New Information Inequalities with Applications to Statistics

Kuan-Yun Lee

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-37
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-37.html

May 4, 2022

New Information Inequalities with Applications to Statistics

by

Kuan-Yun Lee

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Thomas Courtade, Chair
Professor Adityanand Guntuboyina
Professor Jiantao Jiao
Professor Kannan Ramchandran

Spring 2022

New Information Inequalities with Applications to Statistics

Abstract

New Information Inequalities with Applications to Statistics

by

Kuan-Yun Lee

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Thomas Courtade, Chair

We introduce, under a parametric framework, a family of inequalities between mutual information and Fisher information. These inequalities are indexed by reference measures satisfying a log-Sobolev inequality (LSI), and reveal previously unknown connections between LSIs and statistical inequalities. One such connection is shown for the celebrated van Trees inequality by recovering under a Gaussian reference measure a stronger entropic inequality due to Efroimovich. We further present two new inequalities for log-concave priors that do not depend on the Fisher information of the prior and are applicable under certain scenarios where the van Trees inequality and Efroimovich's inequality cannot be applied. We illustrate a procedure to establish lower bounds on risk under general loss functions, and apply it under several statistical settings, including the Generalized Linear Model and a general pairwise comparison framework.

To my family.

Contents

Acknowledgments

The dissertation could not have been completed without the support of my advisor, Professor Thomas Courtade. For one, Tom accepted me into his group many years ago when I had little academic accomplishments nor experience. Tom allowed me to independently explore and dive deep into uncharted areas, and provided invaluable guidance that led to this dissertation. For another, Tom has been supportive through all the ups and downs. I am very grateful for all the support and guidance I have received from Tom throughout the five years at Berkeley.

I would like to thank the rest of the committee members, who have been supportive since the early months of the pandemic: Professor Adityanand Guntuboyina, Professor Jiantao Jiao, and Professor Kannan Ramchandran. Their mentorship and insightful comments have helped shape the dissertation tremendously.

Thanks to my undergraduate advisor Professor I-Hsiang Wang, who taught my first lessons in information theory and guided me through several projects many years back. Without his guidance, I would probably never have stepped into this field.

Thanks to co-authors Professor Ashwin Pananjady and Efe Aras for bearing with many long discussions and "short-loops" that eventually built the foundations of this dissertation.

To the dozens of friends in BLISS, BWRC, ICSI, SWARM and others throughout the school and across the globe — thank you for the wonderful grad school memories. I will never forget the afternoon boba runs that we often had and the fun memories of attending pre-covid social hours. I will miss the many occasions where we chatted about life.

Finally, big thanks to my family for being supportive as always.

Introduction

Many important inference tasks in science and engineering can be modeled parametrically. In such models, there is an underlying **parameter** θ, and an **observation** (or, measurement) modeled by a random variable X, with distribution indexed by the parameter. A Bayesian setting may also be adopted, where the parameter θ is itself a random variable with known **prior distribution**. Based on the observation X, the basic task is to infer θ by designing an **estimator** (a Borel measurable function) $\hat{\theta} : X \mapsto \hat{\theta}(X)$ that minimizes the expected risk

$$\hat{\theta} \mapsto \mathbb{E}[\mathcal{L}(\hat{\theta}, \theta)],$$

where $\mathcal{L}$ is a penalty function suitable for the application at hand.

Despite its abstract formulation, this basic parametric model captures the spirit of many interesting questions encountered in practice. For example, in finance, one may be interested in estimating the volatility of a stochastic process; in this case, θ is the volatility of the process, and X may be the observed time series. In signal processing, one may be interested in recovering a latent signal from noisy measurements; here, θ corresponds to the latent signal, and X corresponds to the measurements. In communications, θ may be one of a discrete set of messages sent across a noisy channel, and X is the observation at the receiver.

