal{O}\left(\frac{\mathrm{poly}(d, k, 1/\sigma_{\min}) \log(1/\delta)}{n}\right) \quad \text{with probability at least } 1 - \delta. \tag{55}$$

Here, $\|\cdot\|_F$ denotes the matrix Frobenius norm. In general, we cannot guarantee that $\sigma_{\min} > 0$. However, if we add a small perturbation on the true component means, then the assumption is satisfied almost surely. More precisely, we assume that there is a matrix $\theta^* \in \mathbb{R}^{d\times k}$ so that each entry of matrix θ is sampled from $\theta_{ij} \sim N(\theta^*_{ij}; \rho^2)$. The following lemma lower bounds the smallest singular value.

Lemma 18 (Ge et al. (2015), Lemma G.16) *Let $\theta^* \in \mathbb{R}^{d\times k}$ and suppose that $d \geq 3k$. If all entries of θ^* are independently perturbed by $N(0, \rho^2)$ to yield matrix θ. For any $\epsilon > 0$, with probability at least $1 - c_1(c_2\epsilon)^d$, the smallest singular value of matrix θ is lower bounded by:*

$$\sigma_{\min} > \epsilon\rho\sqrt{d}.$$

Here, c_1, c_2 are universal constants.

We choose $\epsilon, \rho \sim n^{-c}$ for a sufficiently small $c > 0$, then the perturbation diminishes to zero, and if $\sigma_{\min} > \epsilon\rho\sqrt{d}$ holds, then the right-hand side of equation (55) converges to zero at a polynomial rate as $n \to \infty$. Lemma 18 implies that the probability of this event is at least $1 - \mathcal{O}(n^{-cd})$. Thus, with high probability, the estimator $\widehat{\theta}$ is consistent under the smoothed analysis.

The convergence rate of the estimator $\widehat{\theta}$ can be improved if we add a mild assumption that $D = \widetilde{\mathcal{O}}(\sqrt{\log(nk)})$. Although the main focus of the paper is on lower bounds, the upper bound result on the estimation of $\widehat{\theta}$ in learning mixture of Gaussians is of its independent interest. To obtain the upper bound on $\mathbb{E}[\|\widehat{\theta} - \theta\|_F^2]$, we first establish the following lemma:

Lemma 19 *Let the mutual distance satisfy $D \geq c_1\sqrt{\log(nk/\delta)} \geq 3$ for a sufficiently large constant c_1. With probability at least $1 - \delta$, the inequality $\|x_i - \theta_j\|_2 - \|x_i - \theta_{z_i}\|_2 \geq c_2(d\log(nk/\delta))^{-1/2}$ holds for a constant $c_2 > 0$, for any $i \in [n]$ and any $j \in [k]\backslash\{z_i\}$.*

The proof of this technical lemma is relegated to Appendix D. Lemma 19 shows that with high probability, the distance of a random sample to its true component mean is significantly less than the distance to any other means. Let $\widehat{\theta}_j$ represent the j-th column of $\widehat{\theta}$. When the sample size n is sufficiently large, the method of Hsu and Kakade (2013) guarantees that $\|\widehat{\theta}_j - \theta_j\|_2 < o((d\log(nk/\delta))^{-1/2})$ for any $j \in [k]$. Thus, Lemma 19 implies that the distance of x_i to $\widehat{\theta}_{z_i}$ is smaller than the distance to any other estimated centers. As a consequence, we may recover the membership of instances by computing the center that is the closest to them.

$$\widehat{z}_i = \arg\min_{j\in[k]} \|x_i - \widehat{\theta}_j\|_2.$$

According to Lemma 19, with high probability we have $\widehat{z}_i = z_i$ for any $i \in [n]$. Given the membership, we refine the mean estimates by:

$$\widehat{\theta}_j \leftarrow \frac{\sum_{i:\widehat{z}_i=j} x_i}{|\{i : \widehat{z}_i = j\}|}.$$

Since the membership is uniformly assigned, with high probability the sample size of the j-th Gaussian component is lower bounded by $\frac{n}{2k}$. Thus, with high probability the squared error of $\widehat{\theta}_j$ will be upper bounded by $\mathcal{O}(dk/n)$. Since there are k components, the overall squared error is bounded by $\mathcal{O}(dk^2/n)$. Putting pieces together, we have an upper bound on the mean-squared error of parameter estimation.

Proposition 20 *Suppose that $d \geq 3k$ and n is greater than a fixed polynomial function of $(d, k, 1/\rho)$. Let the true parameter θ be ρ-perturbed from an arbitrary matrix $\theta^* \in \mathbb{R}^{d\times k}$. In addition, assume that the distances between the columns of θ^* are at least $D = c\sqrt{\log(nk)}$ for some universal constant c. Then there is a universal constant C such that the estimator $\widehat{\theta}$ described above achieves mean-square error:*

$$\mathbb{E}[\|\widehat{\theta} - \theta\|_F^2] \leq \frac{Cdk^2}{n}.$$

6.2 Minimax Risk of Smoothed Analysis

In this section, we formalize the notion of minimax risk under smoothed analysis. Similar to the classical statistical setting, the *minimax risk under smoothed analysis* can be defined in a game theoretic way. The learner first chooses an estimator $\widehat{\theta}$, then the adversary chooses a parameter θ^* from the parameter space Θ, which is randomly perturbed to form the true parameter θ. The data X is generated with respect to θ. Under this random perturbation framework, the minimax risk is defined as:

$$R_{\text{minimax}} := \inf_{\widehat{\theta}} \sup_{\theta^*\in\Theta} \mathbb{E}_\theta[L(\widehat{\theta}(X), \theta)] \tag{56}$$

where $L(\cdot, \cdot)$ is the loss function. In our GMM application, the parameters are the means of mixture components. The parameter space is the set of means whose mutual distances are lower bounded by D. The true parameter is generated by a random Gaussian perturbation with variance ρ^2. The loss is the Frobenius norm of the difference of matrices.

We note that the minimax risk (56) differs from the classical notion of minimax risk in that the adversary is not able to explicitly choose the true parameter θ. Instead, the true parameter is sampled from a prior distribution parametrized by θ^*. This Bayes nature makes it hard to lower bound the minimax risk (56) using the traditional Le Cam's or the Fano's method. In particular, both the Le Cam's method and the Fano's method lower bound the minimax risk by assuming a uniform prior over a carefully constructed discrete set. However, in our GMM setting, the prior distribution of parameter θ is always continuous.

Our Bayes risk lower bound naturally fits into the setting of smoothed analysis. Let w^* be an arbitrary prior distribution over θ^*. Since θ is perturbed from θ^*, the prior w^* induces a prior w over θ. It is easy to see that the Bayes risk with respect to w is a lower bound on the minimax risk (56). Thus, it suffices to lower bound the Bayes risk:

$$R_{\text{Bayes}}(w, L; \Theta) := \inf_{\widehat{\theta}} \mathbb{E}_{\theta\sim w}[L(\widehat{\theta}(X), \theta)].$$

For the GMM example, we construct the prior distribution w^* as follow: the j-th column of θ^*, namely the vector $\theta_j^* \in \mathbb{R}^d$, is sampled from the normal distribution $N(De_j; I_{d\times d})$,

where e_j is the unit vector of the j-th coordinate. As a consequence, the prior distribution w samples the j-th column of θ from the normal distribution $N(De_j; (1+\rho^2)I_{d\times d})$.

In the GMM setting, the membership variables z_i are unknown to the estimator. If we assume that the memberships are given to the estimator, it makes the problem easier so that the associated Bayes risk is a smaller than or equal to the original Bayes risk. Since we want to derive a lower bound, we make the assumption that the memberships are given, then partition the instances into k disjoint subsets according to their memberships. Let the j-th subset S_j be defined as $S_j := \{x_i : z_i = j\}$. Conditioning on the memberships, the distributions of $\{(\theta_j, S_j)\}_{j=1}^k$ are mutually independent. Thus, we have

$$R_{\text{Bayes}}(w, L; \Theta) \geq \sum_{j=1}^{k} \inf_{\widehat{\theta}_j} \mathbb{E}_{\theta_j \sim w_j}[L(\widehat{\theta}_j(S_j), \theta_j)] \geq \sum_{j=1}^{k} \mathbb{E}\Big[\inf_{\widehat{\theta}_j} \mathbb{E}_{\theta_j \sim w_j}[L(\widehat{\theta}_j(S_j), \theta_j)|n_j]\Big] \quad (57)$$

where w_j is the prior distribution $N(De_j; (1+\rho^2)I_{d\times d})$ and n_j is the cardinality of S_j. We focus on the inner term on the right-hand side, namely $\inf_{\widehat{\theta}_j} \mathbb{E}_{\theta_j \sim w_j}[L(\widehat{\theta}_j(S_j), \theta_j)|n_j]$, and find that it is the Bayes risk of Gaussian mean estimation with n_j i.i.d. samples, with the true parameter θ_j satisfying a Gaussian prior w_j. This Bayes risk can be easily lower bounded by the techniques that we develop in this paper.

Lemma 21 *Suppose that the standard deviation of normal perturbation $\rho \leq 1$ and $n_j \geq 1$. For a universal constant c, the Bayes risk is lower bounded by*

$$R_{\text{Bayes}}(w_j, n_j) := \inf_{\widehat{\theta}_j} \mathbb{E}_{\theta_j \sim w_j}[L(\widehat{\theta}_j(S_j), \theta_j)|n_j] \geq \frac{cd}{n_j}.$$

Proof [Proof of Lemma 21]

We denote the distribution of instances in S_j by P_{θ_j} and let $\mathcal{P}$ be the set of such distributions. Since the support of w_j is $\mathbb{R}^d$, we start by defining a prior whose support is an Euclidean ball of radius $\Gamma := \sqrt{2d}$. Let $\overline{w}$ be the truncated prior satisfying:

$$\overline{w}(x) = \begin{cases} w_j(x)/c_1 & \text{if } \|x - De_j\|_2 \leq \Gamma \\ 0 & \text{otherwise.} \end{cases}$$

The normalization factor c_1 is equal to the total mass of w in the ball $\{x : \|x - De_j\|_2 \leq \Gamma\}$. It is straightforward to verify that the radius Γ is sufficiently large so that c_1 is lower bounded by a universal constant. The prior $\overline{w}$ can be viewed as restricting the original prior in a finite radius. According to Remark 11, we may lower bound the Bayes risk by

$$R_{\text{Bayes}}(w_j, n_j) \geq c_1 \cdot R_{\text{Bayes}}(\overline{w}, n_j).$$

Thus, it suffices to lower bound the second term on the right-hand side.

We follow the similar steps of Example 1 to establish the lower bound. We start by upper bounding the terms $\sup_{a\in\mathcal{A}} \overline{w}(B_t(a, L))$ and the chi-squared informativity $I_{\chi^2}(\overline{w}, \mathcal{P})$. Using definition of the multivariate normal distribution, it is easy to see that

$$\sup_{a\in\mathcal{A}} \overline{w}(B_t(a, L)) = \overline{w}(B_t(De_j, L)) \leq \frac{V(\sqrt{t})}{c_1(2\pi(1+\rho^2))^{d/2}}$$

where $V(\sqrt{t})$ represents the volumn of the Euclidean ball of radius $\sqrt{t}$. Thus, there is a universal constant c_2 such that $\sup_{a\in\mathcal{A}} \overline{w}(B_t(a,L)) \le (c_2\sqrt{t}/\Gamma)^d$. On the other hand, we follow the same steps of Example 1 to upper bound the chi-square informativity. Note that our setup has n_j i.i.d. observations, but in Example 1 there is only one observation. In this generalized setup, the chi-square distance $\chi^2(P_\theta \| P_{\theta'})$ is equal to $\exp\left(n_j\|\theta-\theta'\|_2^2/\sigma^2\right) - 1$. Plugging this formula into the argument of Example 1, we obtain the upper bound $I_{\chi^2}(\overline{w},\mathcal{P}) \le (3e\Gamma\sqrt{n_j/d})^d - 1$.

Let I_f^{up} be the obtained informativity upper bound. If we choose $t = cd/n_j$ for a sufficiently small constant $c > 0$, then we have $\sup_{a\in\mathcal{A}} \overline{w}(B_t(a,L)) < \frac{1}{4}(1+I_f^{\mathrm{up}})^{-1}$. Corollary 12 then gives $R_{\mathrm{Bayes}}(\overline{w}, n_j) \ge cd/n_j$. ■

Combining inequality (57) and Lemma 21, we have

$$R_{\mathrm{Bayes}}(w, L; \Theta) \ge \sum_{j=1}^{k} \frac{cdk}{2n} \mathbb{P}(n_j \le 2n/k).$$

Recall that every n_j satisfies a binomial distribution $B(n, 1/k)$, which has median $\lfloor n/k \rfloor$ or $\lceil n/k \rceil$, thus the probability $\mathbb{P}(n_j \le 2n/k)$ will be at least $1/2$. It implies that the Bayes risk is lower bounded by $\Omega(dk^2/n)$. Putting pieces together, we have the following lower bound on the minimax risk.

Proposition 22 *Assume that the standard deviation of normal perturbation $\rho \le 1$, then for some universal constant c the minimax risk of smoothed analysis is lower bounded by $R_{\mathrm{minimax}} \ge c\frac{dk^2}{n}$.*

Comparing proposition 20 and proposition 22, we find that both the upper bound and the lower bound are tight. More precisely, under the assumptions of proposition 20, the minimax risk of smoothed analysis is precisely on the order of dk^2/n.

7. Bayes Risk Lower Bounds for Sparse Linear Regression

Linear regression is a canonical problem in machine learning and statistics. For a fixed design matrix $X \in \mathbb{R}^{n\times d}$ and an unknown parameter $\theta \in \mathbb{R}^d$, the learner observes a noise-corrupted response vector $y = X\theta + \varepsilon$, where ε satisfies an isotropic normal distribution $N(0, \sigma^2 I_{d\times d})$. The goal is to take the response vector as input and find an estimator $\widehat{\theta} \in \mathbb{R}^d$ for the true parameter θ. The risk is measured either by the estimation error $L_{\mathrm{est}}(\theta, \widehat{\theta}) := \|\widehat{\theta} - \theta\|_2^2$, or by the prediction error $L_{\mathrm{pre}}(\theta, \widehat{\theta}) := \|X\widehat{\theta} - X\theta\|_2^2$. Both errors will be studied in this section.

For high-dimensional linear regression, the dimension d can be much greater than the sample size n. In order to prevent over-fitting, one needs to impose structural assumptions on the true parameter, for example, assuming that the the number of non-zero entries in vector θ is at most k ($k \ll d$). Formally, we use $\mathbb{B}_0(k)$ to represent the set of k-sparse vectors in $\mathbb{R}^d$, and assume that $\theta \in \mathbb{B}_0(k)$. Under this setting, we want to compute an estimator $\widehat{\theta} \in \mathbb{R}^d$ to minimize the estimation error or the prediction error. Note that the estimator $\widehat{\theta}$ does not need to be k-sparse. Hence, our theoretical framework includes *improper learners* which are allowed to output non-sparse estimates whenever they achieve small risks.

The minimax risks of sparse linear regression have been well-studied. Under the same problem setting, Raskutti et al. (2011) proved information theoretic lower bounds on both the estimation error and the prediction error. Certain lower bounds have also been proved under the computation tractability constraint Zhang et al. (2014), or proved for the family of regularized M-estimators Zhang et al. (2015). All these lower bounds handle the worst-case scenario — given an arbitrary estimator, they prove the existence of a parameter θ that attains the lower bound. This setting might be too pessimistic in practice. The goal of this section is to study the Bayes risk of sparse linear regression under a natural prior, whose construction is described in the next subsection.

7.1 Prior Definition and Assumptions

We define a prior over k-sparse d-dimensional vectors for the true parameter $\theta \in \mathbb{R}^d$, referred to as distribution w, as follows:

1. Uniformly sample a subset of k indices from the integer set $\{1, 2, \ldots, d\}$, naming this subset by K.

2. For every index $i \in K$, the coordinate θ_i is generated by sampling from the normal distribution $N(0, \tau^2)$. For any $i \notin K$, define $\theta_i := 0$.

Given an index set K, we use θ_K as a shorthand notation to denote the coordinates of the vector $\theta \in \mathbb{R}^d$ whose indices belong to the set K. Similarly, we use θ_{-K} to denote the subvector whose indices are not in K. Then the the second step of the above generative process can be rephrased as generating $\theta_K \sim N(0, \tau^2 I_{k\times k})$ and defining $\theta_{-K} = 0$. It is clear that the sampled θ belongs to the k-sparse ℓ_0-ball $\mathbb{B}_0(k) := \big\{\theta \in \mathbb{R}^d : \ \|\theta\|_0 \le k\big\}$.

One may consider variants of the the prior defined above. For example, one can assume that the number of non-zero entries of the vector θ is not exactly equal to k, but random sampled from a Poisson distribution with mean k. One may also redefine the prior of non-zero entries to be a non-Gaussian distribution. However, these variants don't add essential technical challenge to the analysis, thus we focus on the the prior w as a concrete example for illustrating the general idea.

We make an additional assumption on the design matrix X that is important for characterizing the minimax risk (see, e.g. Raskutti et al., 2011), and in this section, we study their effects on the Bayes risk. Specifically, the design matrix X satisfies the *sparse eigenvalue conditions* with parameter (κ_u, κ_ℓ) if:

$$\kappa_\ell \|\beta\|_2 \le \frac{\|X\beta\|_2}{\sqrt{n}} \le \kappa_u \|\beta\|_2 \quad \text{for any } (2k)\text{-sparse vector } \beta \in \mathbb{R}^d. \tag{58}$$

Here, both κ_u and κ_ℓ are positive constants. As a concrete example, if entries of the matrix X are i.i.d. sampled from a normal distribution, then the matrix is called a *Gaussian random design*. This type of matrices have been extensively studied for sparse linear regression (Candes et al., 2006; Guédon et al., 2008), and proved to satisfy condition (58) with $\kappa_u/\kappa_\ell = \mathcal{O}(1)$ (Raskutti et al., 2010). For the rest of this section, we assume that the design matrix X satisfies the condition (58).

7.2 Bayes Risk Lower Bounds

For sparse linear regression, we denote the parameter space and action space by $\Theta = \mathbb{B}_0(k)$ and $\mathcal{A} = \mathbb{R}^d$, respectively. We present a Bayes risk lower bound with respect to the prior distribution defined in Section 7.1, then demonstrate its consequences.

Theorem 23 *Assume that the design matrix X satisfies the sparse eigenvalue condition* (58)*, and that $d > k^3$. There are universal constants $c', c'' > 0$ such that for any $\tau > 0$, we have Bayes risk lower bounds: $R_{\mathrm{Bayes}}(w, L_{\mathrm{est}}; \Theta) \geq c'\, T(\tau)$ and $R_{\mathrm{Bayes}}(w, L_{\mathrm{pre}}; \Theta) \geq c''\, \kappa_\ell^2 T(\tau)$, where $T(\tau)$ is a term defined by*

$$T(\tau) := k\tau^2 \max\Big\{ \frac{1}{1 + \kappa_u^2 \tau^2 n/\sigma^2}, \exp\Big(- \frac{4\kappa_u^2 n}{\sigma^2} \Big[\tau^2 - \frac{\sigma^2 \log(d/k)}{16 \kappa_u^2 n} \Big]_+ \Big) \Big\}. \tag{59}$$

The proof of Theorem 23 follows the general strategy that we sketched in earlier sections: first, we bound the mutual informativity using the techniques described in Section 5, then we upper bound the probability $\sup_{a \in \mathcal{A}} w(B_t(a, L))$ for a specific scalar $t > 0$. Combining the two upper bounds with Corollary 12 establishes the theorem. See Appendix E for the proof. We make a few important remarks of this result in the below.

Estimation versus prediction By Theorem 23, the lower bounds on the estimator error and the prediction error differ by a factor κ_ℓ^2. As a consequence, if we multiply a constant to the design matrix, then the term κ_ℓ^2 will also be scaled. If the scalar is very small, then the lower bound on the prediction error will be close to zero, but the lower on the estimation error won't. These are the right scaling for both risks. Indeed, when the design matrix converges to an all-zero matrix, the true parameters will be hard to identify, but the constant estimator $\widehat{\theta} \equiv 0$ will be able to achieve a small prediction error.

Comparison with minimax risk lower bounds It is worth comparing Theorem 23 with the well-studied minimax risk lower bound. Under the sparse eigenvalue condition (58), Raskutti et al. (2011) proved the follow minimax risk lower bound:

$$\inf_{\widehat{\theta}} \max_{\theta \in \mathbb{B}_0(k)} \mathbb{E}[L_{\mathrm{est}}(\theta, \widehat{\theta})] \geq c' \frac{\sigma^2 k \log(d/k)}{\kappa_u^2 n} \quad \text{and} \quad \inf_{\widehat{\theta}} \max_{\theta \in \mathbb{B}_0(k)} \mathbb{E}[L_{\mathrm{pre}}(\theta, \widehat{\theta})] \geq c'' \frac{\kappa_\ell^2 \sigma^2 k \log(d/k)}{\kappa_u^2 n}, \tag{60}$$

where c' and c'' are universal constants. These bounds are matched by Theorem 23. In particular, if we assume $d > k^3$ and consider the prior distribution with variance:

$$\tau^2 = \left(\tau_*^2 := \frac{\sigma^2 \log(d/k)}{16 \kappa_u^2 n} \right), \tag{61}$$

then expression (59) implies $T(\tau) = k\tau^2$, and as a consequence, we have

$$R_{\mathrm{Bayes}}(w, L_{\mathrm{est}}; \Theta) \geq c' \frac{\sigma^2 k \log(d/k)}{\kappa_u^2 n} \quad \text{and} \quad R_{\mathrm{Bayes}}(w, L_{\mathrm{pre}}; \Theta) \geq c'' \frac{\kappa_\ell^2 \sigma^2 k \log(d/k)}{\kappa_u^2 n}, \tag{62}$$

where c' and c'' are universal constants. The minimax risk lower bounds (60) and the Bayes risk lower bounds (62) thus match by a universal constant factor. Therefore, using our

technique, we can directly obtain this classical minimax result on sparse linear regression. It is worth noting that the lower bounds of Raskutti et al. (2011) were proved by constructing a uniform prior over a discrete packing set over the parameter space. The existence of the proper packing set was proved in a non-constructive, worst-case fashion, which might be too pessimistic in practice. In contrast, our lower bound was established for a realistic and flexible prior which admits a simple closed-form definition and allows for different levels of variance. The theorem also shows that the prior w with the variance level (61) is in fact *a least favorable prior* for sparse linear regression.

Bayes risk on the spectrum of priors Besides the least-favorable setting (61), let us consider the Bayes risk under other choices of the parameter τ^2. When $\tau^2 < \tau_*^2$, Theorem 23 implies

$$R_{\text{Bayes}}(w, L_{\text{est}}; \Theta) \geq c' \, k\tau^2 \quad \text{and} \quad R_{\text{Bayes}}(w, L_{\text{pre}}; \Theta) \geq c'' \, \kappa_\ell^2 k\tau^2. \tag{63}$$

When $\tau^2 \to +\infty$, Theorem 23 implies

$$R_{\text{Bayes}}(w, L_{\text{est}}; \Theta) \geq c' \, \frac{k\sigma^2}{\kappa_u^2 n} \quad \text{and} \quad R_{\text{Bayes}}(w, L_{\text{pre}}; \Theta) \geq c'' \, \frac{\kappa_\ell^2 k\sigma^2}{\kappa_u^2 n}. \tag{64}$$

In both cases, the Bayes risk lower bounds can be significantly smaller than the minimax risk. We argue that these lower bounds are essentially tight under specific assumptions. That is, when taking the prior information into account, we can indeed achieve better rates than the minimax rate.

First, notice that the upper bound:

$$\mathbb{E}_{\theta \sim w}[L_{\text{est}}(\theta, \widehat{\theta})] \leq k\tau^2 \quad \text{and} \quad \mathbb{E}_{\theta \sim w}[L_{\text{est}}(\theta, \widehat{\theta})] \leq \kappa_u^2 k\tau^2.$$

can always be achieved using the constant estimator $\widehat{\theta} \equiv 0$. It means that for the case of $\tau^2 < \tau_*^2$, the lower bounds (63) are tight under the assumption $\kappa_u / \kappa_\ell = \mathcal{O}(1)$.