In the classical setting of ℓ_2 loss (i.e., $\mathcal{L}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$), the classical Cramér–Rao bound and its Bayesian counterpart — the van Trees inequality — can be applied to lower bound the expected loss. The Cramér–Rao bound is applicable to unbiased estimators (i.e., $\mathbb{E}\hat{\theta} = \theta$), and the van Trees inequality applies when the parameter admits a prior distribution with sufficiently smooth density. The utility of the Cramér–Rao bound and the related van Trees inequality is, in large part, due to the fact that they depend only on basic Fisher information quantities associated to the parametric model, and often give good bounds on ℓ_2 loss. Nevertheless, it remains a problem of significant interest and practical importance to establish lower bounds on expected risk when the loss function is not squared loss, and in situations where the Cramér–Rao bound and the van Trees inequality do not otherwise apply. This is one primary motivation for this thesis.

Our first development is a family of upper bounds on mutual information $I(\pi; P_\theta)$, with $\theta \sim \pi$ and $X \sim P_\theta$, to be covered in Chapter 2. Upper bounds on $I(\pi; P_\theta)$ have natural applications in information theory and statistics. When combined with tools in rate-distortion theory, upper bounds on $I(\pi; P_\theta)$ can give *lower bounds* on expected risk, and therefore understanding bounds within this family can lead to potential advances in techniques to lower

bound risk. While there is a vast volume of literature related to lower bounds on risk using Fano's method and metric entropy, there is comparably less focus from a rate distortion perspective despite well known results regarding the Shannon lower bound [Shannon, 1959]. One reason is that these results are often expressed in terms of $I(\pi; P_\theta)$, which is difficult to control outside of certain special cases. On the contrary, our new upper bounds on $I(\pi; P_\theta)$ can be easily integrated with the Shannon lower bound, yielding interesting applications to lower bounds on risk as we will demonstrate later.

Our family of upper bounds on $I(\pi; P_\theta)$ are indexed by *reference measures* satisfying an LSI. In other words, choices of reference measure affect the corresponding upper bound. Our next development is an analysis of the special case where the reference measure is Gaussian. Interestingly, we recover an inequality due to Efroimovich [1979], which we will be referring to as **Efroimovich's inequality**. Efroimovich's inequality is an entropic generalization of the van Trees inequality, and therefore, both the van Trees inequality and Efroimovich's inequality are inequalities within our family of upper bounds on $I(\pi; P_\theta)$. This unveils a previously unknown relation between the LSI and the van Trees inequality.

It is worth noting that Efroimovich's inequality, and consequently the van Trees inequality, is not applicable when the Fisher information $\mathcal{J}(\pi)$ of the prior π is not well-behaved. Towards a development of bounds that are applicable under these constraints, we present in Chapter 3 two new upper bounds on $I(\pi; P_\theta)$ given log-concave π. These new bounds do not depend on $\mathcal{J}(\pi)$ and, therefore, can be applied in situations where $\mathcal{J}(\pi)$ is undesirable.

A general procedure to apply our upper bounds on $I(\pi; P_\theta)$ to lower bounds on risk using rate distortion theory will be discussed in Chapter 4. Applications to the Generalized Linear Model, including an array of Bayes risk and minimax risk lower bounds, will be discussed in Chapter 5. We also provide an application to a generalized pairwise comparison framework, for which we derive lower bounds on minimax risk that hold uniformly over the broad class of models we consider. Details are available in Chapter 6. We conclude with our outlook on future directions and final remarks in Chapter 7.

A summary of key contributions is presented below.

1. We introduce a family of inequalities indexed by reference measures satisfying an LSI. We expose interesting relations between LSIs and quantities of statistical interest, namely mutual information and Fisher information.

2. We show that Efroimovich's inequality is a special case within our family of inequalities, indexed by a Gaussian reference measure. This illuminates a previously unknown connection between the Gaussian LSI and the celebrated van Trees inequality, by nature of it being a corollary of Efroimovich's inequality.

3. We present two new upper bounds on $I(\pi; P_\theta)$, provided the prior π on $\theta \sim \pi$ is log-concave. These inequalities do not depend on the Fisher information of the prior, and therefore can be applied to certain scenarios where Efroimovich's inequality and the van Trees inequality fail.

4. We provide a general procedure to transform our upper bounds on $I(\pi; P_\theta)$ into lower bounds on risk using tools from rate distortion theory.