For the case of $\tau^2 \to +\infty$, we consider the ℓ_0-norm constrained estimator:

$$\widehat{\theta} := \arg \inf_{\beta \in \mathbb{B}_0(k)} \|X\beta - y\|_2^2. \tag{65}$$

Whenever $\kappa_u / \kappa_\ell = \mathcal{O}(1)$, Raskutti et al. (2011) showed that the estimator (65) achieves an error bound $\|\widehat{\theta} - \theta\|_2^2 \leq c \, \frac{k \log(d)}{n}$ with high probability for a constant $c > 0$. Suppose that $\tau^2 = C \, \frac{k \log(d)}{n}$ with a scaling factor $C > c$. For any $i \in K$, the expectation of θ_i^2 is equal to τ^2, so that the probability of $\theta_i^2 \leq c \, \frac{k \log(d)}{n}$ is bounded by $\mathcal{O}(c/C)$. It means that by choosing a large enough C (specifically, choosing $C \gg ck$), the lower bound $\theta_i^2 > c \, \frac{k \log(d)}{n}$ will hold for every $i \in K$ with a probability close to 1. Combining this fact with the bound $\|\widehat{\theta} - \theta\|_2^2 \leq c \, \frac{k \log(d)}{n}$, we find that the support of $\widehat{\theta}$ must agree with K, so that the estimator must satisfy:

$$\widehat{\theta}_K = \arg \inf_{\beta \in \mathbb{R}^k} \|X_K \beta - y\|_2^2 \quad \text{and} \quad \widehat{\theta}_{-K} = 0,$$

where X_K is a submatrix of X consisting of columns indexed by K. In other words, the vector $\widehat{\theta}_K$ is the least-square estimator for a k-dimensional linear regression problem. For estimators taking this form, both the estimation error and the prediction error are known to match the lower bound (64) with high probability.

8. Conclusions

In this paper, we presented lower bounds for the Bayes risk in abstract decision-theoretic problems. Our bounds are quite general and only require upper bounds on $\sup_{a\in\mathcal{A}} w(B_t(a,L))$ and the f-informativity $I_f(w,\mathcal{P})$ for their application. Because of the generality, the bounds are not always tight however. For example, the bounds involve $\sup_{a\in\mathcal{A}} w(B_t(a,L))$ and this quantity becomes large when the prior w has a spike. In such situations, our main Bayes risk lower bound in Theorem 9 will not be tight. In specific examples, this looseness can be remedied by adhoc fixes such as the one described in Remark 11. Obtaining tight lower bounds for the Bayes risk in the generality considered in this paper is a challenging open problem.

Acknowledgments

Adityanand Guntuboyina is supported by NSF Grant DMS-1309356. The authors would like to thank Michael I. Jordan and Sivaraman Balakrishnan for helpful discussions.

Appendix A. Proofs and Additional Results for Section 3 on Bayes Risk Lower Bound for Zero-one Loss

A.1 Proof of Lemma 1

Recall the expression (13) of $\phi_f(a,b)$. We first fix b and show that $g(a) : a \mapsto \phi_f(a,b)$ is a non-increasing for $a \in [0,b]$. There is nothing to prove if $b = 0$ so let us assume that $b > 0$. We will consider the cases $0 < b < 1$ and $b = 1$ separately. For $0 < b < 1$, note that for every $a \in (0,b]$, we have,

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

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 $\frac{a}{b} \le \frac{1-a}{1-b}$ for every $0 \le a \le b$ and f is convex, we see that

$$g'_L(a) \le f'_R\left(\frac{a}{b}\right) - f'_R\left(\frac{1-a}{1-b}\right) \le 0$$

for every $a \in (0,b]$ which implies that $g(a)$ is non-increasing on $[0,b]$.

When $b = 1$, we have $g'_L(a) = f'_L(a) - f'(\infty)$ which is always ≤ 0 because f is convex (note that $f'(\infty) = \lim_{x\uparrow\infty} f(x)/x = \lim_{x\uparrow\infty}(f(x) - f(1))/(x-1)$).

The convexity and continuity of g follow from the convexity of f and the expression for ϕ_f.

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

$$h'_R(b) = f\left(\frac{a}{b}\right) - \frac{a}{b}f'_L\left(\frac{a}{b}\right) - f\left(\frac{1-a}{1-b}\right) + \frac{1-a}{1-b}f'_R\left(\frac{1-a}{1-b}\right), \tag{66}$$

where h'_R represents the right derivative of h. By the convexity of f,

$$f\left(\frac{a}{b}\right) - f\left(\frac{1-a}{1-b}\right) \ge f'_R\left(\frac{1-a}{1-b}\right)\left(\frac{a}{b} - \frac{1-a}{1-b}\right). \tag{67}$$

Combining (66) with (67), we obtain that,

$$h'_R(b) \ge \frac{a}{b}\left(f'_R\left(\frac{1-a}{1-b}\right) - f'_L\left(\frac{a}{b}\right)\right) \ge \frac{a}{b}\left(f'_L\left(\frac{1-a}{1-b}\right) - f'_L\left(\frac{a}{b}\right)\right) \ge 0,$$

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

A.2 A Variant of Fano's Inequality from Braun and Pokutta (2014)

One of the main results in Braun and Pokutta (2014) (Proposition 2.2) establishes the following variant of Fano's inequality. Consider the setting of Lemma 3. In particular, recall the quantities $R^{\mathfrak{d}}$ and $R^{\mathfrak{d}}_Q$ from (21) and also the sets $B(a), a \in \mathcal{A}$ from (16). (Braun and Pokutta, 2014, Proposition 2.2) proved the following: for any decision rule $\mathfrak{d}$,

$$R^{\mathfrak{d}} \ge \frac{-I(w,\mathcal{P}) - H(R^{\mathfrak{d}}) - \log w_{\max}}{\log\left[(1-w_{\min})/w_{\max}\right]}, \tag{68}$$

where $H(x) := -x\log x-(1-x)\log(1-x)$, $w_{\min} := \inf_{a\in\mathcal{A}} w(B(a))$ and $w_{\max} := \sup_{a\in\mathcal{A}} w(B(a))$.

Below we provide a proof of this inequality using Lemma 3. The proof given in Braun and Pokutta (2014) is quite different proof. Using (20) from Lemma 3 with $f(x) = x\log x$, we have for any decision rule

$$\int_\Theta D_f(P_\theta\|Q)w(d\theta) \geq R^{\mathfrak{d}}\log\frac{R^{\mathfrak{d}}}{R^{\mathfrak{d}}_Q} + (1-R^{\mathfrak{d}})\log\frac{1-R^{\mathfrak{d}}}{1-R^{\mathfrak{d}}_Q}.$$

We can rewrite this as

$$\int_\Theta D_f(P_\theta\|Q)w(d\theta) \geq -H(R^{\mathfrak{d}}) - R^{\mathfrak{d}}\log R^{\mathfrak{d}}_Q - (1-R^{\mathfrak{d}})\log(1-R^{\mathfrak{d}}_Q) \tag{69}$$

where $H(x) := -x\log x - (1-x)\log(1-x)$. Since L in Lemma 3 is zero-one valued.

$$R^{\mathfrak{d}}_Q = 1 - \mathbb{E}_Q w(B(\mathfrak{d}(X))) \tag{70}$$

where $\mathbb{E}_Q$ denotes expectation taken under $X\sim Q$ and and $B(\mathfrak{d}(X))$ is defined in (16). As a result, we have

$$1 - \max_{a\in\mathcal{A}} w(B(a)) \leq R^{\mathfrak{d}}_Q \leq 1 - \min_{a\in\mathcal{A}} w(B(a)). \tag{71}$$

Using the bounds in (71) on the right hand side of (69), we deduce

$$\int_\Theta D_f(P_\theta\|Q)w(d\theta) \geq -H(R^{\mathfrak{d}}) - R^{\mathfrak{d}}\log(1-w_{\min}) - (1-R^{\mathfrak{d}})\log w_{\max}.$$

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

$$I(w,\mathcal{P}) \geq -H(R^{\mathfrak{d}}) - R^{\mathfrak{d}}\log(1-w_{\min}) - (1-R^{\mathfrak{d}})\log(w_{\max}).$$

Provided $w_{\min} + w_{\max} < 1$, one can rewrite the above inequality as (68). This completes the proof of (68).

A.3 Proof of Corollary 7

1. **Proof of inequality** (26): Applying Theorem 2 with $f(x) = x^2 - 1$, we obtain

$$I_{\chi^2}(w,\mathcal{P}) \geq \frac{(R_0-R)^2}{R_0(1-R_0)}$$

   Because $R \leq R_0$, we can invert the above to obtain (26).

2. **Proof of inequality** (27): Theorem 2 with $f(x) = |x-1|/2$ gives

$$I_{TV}(w,\mathcal{P}) \geq \frac{R_0}{2}\left|\frac{R}{R_0} - 1\right| + \frac{1-R_0}{2}\left|\frac{1-R}{1-R_0} - 1\right| = R_0 - R,$$

   where the last equality uses the fact that $R \leq R_0$. Inverting the above inequality, we obtain (27).

3. **Proof of inequality** (28): Theorem 2 with $f(x) = f_{1/2}(x) = 1 - \sqrt{x}$ gives

$$I_{f_{1/2}}(w, \mathcal{P}) \geq 1 - \sqrt{RR_0} - \sqrt{(1-R)(1-R_0)}. \tag{72}$$

Assume that P_θ has density p_θ with respect to a common dominating measure μ. We shall show below that

$$I_{f_{1/2}}(w, \mathcal{P}) = 1 - \sqrt{\int_{\mathcal{X}} u^2 d\mu} \quad \text{where } u := \int_\Theta \sqrt{p_\theta} w(d\theta). \tag{73}$$

To see this, fix a probability measure Q that has a density q with respect to μ. We can then write

$$\int_\Theta D_{f_{1/2}}(P_\theta \| Q) w(d\theta) = 1 - \int_{\mathcal{X}} \sqrt{q} \left(\int_\Theta \sqrt{p_\theta} w(d\theta) \right) d\mu = 1 - \int_{\mathcal{X}} \sqrt{qu^2} d\mu$$

It follows then from the Cauchy-Schwarz inequality that

$$\int_\Theta D_{f_{1/2}}(P_\theta || Q) w(\mathrm{d}\theta) = 1 - \int_{\mathcal{X}} \sqrt{qu^2} \, \mathrm{d}\mu \geq 1 - \sqrt{\int_{\mathcal{X}} u^2 \, \mathrm{d}\mu},$$

with equality holding when q is proportional to u^2. This proves (73). We now see that

$$\int_{\mathcal{X}} u^2 \, \mathrm{d}\mu = \int_\Theta \int_\Theta \int_{\mathcal{X}} \sqrt{p_\theta} \sqrt{p_{\theta'}} \, \mathrm{d}\mu \; w(\mathrm{d}\theta) w(\mathrm{d}\theta') = 1 - \frac{1}{2} h^2 \tag{74}$$

where h^2 is defined as

$$h^2 = \int_\Theta \int_\Theta H^2(P_\theta \| P_{\theta'}) w(\mathrm{d}\theta) w(\mathrm{d}\theta'). \tag{75}$$

This, together with (72) and (73), gives the inequality

$$\sqrt{RR_0} + \sqrt{(1-R)(1-R_0)} \geq \sqrt{1 - \frac{h^2}{2}} \tag{76}$$

Now under the assumption $h^2 \leq 2R_0$, the right hand side of the inequality (76) lies between $\sqrt{1-R_0}$ and 1. On the other hand, it can be checked that, as a function in R, the left hand side of (76) is strictly increasing from $\sqrt{1-R_0}$ (at $R = 0$) to 1 at $(R = R_0)$. Therefore, from (76), we know that $R \geq \widehat{R}$ where $\widehat{R} \in [0, R_0]$ is the solution to the equation obtained by replacing the inequality (76) with an equality. One can solve this equation and obtain two solutions. One of two solutions can be discarded by the fact that $R \leq R_0$. The other solution is given by:

$$\widehat{R} = R_0 - (2R_0 - 1)\frac{h^2}{2} - \sqrt{R_0(1-R_0)} \sqrt{h^2(2-h^2)}$$

and thus we have $R \geq \widehat{R}$ which proves inequality (28).

We note that the lower bound on R in (28) only holds under the condition $h^2 \leq 2R_0$. When $h^2 > 2R_0$, inequality (28) holds for every $R \in [0, R_{Q^*}]$ and thus cannot provide a non-trivial lower bound on R. As an example, when $\Theta = \mathcal{A} = \{1, \ldots, N\}$, $L(\theta, a) = \mathbb{I}\{\theta \neq a\}$ and w is the uniform prior on Θ, it is easy to see that $R_0 = 1 - (1/N)$ and

$$h^2 = \frac{1}{N^2} \sum_{\theta \neq \theta'} H^2(P_\theta \| P_{\theta'}) \leq 2 \frac{N(N-1)}{N^2} = 2R_{Q^*}. \tag{77}$$

Inequality (28) therefore is equivalent to

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

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

A.4 Derivations of Le Cam's Inequality (Two Hypotheses) and Assouad's Lemma and other Results from Corollary 7

To demonstrate the application of Corollary 7, we apply it to derive the two hypotheses version of Le Cam's inequality (with total variation distance) and Assouad's lemma (see Theorem 2.12 in (Tsybakov, 2010)).

The simplest version of the Le Cam's inequality, the so-called two-point argument, is an easy corollary of (27). Indeed, applying (27) with $\Theta = \mathcal{A} = \{\theta_0, \theta_1\}$, $L(\theta, a) = \mathbb{I}\{\theta \neq a\}$ and $w\{0\} = w\{1\} = 1/2$ (and note that $R_0 = 1/2$), we obtain that for any distribution Q on $\mathcal{X}$,

$$\frac{1}{2} \left(\|P_{\theta_0} - Q\|_{TV} + \|P_{\theta_1} - Q\|_{TV} \right) \geq I_{TV}(w, \mathcal{P}) \geq 1/2 - R.$$

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

$$R_{\text{minimax}} \geq \frac{1}{2} \left(1 - \|P_{\theta_0} - P_{\theta_1}\|_{TV} \right). \tag{78}$$

The more involved Le Cam's inequality considers $\Theta = \mathcal{A} = \Theta_0 \cup \Theta_1$ for two disjoint subsets Θ_0 and Θ_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 Θ_0 and Θ_1 respectively,

$$R_{\text{minimax}} \geq \frac{1}{2} \left(1 - \|m_0 - m_1\|_{TV} \right) \tag{79}$$

where m_0 and m_1 are marginal densities given by $m_\tau(x) = \int p_\theta(x) w_\tau(\mathrm{d}\theta)$ for $\tau = 0, 1$. To prove (79), consider the prior $w = (w_0 + w_1)/2$. 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 $\tau \sim$ Uniform $\{0, 1\}$ and then $X \sim m_\tau$. The decision $a \in \mathcal{A}$ can be converted into the binary decision $\hat{\tau} = \mathbb{I}(a \in \Theta_1)$. The loss function is $L(\tau, \hat{\tau}) = \mathbb{I}(\tau \neq \hat{\tau})$. The Bayes risk under the prior w can be re-written as,

$$R_{\text{Bayes}}(w, L; \Theta) = \frac{1}{2} \inf_{\hat{\tau}} \sum_{\tau = 0,1} \int_{\mathcal{X}} \mathbb{I}(\tau \neq \hat{\tau}(x)) m_\tau(x) \mu(\mathrm{d}x), \tag{80}$$

which has the same form as the Bayes risk in the earlier binary testing problem. Applying the same argument as for proving (78), we obtain the lower bound on the Bayes risk in (80), $R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2}(1 - \|m_0 - m_1\|_{TV})$, which further implies (79).

Another classical minimax inequality involving the total variation distance is Assouad's inequality (Assouad, 1983) which states that if $\Theta = \mathcal{A} = \{0,1\}^d$ and the loss function L is defined by the Hamming distance, i.e., $L(\theta, a) = \sum_{i=1}^{d} \mathbb{I}(\theta_i \neq a_i)$, then

$$R_{\text{minimax}} \geq \frac{d}{2} \min_{L(\theta,\theta')=1} (1 - \|P_\theta - P_{\theta'}\|_{TV}). \tag{81}$$

This inequality is also a consequence of (27): let w be the uniform probability measure on Θ and $L_1(\theta, a) = \mathbb{I}(\theta_1 \neq a_1)$. Under w, the marginal distribution of the first coordinate is $w_1\{0\} = w_1\{1\} = 1/2$. Let $m_\tau(x) := \sum_{\theta:\theta_1=\tau} p_\theta(x)/2^{d-1}$ 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 (78), we obtain that the minimax risk for the zero-one valued loss function $L_1(\theta, a)$ is bounded below by $\frac{1}{2}(1 - \|m_0 - m_1\|_{TV}) \geq \frac{1}{2}\min_{L(\theta,\theta')=1}(1 - \|P_\theta - P_{\theta'}\|_{TV})$. Repeating this argument for $L_i(\theta, a) := \mathbb{I}\{\theta_i \neq a_i\}$ for $i = 2, \ldots, d$ and adding up the resulting bounds, we obtain (81).

By using Le Cam's inequality (see, e.g., Lemma 2.3 in (Tsybakov, 2010)) which states that:

$$\|P_\theta - P_{\theta'}\|_{TV} \leq \sqrt{H^2(P_\theta \| P_{\theta'}) \left(1 - \frac{1}{4} H^2(P_\theta \| P_{\theta'})\right)},$$

the inequality in (81) further implies the Hellinger distance version of Assouad's inequality in the book Tsybakov (2010, Theorem 2.12), i.e.,

$$R_{\text{minimax}} \geq \frac{d}{2} \min_{L(\theta,\theta')=1} \left\{1 - \sqrt{H^2(P_\theta \| P_{\theta'}) \left(1 - \frac{1}{4} H^2(P_\theta \| P_{\theta'})\right)}\right\}. \tag{82}$$

A.5 Comparison of the Bounds for Different Divergences

We provide some qualitative comparisons of Bayes risk lower bounds given by Theorem 2 for different power divergences. In particular, let us consider the discrete setting where $\Theta = \mathcal{A} = \{\theta_1, \ldots, \theta_N\}$, $L(\theta, a) = \mathbb{I}\{\theta \neq a\}$, and w is the discrete uniform. Note that in such a "multiple testing problem" setup, R_0 is equal to $1 - (1/N)$. We take N sufficiently large so that R_0 is close to 1. To establish minimax lower bounds, a typical approach is to reduce the estimation problem to a multiple hypotheses testing problem in the aforementioned setup, then try to prove that the Bayes risk $R \geq c > 0$ (see Section 2.2. in Tsybakov (2010)). Without loss of generality, we take $c = 1/2$ and we shall see how the three inequalities (25), (26) and (28) work to establish $R \geq 1/2$.

Let us start with (25) corresponding to KL divergence, which is equivalent to the classical Fano's inequality (3) in the discrete setting. To establish $R \geq 1/2$, the following condition should hold:

$$I(w, \mathcal{P}) \leq \frac{1}{2} \log\left(\frac{N}{4}\right). \tag{83}$$

We remark that $I(w, \mathcal{P})$ is at most $\log N$ even if every the pairwise KL divergence $D(P_{\theta_i} \| P_{\theta_j})$ equals ∞ for $i \neq j$. This fact will be clear from the inequality (47) from Section 5 (let $M = N$

and $Q_j = P_{\theta_j}$ for $1 \leq j \leq M$). The upper bound on $I(w, \mathcal{P})$ in (47) further provides a sufficient condition to verify (83).

Now we turn to (26) corresponding to the chi-squared divergence. Since $R_0 = 1-(1/N)$, inequality (26) implies a sufficient condition for $R \geq 1/2$:

$$I_{\chi^2}(w, \mathcal{P}) \leq \frac{N^2}{N-1}\left(\frac{1}{2} - \frac{1}{N}\right)^2. \tag{84}$$

When N is large, the above condition is equivalent to $I_{\chi^2}(w, \mathcal{P}) \leq N/4$. Note that the maximum possible value of $I_{\chi^2}(w, \mathcal{P})$ in this discrete setting is $N-1$ (even when $\chi^2(P_{\theta_i} \| P_{\theta_j}) = \infty$ for every $i \neq j$) and this follows from our upper bounds on f-informativity for a class of power divergences in (50) (see Section 5).

The conditions (83) and (84) don't imply each other. The chi-squared divergence is always greater than the KL divergence (see Lemma 2.7 in Tsybakov (2010)), but the upper bound required by (84) is also weaker than that required by (83). For both divergences, constructing more hypotheses (i.e., choosing $N > 2$) is often helpful for showing $R \geq 1/2$.

For the Hellinger distance (inequality (28)), we claim that it gives no more useful bounds than those obtained by a simple two point argument. To see this, since $R_0 = 1 - (1/N)$, inequality (28) implies

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

where $h^2 = \sum_{i,j} H^2(P_{\theta_i}||P_{\theta_j})/N^2$. When N is large, the above inequality reduces to effectively $R \geq 1 - (h^2/2)$. Therefore a sufficient condition for $R \geq 1/2$ is $h^2 \leq 1$, which is equivalent to,

$$\frac{1}{N(N-1)/2}\sum_{i<j} H^2(P_{\theta_i}||P_{\theta_j}) \leq \frac{N}{N-1}.$$

When N is large, the above displayed condition implies the existence of $i < j$ for which $H^2(P_{\theta_i}||P_{\theta_j}) \leq 1$. Let $\tilde{w}$ denote the prior $\tilde{w}\{i\} = \tilde{w}\{j\} = 1/2$. It is easy to see that the Bayes risk for $\tilde{w}$ equals $R_{\text{Bayes}}(\tilde{w}) = \frac{1}{2}\left(1 - \|P_{\theta_i} - P_{\theta_j}\|_{TV}\right)$. By Le Cam's inequality (see Lemma 2.3 in Tsybakov (2010)), we have,

$$R_{\text{Bayes}}(\tilde{w}) \geq \frac{1}{2}\left(1 - H(P_{\theta_i}||P_{\theta_j})\sqrt{1 - \frac{H^2(P_{\theta_i}||P_{\theta_j})}{4}}\right)$$

Since $H(P_{\theta_i}||P_{\theta_j}) \leq 1$, it is easy to verify from the above that $R_{\text{Bayes}}(\tilde{w}) \geq 1/8$. Therefore in this discrete setting, if inequality (28) implies $R_{\text{Bayes}}(w) \geq 1/2$, then there is a much simpler two point prior $\tilde{w}$ for which $R_{\text{Bayes}}(\tilde{w}) \geq 1/8$. It shows that for Hellinger distance, considering $N > 2$ hypotheses is not more useful than using a pair of hypotheses. The reason is that the Hellinger informativity can be written as an expression involving pairwise Hellinger distances. In particular, it can be seen from the proof of inequality (28) that

$$I_{f_{1/2}}(w, \mathcal{P}) = 1 - \left(1 - \frac{1}{2N^2}\sum_{i,j} H^2(P_{\theta_i}||P_{\theta_j})\right)^{1/2}.$$

In contrast, the mutual information, $I(w, \mathcal{P})$, cannot be written in terms of $D(P_{\theta_i} \| P_{\theta_j})$ for $i \neq j$ (recall that $I(w, \mathcal{P})$ is always at most $\log N$ even when $D(P_{\theta_i} \| P_{\theta_j}) = \infty$ for all $i \neq j$). The same holds for $I_{\chi^2}(w, \mathcal{P})$ as well (which is always at most $N-1$ even if $\chi^2(P_{\theta_i} \| P_{\theta_j}) = \infty$ for all $i \neq j$).

If the eventual goal of obtaining Bayes risk lower bounds is to obtain lower bounds up to multiplicative constants on the minimax risk, then the bound in (28) gives no more useful bounds than those obtained by the simple two point argument. In this sense, inequality (28) induced by Hellinger distance is not as useful as inequalities (25) and (26). In fact, the Hellinger distance is seldom used in lower bounding minimax risk involving many hypotheses (for example, none of the minimax rates in the examples of Tsybakov (2010) involving multiple hypotheses testing are established via Hellinger distance).