5. We establish lower bounds on risk for the Generalized Linear Model and a general pairwise comparison framework. In many cases our bounds have improved topological dependence compared to previously known results.

Parts of the dissertation are based on [Aras et al., 2019, Lee and Courtade, 2020, Lee and Courtade, 2020, Lee and Courtade, 2021].

Chapter 1

Preliminaries

In this chapter, we fix basic notation and assumptions that will be used throughout the dissertation. In particular, we precisely formulate what is meant by a parametric model, and introduce the standard regularity assumption that we adopt throughout. We define both Shannon and Fisher information quantities that will be needed frequently in our development. Such quantities include (Shannon) entropy, mutual information, relative entropy, Fisher information matrices, and relative Fisher information. The quantities of (Shannon) relative entropy and relative Fisher information are connected through logarithmic Sobolev inequalities, which are also briefly reviewed in the last section of this chapter.

As a matter of convention, terms that are being defined are written in bold font throughout this dissertation.

1.1 The basic parametric model

Consistent with the notation introduced in the introduction, we adopt a parametric model with parameter $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ and observation X taking values in some set $\mathcal{X}$. The model is defined by a family of dominated probability measures $(P_\theta)_{\theta \in \mathbb{R}^d}$ on a measurable space $(\mathcal{X}, \mathcal{F})$, with dominating σ-finite measure λ. For each P_θ, we associate a density $f(\cdot;\theta)$ with respect to λ according to

$$dP_\theta(\cdot) = f(\cdot;\theta) d\lambda(\cdot). \tag{1.1}$$

We shall *always* make the following regularity assumption of $(P_\theta)_{\theta \in \mathbb{R}^d}$, and will not explicitly refer to it in statements of results:

Assumption 1. *The function $\theta \mapsto f(x;\theta)$ is differentiable for λ-a.e. $x \in \mathcal{X}$, and f is sufficiently regular to permit the following exchange of integration and differentiation:*

$$\int_{\mathcal{X}} \nabla_\theta f(x;\theta) d\lambda(x) = 0, \quad \forall \theta \in \mathbb{R}^d. \tag{1.2}$$

Here, ∇_θ denotes the gradient with respect to θ.

In other words, the orders of differentiation with respect to θ and integration with respect to x can be exchanged via the Leibniz rule. This assumption is common in the context of the Cramér–Rao bound, and is implied by continuity of Fisher information with respect to θ; see, e.g., Cramér [1999], Gill and Levit [1995]. One immediate consequence of the regularity assumption is that $f : (x, \theta) \mapsto f(x; \theta)$ is a Carathéodory function, and is therefore jointly measurable in its arguments [Aliprantis and Border, 2006, Lemma 4.51]. In particular, this ensures $\theta \mapsto P_\theta(F)$ is (Borel) measurable for every $F \in \mathcal{F}$ by Tonelli's theorem.

We let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$ equipped with the Borel σ-algebra $\mathcal{B}$, and let $\pi \in \mathcal{P}(\mathbb{R}^d)$ denote the prior distribution of θ. Given θ, the observation has distribution $X \sim P_\theta$. In this way, π and $(P_\theta)_{\theta \in \mathbb{R}^d}$ induce a joint probability distribution $\mathbb{P}$ on $(\mathbb{R}^d \times \mathcal{X}, \mathcal{B} \times \mathcal{F})$, defined by

$$\Pr\{\theta \in B, X \in F\} \equiv \mathbb{P}(B \times F) = \int_B P_\theta(F) d\pi(\theta), \quad \forall B \in \mathcal{B}, \forall F \in \mathcal{F}. \tag{1.3}$$

Example 1 (Gaussian location model)**.** *Take $\mathcal{X} = \mathbb{R}^d$, λ equal to Lebesgue measure, and*

$$f(x; \theta) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\theta)^T \Sigma^{-1} (x-\theta)}.$$

This describes the Gaussian location model $X = \theta + Z$, where $Z \sim \mathcal{N}(0, \Sigma)$ is independent of $\theta \sim \pi$.