Appendix B. Proofs and Additional Results for Section 5 on Upper Bounds on f-informativity

B.1 Proof of Lemma 15

Let $\phi(t) \equiv t^r$ with $\phi'(t) = rt^{r-1}$ and $\phi''(t) = r(r-1)t^{r-2}$ and $\varphi(t) = t^{1/r}$ with $\varphi'(t) = \frac{1}{r}t^{(1-r)/r}$. Then

$$f(u) = \varphi\left(\int_T \phi(u(t))\mu(\mathrm{d}t)\right).$$

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

$$h(s) = \varphi\left(\int_T \phi(u(t) + sv(t))\mu(\mathrm{d}t)\right), \tag{85}$$

for arbitrary $u, v \in L^r_\mu(T)$. 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 $g(s) = \int_T \phi(u(t) + sv(t))\mu(\mathrm{d}t)$,

$$\begin{aligned}
h'(s) =& \varphi'(g(s)) \int_T \phi'(u(t) + sv(t))v(t)\mu(\mathrm{d}t) \\
h''(s) =& \varphi''(g(s)) \left(\int_T \phi'(u(t) + sv(t))v(t)\mu(\mathrm{d}t)\right)^2 \\
&+ \varphi'(g(s)) \int_T \phi''(u(t) + sv(t))v^2(t)\mu(\mathrm{d}t)
\end{aligned}$$

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

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

Applying the Cauchy-Schwarz inequality

$$\left(\int_T a(t)b(t)\mu(\mathrm{d}t)\right)^2 \leq \left(\int_T a(t)^2\mu(\mathrm{d}t)\right)\left(\int_T b(t)^2\mu(\mathrm{d}t)\right)$$

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

B.2 Example Demonstrating the Effectiveness of Theorem 14

In this example, we show the tightness of the upper bound in (49) 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_ξ be the n-fold product of $N(\xi,1)$ where $\xi \sim N(0,1)$. 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 $\chi^2(P||Q_\xi) = \exp(n\xi^2) - 1$, the right hand side of (49) equals to $\sqrt{2n+1} - 1$. The term $\chi^2(P||\bar{Q})$ on the left hand side of (49) is difficult to evaluate. However, we can lower bound $\chi^2(P||\bar{Q})$ using the following standard inequality $\exp\left(D(P||\bar{Q})\right) - 1 \le \chi^2(P||\bar{Q})$ (see Lemma 2.7 in Tsybakov (2010)). By the closed-form expression for KL divergence between two multivariate Gaussian distributions, we have $D(P||\bar{Q}) = \frac{1}{2}\left(\log(n+1) - n/(n+1)\right)$ and thus

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

As we can see, the upper bound $\sqrt{2n+1} - 1$ in (49) is quite tight and $\chi^2(P||\bar{Q})$ is on the order of $\sqrt{n}$.

B.3 Proof of Corollary 17

Fix $0 < \delta \le A^{-1/2}$. Partition the entire parameter space Θ into small hypercubes each with side length δ. For each such hypercube S and let π_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 $\mathfrak{d}(X)$, clearly

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

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

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

The proof will therefore be completed if we show that

$$R_{\mathrm{Bayes}}(\pi_S, L; S) \ge \frac{1}{2} e^{-2p} 8^{-p/d} \delta^p V^{-p/d} \int_S \left(\frac{1}{r_\delta(\theta)}\right)^{p/d} \pi_S(d\theta) \tag{86}$$

for every fixed hypercube S. So let us fix S and, for notational simplicity, let $\pi := \pi_S$. We will use (39) to prove a lower bound on $R_{\mathrm{Bayes}}(\pi_S, L; S)$. Note first that

$$\begin{aligned}\inf_Q \int_S D(P_\theta||Q)\pi(\mathrm{d}\theta) &\le \int_S \int_S D(P_\theta||P_{\theta'})\pi(\mathrm{d}\theta)\pi(\mathrm{d}\theta') \\ &\le A \max_{\theta \in S, \theta' \in S} \|\theta - \theta'\|_2^2 \le Ad\delta^2 =: I_f^{\mathrm{up}}. \end{aligned} \tag{87}$$

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

$$\sup_{a \in S} \pi(B_t(a, L)) \leq \frac{f_w^{\max}}{w(S)} \mathrm{Vol}(B_t(a, L)) \leq \frac{f_w^{\max} V t^{d/p}}{f_w^{\min} \delta^d}.$$

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

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

Thus, by (87), the choice

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

leads to $\sup_{a \in S} \pi(B_t(a, L)) < \frac{1}{4} e^{-2I_f^{\mathrm{up}}}$. Employing (39), we deduce

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

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

$$\begin{aligned} R_{\mathrm{Bayes}}(\pi, L; S) &\geq \frac{1}{2} e^{-2p} \delta^p (8V)^{-p/d} \sup_{\tilde{\theta} \in S} \left(\frac{1}{r_\delta(\tilde{\theta})} \right)^{p/d} \\ &\geq \frac{1}{2} e^{-2p} \delta^p (8V)^{-p/d} \int_S \left(\frac{1}{r_\delta(\theta)} \right)^{p/d} \pi(\mathrm{d}\theta). \end{aligned}$$

This proves (86).

Appendix C. More Examples on Bayes Risk Lower Bounds

In this section, we provide more examples on the applications of derived Bayes risk lower bound in Theorem 9 and Corollary 12. For the clarity of the presentation, in each example, we will first present the Bayes risk lower bound and then provide the proof.

C.1 Generalized Linear Model

Fix $d \geq 1$ and let $\Theta = \mathcal{A} = \mathbb{R}^d$ 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 whose rows are written as $x_1^T, \ldots, x_n^T$. As in the last example, $\lambda_{\max}$ denotes the maximum eigenvalue of $X^T X/n$.

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

$$\exp\left[\frac{y\beta_i - b(\beta_i)}{a(\phi)} + c(y, \phi) \right] \quad \text{for } y \in \mathbb{R} \tag{88}$$

with $\beta_i = x_i^T\theta$ for $i = 1, \ldots, n$. The parameter ϕ 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 β where $b''(\cdot)$ is the second derivative of $b(\cdot)$. This assumption indeed holds for many generalized linear models (e.g., binomial, Gaussian) and we will discuss the case (i.e., Poisson) where this assumption fails at the end of this example.

Let w denote the Gaussian prior with mean zero and covariance matrix $\tau^2 I_d$. Using Corollary 17, we can prove that

$$R_{\text{Bayes}}(w, L; \Theta) \geq C \left[d \min\left(\frac{a(\phi)}{nK}, \tau^2 \right) \right]^{p/2} \tag{89}$$

for a constant C that depends only on p. Let us illustrate this lower bound by considering a simple case of $p = 2$. We note that the term $\frac{da(\phi)}{nK}$ is the well-known minimax risk of generalized linear model under the squared loss. The parameter τ characterizes the strength of the prior information. In fact, since $\tau^2 I$ is the variance of the Gaussian prior distribution, a small value of τ provides strong prior information that each θ_j should be concentrated around 0. When τ is large, i.e., with less prior information, the lower bound of the Bayes risk in (89) is the same as the minimax risk up to a constant factor. On the other hand, when τ is small, i.e., with strong prior information, the lower bound of the Bayes risk becomes $d\tau^2$, which is smaller than the minimax risk.

The proof of (89) will involve Corollary 17 for which we need to determine A, V and $r_\delta(\theta)$. As before, it is easy to check that $V = \text{Vol}(B)$. To determine A, fix a pair θ_1, θ_2 and, letting $\beta_i^{(j)} = x_i^T\theta_j$ for $j = 1, 2$ and $i = 1, \ldots, n$, observe that

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

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

$$\begin{aligned} D(P_{\theta_1} \| P_{\theta_2}) &\leq \frac{K}{2a(\phi)} \|\beta^{(1)} - \beta^{(2)}\|_2^2 = \frac{K}{2a(\phi)} (\theta_1 - \theta_2)^T X^T X (\theta_1 - \theta_2) \\ &\leq \frac{nK\lambda_{\max}}{2a(\phi)} \|\theta_1 - \theta_2\|^2. \end{aligned}$$

We can thus take $A = nK\lambda_{\max}/(2a(\phi))$ in Corollary 17. Next we control $r_\delta(\theta)$. For given θ and δ,

$$r_\delta(\theta) = \sup \left\{ \exp\left(-\frac{1}{2\tau^2} \left(\|\theta_1\|_2^2 - \|\theta_2\|_2^2 \right) \right) : \|\theta_i - \theta\|_2 \leq \sqrt{d}\delta \right\}.$$

For θ_1, θ_2 with $\|\theta_i - \theta\|_2 \le \sqrt{d}\delta,\ i = 1, 2$, we have

$$\begin{aligned}\left|\|\theta_1\|_2^2 - \|\theta_2\|_2^2\right| &= \left|\|\theta_1 - \theta\|_2^2 + 2\theta^T(\theta_1 - \theta) - \|\theta_2 - \theta\|_2^2 - 2\theta^T(\theta_2 - \theta)\right| \\ &\le \left|\|\theta_1 - \theta\|_2^2 - \|\theta_2 - \theta\|_2^2\right| + 2\|\theta\|_2\left(\|\theta_1 - \theta\|_2 + \|\theta_2 - \theta\|_2\right) \\ &\le d\delta^2 + 4\sqrt{d}\delta\|\theta\|_2.\end{aligned}$$

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

$$\begin{aligned}\int_\Theta \left(\frac{1}{r_\delta(\theta)}\right)^{p/d} w(\mathrm{d}\theta) &\ge \exp\left(-\frac{p\delta^2}{2\tau^2}\right)\int_\Theta \exp\left(-\frac{2p\delta}{\tau}\frac{\|\theta\|_2}{\tau\sqrt{d}}\right) w(\mathrm{d}\theta) \\ &\ge \exp\left(-\frac{p\delta^2}{2\tau^2} - \frac{4p\delta}{\tau}\right)\int_\Theta \mathbb{I}\left\{\|\theta\|_2 < 2\tau\sqrt{d}\right\} w(\mathrm{d}\theta).\end{aligned}$$

By Chebyshev's inequality, we have

$$\int_\Theta \mathbb{I}\left\{\|\theta\|_2 \ge 2\tau\sqrt{d}\right\} w(\mathrm{d}\theta) \le \frac{1}{4\tau^2 d}\int_\Theta \|\theta\|_2^2 w(\mathrm{d}\theta) = \frac{1}{4}. \tag{90}$$

Consequently,

$$\int_\Theta \left(\frac{1}{r_\delta(\theta)}\right)^{p/d} w(\mathrm{d}\theta) \ge \frac{3}{4}\exp\left(-\left(\frac{p\delta^2}{2\tau^2} + \frac{4p\delta}{\tau}\right)\right). \tag{91}$$

Corollary 17 therefore gives

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

We make the choice

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

which implies that the exponential term in the right hand side of (91) is bounded from below by $\exp(-9p/2)$. We thus have

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

The inequality (89) now follows because $V^{1/d} \asymp d^{-1/2}$.

The assumption that $b''(\beta) \le K$ which was used for the proof of (89) holds under some widely used densities of Y_i in (88). For Gaussian distribution in (88), 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 $b''(\beta) = \frac{\exp(\beta)}{(1+\exp(\beta))^2} \le \frac{1}{4}$ 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 $\widetilde{\Theta} = \{\theta \in \Theta : \|\theta\|_2 \le 2\tau\sqrt{d}\}$ and define the re-scaled prior distribution π on $\widetilde{\Theta}$ as $\pi(S) = w(S)/w(\widetilde{\Theta})$ for any measurable set $S \subseteq \widetilde{\Theta}$. Let $B = \max_{i=1,\ldots,n}\|x_i\|_2$. For any $\beta = x_i^T\theta$ for some $i = 1, \ldots, n$ and $\theta \in \widetilde{\Theta}$, we have $b''(\beta) \le \exp(2\tau\sqrt{d}B) := K$. 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 $b''(\beta) \le K$ when $\theta \in \widetilde{\Theta}$, applying the same argument, we obtain the lower bound on $R_{\mathrm{Bayes}}(\pi, L; \widetilde{\Theta})$. By (90), we have $w(\widetilde{\Theta}) \ge 3/4$ and the lower bound on $R_{\mathrm{Bayes}}(w, L; \Theta)$ can be easily established by noticing that $R_{\mathrm{Bayes}}(w, L; \Theta) \ge w(\widetilde{\Theta})R_{\mathrm{Bayes}}(\pi, L; \widetilde{\Theta}) \ge \frac{3}{4}R_{\mathrm{Bayes}}(\pi, L; \widetilde{\Theta})$.

C.2 Spiked Covariance Model

Fix $\Theta = \mathcal{A} = B$ where B is the unit Euclidean closed ball of radius one and let $L(\theta, a) := \|\theta - a\|_2^p$ for a fixed $p > 0$. Also fix $n \geq d/2$. For $\theta \in \Theta$, let P_θ denote the joint distribution of independent and identically distributed observations $X_1, \ldots, X_n$ satisfying the Gaussian distribution with zero mean and covariance matrix $\Sigma_\theta := I_d + \theta\theta^T$. 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 that

$$R_{\text{Bayes}}(w, L; \Theta) \geq C \left[\min\left(\frac{1}{2}, \frac{d}{n}\right)\right]^{p/2} \tag{92}$$

where C only depends on p.

The proof is based on the application of (32) with $f(x) = x^2 - 1$, i.e., on inequality (40).

For this, we need to bound the term $\sup_{a \in \mathcal{A}} w(B_t(a, L))$ and the f-informativity corresponding to the chi-squared divergence. It is easy to see that $\sup_{a \in \mathcal{A}} w(B_t(a, L)) \leq t^{d/p}$.

For the f-informativity, we will use the bound (51) with $\alpha = 2$ which requires bounding $M_{\chi^2}(\epsilon, \Theta)$. According to (Guntuboyina, 2011a, Theorem 4.6.1), for two Gaussian distributions with mean zero and covariance matrices Σ_1 and Σ_2 such that $2\Sigma_1^{-1} - \Sigma_2^{-1}$ is positive definite and $\|\Sigma_1 - \Sigma_2\|_F^2 \leq \frac{1}{2}\lambda_{\min}^2(\Sigma_2)$, we have

$$\chi^2\left(N_d(0, \Sigma_1) || N_d(0, \Sigma_2)\right) \leq \exp\left(\frac{\|\Sigma_1 - \Sigma_2\|_F^2}{\lambda_{\min}(\Sigma_2)^2}\right) - 1. \tag{93}$$

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 $\lambda_{\min}(\Sigma_\theta) = 1$ for all θ),

$$\chi^2\left(P_{\theta_1} || P_{\theta_2}\right) \leq \exp\left(n\|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|_F^2\right) - 1, \tag{94}$$

provided

$$2\Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1} \text{ is positive definite and } \|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|_F^2 \leq 1/2. \tag{95}$$

In the sequel, whenever we employ (94), the conditions (95) hold. But, for ease of presentation, instead of verifying (95) for every application of (94), we will simply assume (94) and verify the necessary conditions at the end of the proof. Assuming (94), we see that $\chi^2(P_{\theta_1} \| P_{\theta_2}) \leq \epsilon^2$ provided $\|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|_F^2 \leq \log(1 + \epsilon^2)/n$. Now for $\theta_1, \theta_2 \in \Theta$

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

It follows therefore that the ϵ^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

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

We now set ϵ to satisfy $\log(1+\epsilon^2) = \min(n/2, d)$ so that Corollary 16 gives

$$\begin{aligned} I_{\chi^2}(w, \mathcal{P}) &\leq M_{\chi^2}(\epsilon)(1+\epsilon^2) - 1 \\ &\leq \exp\left(\min\left(\frac{n}{2}, d\right)\right)\left[36\max\left(2, \frac{n}{d}\right)\right]^{d/2} - 1 =: I_f^{\text{up}}. \end{aligned}$$

It follows that $\sup_{a\in\mathcal{A}} w(B_t(a, L)) < \frac{1}{4}(1 + I_f^{\text{up}})^{-1}$ provided $t = (4(1 + I_f^{\text{up}}))^{-p/d}$. Inequality (40) then proves

$$R_{\text{Bayes}}(w, L; \Theta) \geq \frac{1}{2}\left(4(1 + I_f^{\text{up}})\right)^{-p/d} \geq \frac{1}{2}(24e)^{-p}\left[\min\left(\frac{1}{2}, \frac{d}{n}\right)\right]^{p/2}$$

which implies (92).

It remains to justify the conditions (95) when we used (94). It should be clear that for this, we only need to verify (95) when

$$\|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|_F^2 \leq \frac{\log(1+\epsilon^2)}{n} = \min\left(\frac{1}{2}, \frac{d}{n}\right). \tag{96}$$

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,

$$\lambda_{\min}\left(2\Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1}\right) \geq \lambda_{\min}\left(2\Sigma_{\theta_1}^{-1}\right) - \lambda_{\max}\left(\Sigma_{\theta_2}^{-1}\right) = \frac{2}{1 + \|\theta_1\|_2^2} - 1 \geq 0.$$

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 θ_2 is orthogonal to θ_1. However, when $\|\theta_1\|_2 = 1$ and θ_2 is orthogonal to θ_1, $\|\Sigma_{\theta_1} - \Sigma_{\theta_2}\|_F^2 = \|\theta_1\|_2^2 + \|\theta_2\|_2^2 > 1$, which contradicts (96). Therefore $2\Sigma_{\theta_1}^{-1} - \Sigma_{\theta_2}^{-1}$ is positive definite and this completes the proof of (92).

C.3 Gaussian Model with General Loss

In this example, we consider Gaussian location model with continuous prior with a bounded Lebesgue density and general loss functions. Here, we do not specify the form of the prior and loss. We only present this example to illustrate applications of Theorem 9 and Corollary 12. Our main bound is inequality (97). This bound however might be suboptimal for specific priors w because we do not use knowledge about the specific form of w. However, when the specific form of w is available, the argument can often be easily modified to improve inequality (97). We provide examples of this at the end of this subsection.

C.3.1 Gaussian Model with Squared Loss

Fix $d \geq 1$. Suppose $\Theta = \mathcal{A} = \mathbb{R}^d$ 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_θ denote the Gaussian distribution with mean θ and covariance matrix $\sigma^2 I_d$ ($\sigma^2 > 0$ is a constant). For every prior w on $\mathbb{R}^d$ with a Lebesgue density bounded by $W > 0$, we have

$$R_{\text{Bayes}}(w, L; \Theta) \gtrsim \frac{d\sigma^4 W^{-2/d}}{(\sigma^2 + V)^2} \tag{97}$$

where

$$V := \min_{s \in \mathbb{R}^d} \int_\Theta \frac{1}{d} \sum_{i=1}^d (\theta_i - s_i)^2 w(\mathrm{d}\theta). \tag{98}$$

To prove (97), we shall apply (32) with $f(x) = x \log x$, i.e., we apply (39). The resulting f-informativity (a.k.a mutual information) can be bounded in the following way. Because $I(w, \mathcal{P}) \leq \int D(P_\theta \| Q) w(\mathrm{d}\theta)$ for every Q. In particular, we take Q to be the Gaussian distribution with mean t and covariance matrix $(\sigma^2 + V) I_d$, where $t = \operatorname{argmin}_{s \in \mathbb{R}^d} \int_\Theta \frac{1}{d} \sum_{i=1}^d (\theta_i - s_i)^2 w(\mathrm{d}\theta)$, i.e., $t_i = \int_\Theta \theta_i w(\mathrm{d}\theta)$ is "center" of the prior. Then, we obtain

$$I(w, \mathcal{P}) \leq \int_\Theta D\left(N\left(\theta, \sigma^2 I_d\right) || N\left(t, \left(\sigma^2 + V\right) I_d\right)\right) w(\mathrm{d}\theta).$$

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

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

which by (98) implies that

$$I(w, \mathcal{P}) \leq \frac{d}{2} \log \frac{\sigma^2 + V}{\sigma^2}. \tag{99}$$

Let I_f^{up} denote the right hand side above. To apply (39), we also need an upper bound on $\sup_{a \in A} w\left(B_t(a, L)\right)$. Because of the assumption that the Lebesgue density of w is bounded from above by W, we get

$$\sup_{a \in A} w\left(B_t(a, L)\right) \leq W t^{d/2} \mathrm{Vol}(B) \tag{100}$$

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

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

for a small enough universal positive constant c, ensures $\sup_{a \in A} w\{B_t(a)\} < \frac{1}{4} e^{-2 I_f^{\text{up}}}$ (recall that I_f^{up} is the right hand side of(99)). Consequently, inequality (39) implies that $R_{\text{Bayes}} \geq t/2$. The proof of (97) is now completed using the standard fact: $\mathrm{Vol}(B)^{1/d} \asymp d^{-1/2}$.

However, since the form of the prior w is unspecified in this example, the simple upper bound on $\sup_{a \in A} w\left(B_t(a, L)\right)$ in (100) could be loose. But this can be easily fixed when

the concrete form of the prior is available. For example, for a spiked model with a large W (see an example of mixture prior in Remark 11 in the main text), the lower bound in (97) could be sub-optimal but can be easily tightened using the proposed chaining technique in Remark 11 in the main text. For another example, let w be the uniform prior on the hyper-rectangle $H = [-\epsilon, \epsilon] \times [-1, 1]^{d-1}$ for some very small ϵ. Here inequality (100) is equivalent to

$$\sup_{a \in A} w\left(B_t(a, L)\right) \leq W t^{d/2} \mathrm{Vol}(B).$$

When $\epsilon \to 0$, we have $W \to \infty$ so that the upper bound is fairly loose. However, since H is the support of w, we can also use the following upper bound:

$$\sup_{a \in A} w\left(B_t(a, L)\right) \leq W t^{d/2} \mathrm{Vol}(B \cap H).$$

When $\epsilon \to 0$, we have $W \to \infty$ but $\mathrm{Vol}(B \cap H) \to 0$. In particular, the product limit $\lim_{\epsilon \to 0} W \mathrm{Vol}(B \cap H) \to 0$ is finite. It converges to the maximum value of $w\left(B_t(a, L)\right)$ where w is restricted in a $(d-1)$-dimensional subspace of $\mathbb{R}^d$. Once we replace inequality (100) by the above upper bound, the associated Bayes risk lower bound will be tight.

C.3.2 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:

$$R_{\mathrm{Bayes}}(w, L; \Theta) \gtrsim \frac{\sigma^4 W^{-2/d}}{(\sigma^2 + V)^2} \frac{d^2}{(\mathbb{E}\|Z\|_*)^2}. \tag{101}$$

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 the previous example.

The proof of (101) is largely similar to that of (97). We use (39) along with (99) for controlling $I(w, \mathcal{P})$. To control $\sup_{a \in \mathcal{A}} w(B_t(a, L))$, we again use the fact that the Lebesgue density of w is bounded from above by W to obtain

$$\sup_{a \in A} w\left(B_t(a, L)\right) \leq W \mathrm{Vol}\left\{\theta \in \mathbb{R}^d : \|\theta\| < \sqrt{t}\right\}. \tag{102}$$

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 (2015). Urysohn's inequality gives

$$\left(\frac{\mathrm{Vol}\left\{\theta \in \mathbb{R}^d : \|\theta\| < \sqrt{t}\right\}}{\mathrm{Vol}(B)}\right)^{\frac{1}{d}} \leq \frac{\sqrt{t}}{\sqrt{d}} \mathbb{E}\|Z\|_* \quad \text{with } Z \sim N(0, I_d). \tag{103}$$

Inequalities (102) and (103) together give

$$\sup_{a \in A} w\left(B_t(a, L)\right) \leq W t^{d/2} \mathrm{Vol}(B) \left(\frac{\mathbb{E}\|Z\|_*}{\sqrt{d}}\right)^d.$$

The choice

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

for a small enough universal positive constant c ensures $\sup_{a\in A} w\{B_t(a)\} < \frac{1}{4}e^{-2I_f^{\mathrm{up}}}$ (I_f^{up} is the right hand side of (99)). The proof of (101) is then completed by noting that $\mathrm{Vol}(B)^{1/d} \asymp d^{-1/2}$.

Appendix D. Proof of Lemma 19 in Section 6

Consider the i-th instance $x_i \in \mathbb{R}^d$ sampled from $N(\theta_{z_i}; I_{d\times d})$, where z_i is the membership. Note that for any $j \in [k]\backslash\{z_i\}$, the distance between θ_{z_i} and θ_j is lower bounded by D. We have

$$\|x_i-\theta_j\|_2^2 - \|x_i-\theta_{z_i}\|_2^2 = \|\theta_j-\theta_{z_i}\|_2^2 - 2\langle\theta_j-\theta_{z_i}, x_i-\theta_{z_i}\rangle. \tag{104}$$

The random variable $\langle\theta_j-\theta_{z_i}, x_i-\theta_{z_i}\rangle$ satisfies distribution $N(0;\|\theta_j-\theta_{z_i}\|_2^2)$. Let Φ be the CDF of the standard normal distribution. Then with probability $\Phi(\frac{\|\theta_j-\theta_{z_i}\|_2-1}{2})$, we have

$$\langle\theta_j-\theta_{z_i}, x_i-\theta_{z_i}\rangle \le \|\theta_j-\theta_{z_i}\|_2\cdot\frac{\|\theta_j-\theta_{z_i}\|_2-1}{2}. \tag{105}$$

Combining (104) and (105), we have

$$\|x_i-\theta_j\|_2^2 - \|x_i-\theta_{z_i}\|_2^2 \ge \|\theta_j-\theta_{z_i}\|_2. \tag{106}$$

On the other hand, the triangular inequality implies

$$\|x_i-\theta_j\|_2 + \|x_i-\theta_{z_i}\|_2 \le 2\|x_i-\theta_{z_i}\|_2 + \|\theta_j-\theta_{z_i}\|_2. \tag{107}$$

The random variable $\|x_i-\theta_{z_i}\|_2^2$ satisfies a chi-square distribution with d degrees of freedom. It is upper bounded by βd with probability at least $1-\exp(\frac{d}{2}(1-\beta+\log\beta))$ for any $\beta>1$ (Dasgupta and Gupta, 2003). Putting (106) and (107) together, we have

$$\|x_i-\theta_j\|_2 - \|x_i-\theta_{z_i}\|_2 = \frac{\|x_i-\theta_j\|_2^2-\|x_i-\theta_{z_i}\|_2^2}{\|x_i-\theta_j\|_2+\|x_i-\theta_{z_i}\|_2} \ge \frac{\|\theta_j-\theta_{z_i}\|_2}{2\|\theta_j-\theta_{z_i}\|_2+\sqrt{\beta d}} \ge \frac{3}{6+\sqrt{\beta d}}$$

with probability at least $\Phi(\frac{D-1}{2})-\exp(\frac{d}{2}(1-\beta+\log\beta))$. By choosing $D=c\sqrt{\log(nk/\delta)}$ and $\beta = c\log(nk/\delta)/d$ for a sufficiently large constant c, this probability is lower bounded by $1-\delta/(nk)$. Applying union bound, the inequality holds for any (i,j) pair with probability at least $1-\delta$.