Remark 1. *In the example above, and throughout this dissertation, we let $\mathrm{I_d}$ denote the $d \times d$ identity matrix, and $|\cdot|$ denote the Euclidean metric on $\mathbb{R}^d$.*

Example 2 (Poisson model)**.** *Take $\mathcal{X} = \{0, 1, \ldots\}$, λ equal to the counting measure, and*

$$f(x; \theta) = \frac{\theta^x e^{-\theta}}{x!}.$$

This describes the Poisson model $X \sim Poisson(\theta)$.

Example 3 (Exponential model)**.** *Take $\mathcal{X} = \mathbb{R}^+$, λ equal to Lebesgue measure, and*

$$f(x; \theta) = \theta e^{-\theta x}.$$

This describes the exponential model with rate θ.

1.2 Shannon information quantities

Mutual information and Shannon entropies

Given that the joint distribution of θ and X is well-defined, we define their **mutual information**

$$I(\pi; P_\theta) \equiv I(\theta; X) := \int_{\mathbb{R}^d} \int_{\mathcal{X}} f(x; \theta) \log \frac{f(x; \theta)}{\int_{\mathbb{R}^d} f(x; \theta') d\pi(\theta')} d\lambda(x) d\pi(\theta). \tag{1.4}$$

We shall use the notations $I(\pi; P_\theta)$ and $I(\theta; X)$ interchangeably, depending on which is most convenient in the given context. In particular, the former is typically more convenient when a given parametric model is described in terms of probability distributions π and $(P_\theta)_{\theta \in \mathbb{R}^d}$; the latter is most convenient when describing the model in terms of the random variables θ and X. Throughout this dissertation, logarithms are understood to be taken with respect to the natural base.

When π admits a density $d\pi = \psi d\theta$ with respect to Lebesgue measure, we define its **Shannon entropy**

$$h(\pi) \equiv h(\theta) := -\int_{\mathbb{R}^d} \psi(\theta) \log \psi(\theta) d\theta,$$

provided the integral exists in the Lebesgue sense. A sufficient condition for the entropy $h(\theta)$ to exist is that θ has finite second moments [Shannon, 1948]. When the entropy $h(\theta)$ exists, we define the (Shannon) conditional entropy

$$h(\theta|X) := h(\theta) - I(\theta; X), \tag{1.5}$$

provided the indeterminate form $\infty - \infty$ is avoided.

Let us give some examples of Shannon entropy that will be useful to us in later chapters. The first is the entropy of a uniform distribution.

Example 4 (Entropy of a uniform distribution)**.** *Consider $\pi = \text{Uniform}(a, b)$ with $a < b$, the uniform measure on the set (a, b). Then,*

$$h(\pi) \equiv h(\theta) = \log(b - a).$$

More generally, in higher dimensions, it is often of practical interest to consider the uniform measure over a convex set, a reason being that the uniform measure is the maximum-entropy density over a constrained convex set. The Shannon entropy for the uniform measure on a convex set is determined by the volume of the convex set.

Example 5 (Entropy of a uniform distribution in higher dimensions)**.** *Consider π to be the uniform distribution over a convex set $\Theta \subset \mathbb{R}^d$ with volume $V > 0$. Then,*

$$h(\pi) \equiv h(\theta) = \log V.$$

As another example, let us examine the entropy of the Gaussian distribution, which has practical use in many settings and is the maximum entropy distribution under fixed second moments [Shannon, 1948, Cover and Thomas, 2012]. The entropy of a Gaussian is determined by its covariance.

Example 6 (Entropy of a Gaussian distribution)**.** *Consider $\pi = \mathcal{N}(0, \Sigma)$ with $\Sigma \in \mathbb{R}^{d \times d}$. Then,*

$$h(\pi) \equiv h(\theta) = \frac{1}{2} \log \det (2\pi e \Sigma).$$

From the entropy of the Gaussian one can infer an identity for the mutual information of a Gaussian channel with Gaussian input.

Example 7 (Mutual information of a Gaussian channel with Gaussian input)**.** *Suppose the observation $X = \theta + Z \sim P_\theta$, where $Z \sim \mathcal{N}(0, \Delta)$ is Gaussian noise independent from θ. Suppose $\theta \sim \pi = \mathcal{N}(0, \Sigma)$. Then,*

$$I(\pi; P_\theta) \equiv I(\theta; X) = \frac{1}{2} \log \det(\mathrm{I_d} + \Sigma\Delta^{-1}).$$

This is also the capacity of the Gaussian channel [Cover and Thomas, 2012] under power (covariance) constraint Σ.

Relative entropy

Finally, if $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, with $\nu \ll \mu$ and $d\nu = h d\mu$, define the **relative entropy** as

$$D(\nu \| \mu) := \int_{\mathbb{R}^d} h \log(h)\, d\mu.$$

The relative entropy always exists as a number in $[0, +\infty]$, and we adopt the convention that $D(\nu\|\mu) = +\infty$ if $\nu \not\ll \mu$.

Remark 2. *Relative entropy is also commonly known as the Kullback-Leibler (KL) divergence.*

Example 8 (Relative entropy between two Gaussians)**.** *Suppose $\mu = \mathcal{N}(a, \Sigma_1)$ and $\nu = \mathcal{N}(b, \Sigma_2)$ with $a, b \in \mathbb{R}^d$ and $\Sigma_1, \Sigma_2 \in \mathbb{R}^{d\times d}$. Then,*

$$D(\nu\|\mu) = \frac{1}{2} \log\det\left(\Sigma_1^{-1}\Sigma_2\right) + \frac{d}{2} \mathrm{Tr}\left(\Sigma_1\Sigma_2^{-1}\right) + \frac{1}{2}(b-a)^T\Sigma_2^{-1}(b-a) - \frac{d}{2}.$$

In the case with $\Sigma_1 = \Sigma_2 = \Sigma$,

$$D(\nu\|\mu) = \frac{1}{2}(b-a)^T\Sigma^{-1}(b-a).$$

These definitions of mutual information, Shannon (conditional) entropy, and relative entropy are standard [Cover and Thomas, 2012].

1.3 Fisher information

Parametric Fisher information

For the parametric family $(P_\theta)_{\theta\in\mathbb{R}^d}$, we define the **Fisher information matrix** $\bar{\mathcal{I}}_X$ as the matrix-valued function

$$[\bar{\mathcal{I}}_X(\theta)]_{ij} := \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta_i} f(x;\theta) \frac{\partial}{\partial\theta_j} f(x;\theta)}{f(x;\theta)} d\lambda(x), \qquad 1 \le i, j \le d,\ \theta \in \mathbb{R}^d. \tag{1.6}$$

We define the **Fisher information** $\mathcal{I}_X$ as the trace of this matrix, namely

$$\mathcal{I}_X(\theta) := \int_{\mathcal{X}} \frac{|\nabla_\theta f(x;\theta)|^2}{f(x;\theta)} d\lambda(x) = \mathrm{Tr}\left(\bar{\mathcal{I}}_X(\theta)\right), \quad \theta \in \mathbb{R}^d. \tag{1.7}$$

Let us examine the Fisher information in several statistical models that we will be referring to in later chapters.

Example 9 (Fisher information matrix of a Gaussian observation)**.** *Suppose* $X \sim \mathcal{N}(\theta, \Sigma)$ *with* $\theta \in \mathbb{R}^d$ *and* $\Sigma \in \mathbb{R}^{d\times d}$*. Then,*

$$\begin{aligned}
\bar{\mathcal{I}}_X(\theta) &= \int_{\mathbb{R}^d} \frac{\nabla_\theta f(x;\theta) \nabla_\theta f(x;\theta)^T}{f(x;\theta)} dx \\
&= \Sigma^{-1} \left(\int_{\mathbb{R}^d} (x-\theta)(x-\theta)^T f(x;\theta) dx \right) \Sigma^{-1} \\
&= \Sigma^{-1}.
\end{aligned}$$

Example 10 (Fisher information of an exponential observation)**.** *Let* $X \sim \mathrm{Exponential}(\theta)$*. Then,* $f(x;\theta) = \theta e^{-\theta x}$ *and*

$$\begin{aligned}
\mathcal{I}_X(\theta) &= \int_0^\infty \frac{(1-x\theta)^2}{\theta} e^{-\theta x} \\
&= \int_0^\infty \left(x - \frac{1}{\theta}\right)^2 \theta e^{-\theta x} dx \\
&= \frac{1}{\theta^2}.
\end{aligned}$$

Note that the observation X need not have a continuous density. The Fisher information is well-defined for discrete observation models as well, such as the Bernoulli observation.