Appendix E. Proof of Theorem 23 in Section 7

We start with a simplified case where the random index set K is given to the estimator. Knowing this information makes the problem easier, and makes the Bayes risk lower. In addition, it reduces the d-dimensional regression problem to a k-dimensional problem where a closed-form of the Bayes risk can be derived, which establishes the following lower bound:

Claim 1 *For any $\tau > 0$, the Bayes risk is lower bounded by:*

$$R_{\mathrm{Bayes}}(w, L_{\mathrm{est}}; \Theta) \geq \frac{1}{1 + \kappa_u^2 \tau^2 n/\sigma^2} \cdot k\tau^2, \quad \textit{and} \quad R_{\mathrm{Bayes}}(w, L_{\mathrm{pre}}; \Theta) \geq \frac{1}{1 + \kappa_\ell^2 \tau^2 n/\sigma^2} \cdot \kappa_\ell^2 k\tau^2.$$

See Section E.1 for the proof.

For the rest of this proof, we establish stronger lower bounds using the fact that the index set K is unknown. It is easy to verify that for any random variable X sampled from $N(0,1)$, the probability of $|X| \geq 1/2$ is greater than $1/2$. Consider a subset of the parameter space Θ:

$$\overline{\Theta} := \Big\{\theta \in \Theta : \ \|\theta\|_2^2 \leq 2k\tau^2 \text{ and } \sum_{i=1}^{d} \mathbb{I}[|\theta_i| \geq \tau/2] \geq k/2\Big\}. \tag{108}$$

For a random vector θ sampled from the prior distribution w, the quantity $\|\theta\|_2^2/\tau^2$ satisfies a chi-square distribution with k degrees of freedom. For any $k \geq 1$, the event $\|\theta\|_2^2 \leq 2k\tau^2$ happens with probability at least 0.84. Given an index set K, for any $i \in K$ the random variable $\mathbb{I}[|\theta_i| \geq \tau/2]$ satisfies the Bernoulli distribution with parameter greater than $1/2$, so that the event $\sum_{i=1}^{d} \mathbb{I}[|\theta_i| \geq \tau/2$ happens with probability at least $1/2$. Combining these two lower bounds and applying union bound, we obtain $w(\overline{\Theta}) \geq 1/2 - (1 - 0.84) > 1/4$. As a consequence, if we define a distribution $\overline{w}$ over the subset $\overline{\Theta}$ by $\overline{w}(A) := w(A \cap \overline{\Theta})/w(\overline{\Theta})$, then Remark 11 implies that

$$R_{\mathrm{Bayes}}(w, L; \Theta) \geq w(\overline{\Theta}) \cdot R_{\mathrm{Bayes}}(\overline{w}, L; \overline{\Theta}) \geq \frac{1}{4} R_{\mathrm{Bayes}}(\overline{w}, L; \overline{\Theta}). \tag{109}$$

Hence it suffices to focus on the Bayes risk for the marginal prior $\overline{w}$.

Let the action space $\mathcal{A} := \mathbb{R}^d$ and let the loss function be either the estimation error L_{est} or the prediction error L_{pre}. In order to lower bound the Bayes risk, it suffices to bounded the chi-square informativity $I_{\chi^2}(\overline{w}, \mathcal{P})$ and the quantity $\sup_{a \in \mathcal{A}} \overline{w}(B_t(a, L))$, then applying Corollary 12. We begin with an upper bound on the chi-square informativity.

Claim 2 *For any $\tau > 0$, the chi-square informativity is bounded by:*

$$I_{\chi^2}(\overline{w}, \mathcal{P}) + 1 \leq \exp(2\kappa_u^2 \tau^2 kn/\sigma^2) \tag{110}$$

See Section E.2 for the proof.

Next, we upper bound the quantity $\sup_{a \in \mathcal{A}} \overline{w}(B_t(a, L))$. We begin by claiming a property of all Euclidean balls of small enough radius.

Claim 3 *For any point $a \in \mathbb{R}^d$, let $B(a, r)$ be the Euclidean ball of radius r centering at a. If $r \leq \frac{1}{8}\sqrt{k}\tau$, then there is a universal constant $c > 0$ such that*

$$\sup_{a \in \mathcal{A}} \overline{w}(B(a, r)) \leq \frac{c^k}{(d/k^2)^{k/4}} \Big(\frac{r}{\sqrt{k}\tau}\Big)^k.$$

See Section E.3 for the proof.

Lower bound on estimation error For the estimation error, we obtain by Claim 3 that for any $t \le \frac{1}{64}k\tau^2$, the following upper bound holds:

$$\sup_{a\in\mathcal{A}} \overline{w}(B_t(a, L_{\text{est}})) = \sup_{a\in\mathcal{A}} \overline{w}(B(a, \sqrt{t})) \le \frac{c^k}{(d/k^2)^{k/4}} \Big(\frac{\sqrt{t}}{\sqrt{k}\tau}\Big)^k. \tag{111}$$

Combining Claim 3 with inequality (111), and applying inequality (40) in Corollary 12, we obtain the lower bound:

$$R_{\text{Bayes}}(\overline{w}, L_{\text{est}}; \overline{\Theta}) \ge \frac{1}{2} \sup \left\{ 0 < t \le \frac{k\tau^2}{64} : \Big(\frac{\sqrt{t}}{\sqrt{k}\tau}\Big)^k \le \frac{(d/k^2)^{k/4}}{c^k} \cdot \frac{1}{4} \exp(-2\kappa_u^2\tau^2 kn/\sigma^2) \right\}.$$

The right-hand side is lower bounded by any scalar t satisfying:

$$t \le \frac{1}{64}k\tau^2 \quad \text{and} \quad \frac{\sqrt{t}}{\sqrt{k}\tau} \le \frac{(d/k^2)^{1/4}}{c} \cdot \frac{1}{4^{1/k}} \exp(-2\kappa_u^2\tau^2 n/\sigma^2)$$

It implies that for some universal constant $c' > 0$, we have:

$$\begin{aligned}
R_{\text{Bayes}}(\overline{w}, L_{\text{est}}; \overline{\Theta}) &\ge c'\, k\tau^2 \, \min\Big\{1, \exp(\frac{1}{2}\log(d/k^2) - 4\kappa_u^2\tau^2 n/\sigma^2)\Big\} \\
&= c'\, k\tau^2 \, \exp\Big(\min\Big\{0, \frac{1}{2}\log(d/k^2) - \frac{4\kappa_u^2\tau^2 n}{\sigma^2}\Big\}\Big) \\
&= c'\, k\tau^2 \, \exp\Big(-\frac{4\kappa_u^2 n}{\sigma^2}\Big[\tau^2 - \frac{\sigma^2\log(d/k^2)}{8\kappa_u^2 n}\Big]_+\Big) \\
&\ge c'\, k\tau^2 \, \exp\Big(-\frac{4\kappa_u^2 n}{\sigma^2}\Big[\tau^2 - \frac{\sigma^2\log(d/k)}{16\kappa_u^2 n}\Big]_+\Big),
\end{aligned} \tag{112}$$

where the last inequality uses the assumption $d > k^3$ and its implication $\log(d/k^2) > \frac{1}{2}\log(d/k)$.

Combining inequality (112) with Claim 1 yields the lower bound:

$$R_{\text{Bayes}}(w, L_{\text{est}}; \Theta) \ge c'\, k\tau^2 \max\Big\{\frac{1}{1+\kappa_u^2\tau^2 n/\sigma^2}, \exp\Big(-\frac{4\kappa_u^2 n}{\sigma^2}\Big[\tau^2 - \frac{\sigma^2\log(d/k)}{16\kappa_u^2 n}\Big]_+\Big)\Big\},$$

which completes the proof.

Lower bound on prediction error For the prediction error, we consider an arbitrary vector $a \in \mathbb{R}^d$ and an arbitrary scalar t satisfying $\sqrt{t} \le \frac{\kappa_\ell}{16}\sqrt{k}\tau$. Let θ' be the vector in $\overline{\Theta}$ which minimizes the term $\frac{1}{n}\|X(a-\theta')\|_2^2$. If the inequality $\frac{1}{n}\|X(a-\theta')\|_2^2 > t$ is true, then we have

$$\sup_{a\in\mathcal{A}} \overline{w}(B_t(a, L_{\text{pre}})) = 0. \tag{113}$$

Otherwise, we assume that $\frac{1}{n}\|X(a-\theta')\|_2^2 \le t$. Then for any vector $\theta \in \overline{\Theta}$ satisfying $\frac{1}{n}\|X(\theta-a)\|_2^2 \le t$, we have the upper bound

$$\frac{1}{n}\|X(\theta-\theta')\|_2^2 \le (n^{-1/2}\|X(\theta-a)\|_2 + n^{-1/2}\|X(a-\theta')\|_2)^2 \le (\sqrt{t}+\sqrt{t})^2 \le 4t.$$

It means that $B_t(a, L_{\rm pre}) \subseteq B_{4t}(\theta', L_{\rm pre})$. Since the vector θ' is k-sparse, the the sparse eigenvalue condition implies that for any vector $\theta \in \overline{\Theta}$, if $L_{\rm pre}(\theta, \theta') \le 4t$, then $\|\theta - \theta'\|_2 \le \frac{2\sqrt{t}}{\kappa_\ell}$, so that $B_{4t}(\theta', L_{\rm pre}) \subseteq B(\theta', \frac{2\sqrt{t}}{\kappa_\ell})$. Using Claim 3, we have

$$\sup_{a \in \mathcal{A}} \overline{w}(B_t(a, L_{\rm pre})) \le \sup_{a \in \mathcal{A}} \overline{w}(B(\theta', \frac{2\sqrt{t}}{\kappa_\ell})) \le \sup_{a \in \mathcal{A}} \overline{w}(B_t(a, L_{\rm est})) \le \frac{c^k}{(d/k^2)^{k/4}} \Big(\frac{2\sqrt{t}}{\kappa_\ell \sqrt{k}\tau} \Big)^k. \tag{114}$$

Combining equation (113) and inequality (114) we obtain

$$\sup_{a \in \mathcal{A}} \overline{w}(B_t(a, L_{\rm pre})) \le \frac{c^k}{(d/k^2)^{k/4}} \Big(\frac{2\sqrt{t}}{\kappa_\ell \sqrt{k}\tau} \Big)^k \quad \text{for any } \sqrt{t} \le \frac{\kappa_\ell}{16} \sqrt{k}\tau. \tag{115}$$

Comparing inequalities (111) and (115), we find that they differ by a factor of $(2/\kappa_\ell)^k$. Thus, following the same steps for deriving inequality (112), we can find a universal constant $c'' > 0$ such that:

$$R_{\rm Bayes}(\overline{w}, L_{\rm pre}; \overline{\Theta}) \ge c'' \, \kappa_\ell^2 k \tau^2 \, \exp\Big(- \frac{4\kappa_u^2 n}{\sigma^2} \Big[\tau^2 - \frac{\sigma^2 \log(d/k)}{16 \kappa_u^2 n} \Big]_+ \Big) \tag{116}$$

Combining inequality (116) with Claim 1 yields:

$$R_{\rm Bayes}(w, L_{\rm pre}; \Theta) \ge c'' \, \kappa_\ell^2 k \tau^2 \max\Big\{ \frac{1}{1 + \kappa_\ell^2 \tau^2 n / \sigma^2}, \exp\Big(- \frac{4\kappa_u^2 n}{\sigma^2} \Big[\tau^2 - \frac{\sigma^2 \log(d/k)}{16 \kappa_u^2 n} \Big]_+ \Big) \Big\},$$

which completes the proof.

E.1 Proof of Claim 1

The Bayes risk of the original problem is lower bounded by that of the following simplified problem: estimating θ when the index set K is known, and without loss of generality, we assume that $K = [k]$. For this case, let X' be the submatrix consisting of the first k columns of matrix X, and let θ' be the subvectors consisting of the first k coordinate of vectors θ. Given the response vector y, the posterior distribution of θ' is equal to

$$p(\theta'|y) \propto p(\theta') p(y|\theta') = N(\theta'; 0, \tau^2 I) \, N(y; X\theta', \sigma^2 I) \propto N\Big(\theta'; \Sigma^{-1}(X')^\top y, \sigma^2 \Sigma^{-1}\Big),$$

where $\Sigma := (X')^\top X' + \frac{\sigma^2}{\tau^2} I$ is a shorthand notation. As a consequence, the Bayes estimator $\widehat{\theta}$ is given by $\widehat{\theta}_K = \Sigma^{-1}(X')^\top y$ and $\widehat{\theta}_{-K} = 0$. The Bayes risk on the estimation error is lower bounded by:

$$R_{\rm Bayes}(w, L_{\rm est}; \Theta) = \mathbb{E}[\|\Sigma^{-1}(X')^\top y - \theta'\|_2^2] = \sigma^2 {\rm tr}(\Sigma^{-1}) \ge \frac{\sigma^2}{\kappa_u^2 \tau^2 n + \sigma^2} \cdot k\tau^2,$$

where the last inequality uses the sparse eigenvalue condition — it guarantees that all eigenvalues of the matrix Σ are less than or equal to $n\kappa_u^2 + \sigma^2/\tau^2$.

The Bayes estimator for minimizing the prediction error is also given by $\widehat{\theta}_K = \Sigma^{-1}(X')^\top y$ and $\widehat{\theta}_{-K} = 0$. Thus, the Bayes risk is lower bounded by:

$$\begin{aligned} R_{\mathrm{Bayes}}(w, L_{\mathrm{pre}}; \Theta) &= \frac{1}{n}\mathbb{E}[\|X'(\Sigma^{-1}(X')^\top y - \theta')\|_2^2] = \frac{\sigma^2}{n}\mathrm{tr}(X'\Sigma^{-1}(X')^\top) \\ &\geq \frac{\sigma^2}{\kappa_\ell^2\tau^2 n + \sigma^2} \cdot \kappa_\ell^2 k\tau^2, \end{aligned}$$

where the last inequality uses the sparse eigenvalue condition — it guarantees that all eigenvalues of the matrix $X'\Sigma^{-1}(X')^\top$ are greater than or equal to $\frac{n\kappa_\ell^2}{n\kappa_\ell^2 + \sigma^2/\tau^2}$.

E.2 Proof of Claim 2

Corollary 16 shows that the chi-square informativity can be bounded using the covering number $M_{\chi^2}(\epsilon, \overline{\Theta})$. Consider the zero vector $\theta_0 := 0$ and an arbitrary vector $\theta \in \overline{\Theta}$. Their response vectors are generated from $P_{\theta_0} := N(0, \sigma^2 I)$ and $P_\theta := N(X\theta, \sigma^2 I)$, so that the chi-square divergence between P_{θ_0} and P_θ is equal to $\chi^2(P_{\theta_0}\|P_\theta) = \exp(\|X\theta\|_2^2/\sigma^2) - 1$. By the sparse eigenvalue condition and the fact that $\|\theta\|_2^2 \leq 2k\tau^2$, we have

$$\chi^2(P_{\theta_0}\|P_\theta) \leq \exp(\kappa_u^2 n\|\theta\|_2^2/\sigma^2) - 1 \leq \exp(2\kappa_u^2\tau^2 kn/\sigma^2) - 1.$$

It means that if we choose $\epsilon^2 = \exp(2\kappa_u^2\tau^2 kn/\sigma^2) - 1$, then $M_{\chi^2}(\epsilon, \overline{\Theta}) = 1$, so that the chi-square informativity is bounded by

$$I_{\chi^2}(\overline{w}, \mathcal{P}) + 1 \leq (1+\epsilon^2) M_{\chi^2}(\epsilon, \overline{\Theta}) = \exp(2\kappa_u^2\tau^2 kn/\sigma^2). \tag{117}$$

E.3 Proof of Claim 3

Consider an arbitrary vector $a \in \mathbb{R}^d$, and let I_a be the set of indices defined by:

$$I_a = \{i \in [d] : \ |a_i| \geq \tau/4\}.$$

If $|I_a| > 2k$, then for any $\theta \in \overline{\Theta}$, there are at least $k+1$ coordinates such that $\theta_i = 0$ but $|a_i| \geq \tau/4$. It means that $\|a - \theta\|_2 > \frac{1}{4}\sqrt{k}\tau$. Since $r \leq \frac{1}{8}\sqrt{k}\tau$, we have $\overline{w}(B(a, r)) = 0$.

Otherwise, we assume that $|I_a| \leq 2k$. Given an index set K, let w_K and $\overline{w}_K$ be the conditional version of the prior distribution w and $\overline{w}$, conditioning on the fact that the k-sparse index set is K. Recall that for any θ in the support of $\overline{w}_K$, there are at least $k/2$ coordinates such that $|\theta_i| \geq \tau/2$. If $|I_a \cap K| < k/4$, then there at least $k/4$ coordinates such that $|\theta_i| \geq \tau/2$ but $|a_i| < \tau/4$. It means that $\|a - \theta\|_2 > \frac{1}{8}\sqrt{k}\tau$ for any θ in the support of $\overline{w}_K$, and as a consequence, we have $\overline{w}_K(B(a, r)) = 0$.

Thus, a necessary condition for $\overline{w}_K(B(a, r)) > 0$ to hold is $|I_a \cap K| \geq k/4$. Given $|I_a| \leq 2k$, the number of index set K satisfying this constraint is bounded by $\binom{2k}{k/4}\binom{d-k/4}{3k/4}$. To prove this bound, notice that every set K satisfying $|I_a \cap K| \geq k/4$ can be generated by the following two-step procedure: first, generate $k/4$ element from I_a; second, generate the remaining $3k/4$ elements from the remaining $d - k/4$ integers of $\{1, \ldots, d\}$. There are totally $\binom{2k}{k/4}\binom{d-k/4}{3k/4}$ ways of generating the set. We note that the same K can have multiple ways to generate, so that the above combinatorial number is a strict upper bound on the number of sets.

For any set K satisfying the above constraint, we have:

$$\overline{w}_K(B(a,r)) \leq 4w_K(B(a,r)) \leq 4w_K(B(0,r)), \tag{118}$$

where the last equation holds because w_K represents an isotropic normal distribution in $\mathbb{R}^k$, so that the maximum probability is achieved by centering at the origin. The right-hand side of inequality (118) the probability a k-dimension normal random variable $X \sim N(0, \tau^2 I_{k\times k})$ satisfying $\|X\|_2 \leq r$. As we showed in the proof of Lemma 21, this probability is bounded by $(\frac{cr}{\sqrt{k}\tau})^k$ for a universal constant $c > 0$. Putting pieces together, we have

$$\sup_{a\in\mathcal{A}} \overline{w}(B(a,r)) \leq \frac{\binom{2k}{k/4}\binom{d-k/4}{3k/4}}{\binom{d}{k}} \cdot 4\Big(\frac{c\,r}{\sqrt{k}\tau}\Big)^k.$$

By the definition of the combinatorial numbers, we have:

$$\begin{aligned}\frac{\binom{2k}{k/4}\binom{d-k/4}{3k/4}}{\binom{d}{k}} &= \binom{2k}{k/4}\frac{k!(d-k/4)!}{(3k/4)!d!} \leq \binom{2k}{k/4}\frac{k!}{(3k/4)!}\frac{1}{d^{k/4}} \\ &\leq \frac{(2k)^{k/2}}{d^{k/4}} = \Big(\frac{d}{4k^2}\Big)^{-k/4}.\end{aligned}$$

Combining the two upper bounds above completes the proof.

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.

S. Arora and R. Kannan. Learning mixtures of separated nonspherical Gaussians. *The Annals of Applied Probability*, 15(1A):69–92, 2005.

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

K. B. Athreya and S. N. Lahiri. *Measure Theory and Probability Theory*. Springer, 2006.

C. Banderier, R. Beier, and K. Mehlhorn. Smoothed analysis of three combinatorial problems. In *Mathematical Foundations of Computer Science 2003*, pages 198–207. Springer, 2003.

J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory*, 51(4):1611–1615, 2005.

A. Blum and J. Dunagan. Smoothed analysis of the perceptron algorithm for linear programming. In *Proceedings of the ACM-SIAM symposium on Discrete algorithms (SODA)*, 2002.

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.

E. J. Candes, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8): 1207–1223, 2006.

I. Castillo. Lower bounds for posterior rates with gaussian process priors. *Electronic Journal of Statistics*, 2:1281–1299, 2008.

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

S. Dasgupta. Learning mixtures of gaussians. In *Proceedings of the Symposium on Foundations of Computer Science (FOCS)*, 1999.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

S. Dasgupta and L. J. Schulman. A two-round variant of EM for gaussian mixtures. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2000.

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

J. Dunagan, D. A Spielman, and S. H. Teng. Smoothed analysis of condition numbers and complexity implications for linear programming. *Mathematical Programming*, 126 (2):315–350, 2011.

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

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.

D. Garcıa-Garcıa and R. C. Williamson. Divergences and risks for multiclass experiments. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2012.

R. Ge, Q. Q. Huang, and S. M Kakade. Learning mixtures of Gaussians in high dimensions. *arXiv preprint arXiv:1503.00424*, 2015.

R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 03 1995.

O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Majorizing measures and proportional subsets of bounded orthonormal systems. *Revista matemática iberoamericana*, 24(3):1075–1095, 2008.

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. Gneralizing 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, 1997.

D. Hsu and S. M. Kakade. Learning mixtures of spherical Gaussians: moment methods and spectral decompositions. In *Proceedings of the Conference on Innovations in Theoretical Computer Science*, 2013.

F. Liese. Phi-divergences, sufficiency, Bayes sufficiency, and deficiency. *Kybernetika*, 48(4): 690–713, 2012.

Z. Ma and Y. Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *IEEE Trans. Inform. Theory*, 61(12):6939–6956, 2015.

B. Manthey and R. Reischuk. Smoothed analysis of binary search trees. *Theoretical Computer Science*, 378(3):292–315, 2007.

A. Rakhlin, K. Sridharan, and A. B. Tsybakov. Empirical entropy, minimax regret and minimax risk. *arXiv preprint arXiv:1308.1147*, 2013.

G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13 (1):389–427, 2012.

M. D. Reid and R. C. Williamson. Generalised Pinsker inequalities. *arXiv preprint arXiv:0906.1244*, 2009.

M. D. Reid and R. C. Williamson. Information, divergence and risk for binary experiments. *Journal of Machine Learning Research*, 12:731–817, 2011.

H. Röglin and B. Vöcking. Smoothed analysis of integer programming. *Mathematical programming*, 110(1):21–56, 2007.

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

D. A. Spielman and S. H. Teng. Smoothed analysis. In *Algorithms and data structures*, pages 256–270. Springer, 2003.

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 voulme of convex boides in n-dimensional space. *Mat. Sbornik*, 31:477–486, 1924.

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

S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

B. Vidakovi 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.

A. Xu and M. Raginsky. A new information-theoretic lower bound for distributed function computation. In *Proceedings of IEEE International Symposium on Information Theory*, 2014.

Y. H. Yang. Minimax nonparametric classification. I. rates of convergence. *IEEE Trans. Inform. Theory*, 45(7):2271–2284, 1999.

Y. H. 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.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans. Inform. Theory*, 52(4):1307–1321, 2006.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Proceedings of the Conference on Learning Theory*, 2014.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Optimal prediction for sparse linear models? lower bounds for coordinate-separable M-estimators. *arXiv preprint arXiv:1503.03188*, 2015.

Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102): 1–44, 2016.