Example 11 (Fisher information of a Bernoulli observation)**.** *Suppose* $X \sim \mathrm{Bernoulli}(\theta)$*. Then,* $f(x;\theta) = \theta^X (1-\theta)^{1-X}$ *and*

$$\mathcal{I}_X(\theta) = \frac{1}{1-\theta} + \frac{1}{\theta} = \frac{1}{\theta(1-\theta)}.$$

While the above examples conveniently have Fisher information matrix $\bar{\mathcal{I}}_X$ equal to the inverse of the covariance matrix of X (given θ), this is not universally true. The Cauchy observation model is one such example.

Example 12 (Fisher information of a Cauchy observation)**.** *Suppose* $X \sim \mathrm{Cauchy}(0, \theta)$*. Then,*

$$f(x;\theta) = \frac{1}{\pi\theta\left(1 + \frac{x^2}{\theta^2}\right)},$$

and the Fisher information is calculated as

$$\mathcal{I}_X(\theta) = \frac{1}{2\theta^2}.$$

The variance of X, on the other hand, is undefined.

One of the most important properties of Fisher information is that it is additive on conditionally independent observations.

Proposition 1 (Additivity of Fisher information)**.** *Let $X = (X_1, X_2)$, where X_1 and X_2 are conditionally independent given θ. Then,*

$$\bar{\mathcal{I}}_X(\theta) = \bar{\mathcal{I}}_{X_1}(\theta) + \bar{\mathcal{I}}_{X_2}(\theta).$$

A convenient application of Proposition 1 is in the calculation of the Fisher information of n independent samples.

Example 13 (Fisher information of a binomial observation)**.** *Suppose $X = (X_1, \ldots, X_n) \sim$* Binomial$(n, \theta)$*. Then, since $X_1, \ldots, X_n$ are sampled i.i.d. from* Bernoulli(θ)*, it follows that*

$$\mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta) = \frac{n}{\theta(1-\theta)}.$$

The reader is likely familiar with the classical Cramér–Rao bound, which we state below for convenience. The proof boils down to an application of the Cauchy–Schwarz inequalty; see, for example, the classical results [Cramér, 1946, Rao, 1945, Fréchet, 1943, Darmois, 1945]. When combined with the additivity property above, it gives control on sample complexity for unbiased estimators given independent observations.

Theorem 1 (Cramér–Rao Bound)**.** *Fix a parametric model. If $\hat{\theta} : \mathcal{X} \to \mathbb{R}^d$ is* **unbiased** *in the sense that $\theta = \int_{\mathcal{X}} \hat{\theta}(x) dP_\theta(x)$ for each $\theta \in \mathbb{R}^d$, then*

$$\mathbb{E}_{P_\theta}(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T \geq \bar{\mathcal{I}}_X(\theta)^{-1}, \tag{1.8}$$

where $\mathbb{E}_{P_\theta}$ denotes expectation with respect to P_θ.

A canonical example of an unbiased estimator that satisfies the Cramér–Rao bound with equality is the sample mean estimator under the Gaussian observation model.

Example 14 (Sample mean estimator under Gaussian observations)**.** *Suppose $X = (X_1, \ldots, X_n)$ are sampled i.i.d. from $P_\theta = \mathcal{N}(\theta, \Sigma)$. The Fisher information matrix is*

$$\bar{\mathcal{I}}_X(\theta) = n\Sigma^{-1},$$

while the sample mean estimator $\hat{\theta}(X) := \frac{1}{n}\sum_{i=1}^n X_i$ *satisfies*

$$\begin{aligned}\mathbb{E}_{P_\theta^{\otimes n}}\left(\hat{\theta}-\theta\right)\left(\hat{\theta}-\theta\right)^T &= \mathbb{E}_{P_\theta^{\otimes n}}\left(\frac{1}{n}\sum_{i=1}^n X_i-\theta\right)\left(\frac{1}{n}\sum_{i=1}^n X_i-\theta\right)^T \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{P_\theta}\left(X_i-\theta\right)\left(X_i-\theta\right)^T \\ &= \frac{1}{n}\Sigma,\end{aligned}$$

satisfying the Cramér–Rao bound with equality.

We remark that the map $\bar{\mathcal{I}}_X$ is a function of the parametrization $\theta \mapsto P_\theta$. This is implied by the subscript X, which is understood to represent the observation model. This notation is convenient, for instance, in stating the additivity property of Fisher information. Sometimes, however, it is necessary to make the dependence on the parametrization $\theta \mapsto P_\theta$ explicit, in which case we write $\bar{\mathcal{I}}_{X;\theta}$ to denote the Fisher information of the observation X with the parametrization $\theta \mapsto X \sim P_\theta$. This long form notation is important when we reparametrize the model, as in the following proposition, which amounts to a simple change of variables in (1.6).

Proposition 2 (Reparametrization of Fisher information)**.** *Let* $V \in \mathbb{R}^{d\times d}$ *be an invertible matrix, and define the parameter* $\eta(\theta) = V\theta$*. In this case, Fisher information enjoys the matrix congruence*

$$\bar{\mathcal{I}}_{X;\eta}(\eta(\theta)) = V^{-T}\bar{\mathcal{I}}_{X;\theta}(\theta)V^{-1}.$$

Remark 3. *If Assumption 1 holds for parametrization in terms of* θ*, then it holds for all linear (or, more generally, affine) reparametrizations as above.*

Frequently, we will consider the expectation of Fisher information under a given prior. Therefore, for a given parameterization $\theta \mapsto P_\theta$ and prior distribution $\theta \sim \pi \in \mathcal{R}^d$, we write

$$\mathbb{E}_\pi[\bar{\mathcal{I}}_X] \equiv \mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}] := \int_{\mathbb{R}^d} \bar{\mathcal{I}}_X(\theta')d\pi(\theta').$$

Fisher information associated to a prior

When the prior π admits a density ψ, we associate to it the **(information theorists') Fisher information matrix**, or **Fisher information matrix of the prior**

$$[\bar{\mathcal{J}}(\pi)]_{ij} := \int_{\mathbb{R}^d} \frac{\frac{\partial}{\partial\theta_i}\psi(\theta)\frac{\partial}{\partial\theta_j}\psi(\theta)}{\psi(\theta)}d\theta, \qquad 1 \le i,j \le d,\ \theta \in \mathbb{R}^d. \tag{1.9}$$

Similar to before, we define the trace of this matrix to be the **(information theorists') Fisher information**, or **Fisher information of the prior**

$$\mathcal{J}(\pi) \equiv \mathcal{J}(\psi) := \mathrm{Tr}(\bar{\mathcal{J}}(\pi)) = \int_{\mathbb{R}^d} \frac{|\nabla_\theta \psi(\theta)|^2}{\psi(\theta)} d\theta. \tag{1.10}$$

Note that finiteness of the Fisher information boils down to the function $\nabla\sqrt{\psi}$ being square-integrable, which imposes a relatively strong degree of smoothness on ψ. Indeed, by monotone convergence, the following identity can be justified, which explains how (1.10) should be interpreted whenever ψ does not have full support:

$$\mathcal{J}(\pi) \equiv \mathcal{J}(\psi) = \int_{\mathbb{R}^d} \frac{|\nabla_\theta \psi(\theta)|^2}{\psi(\theta)} d\theta = 4 \int_{\mathbb{R}^d} |\nabla_\theta \sqrt{\psi(\theta)}|^2 d\theta.$$

When π does not admit a density, or if $\sqrt{\psi}$ is not differentiable, we adopt the convention $\mathcal{J}(\pi) = +\infty$. Not all priors of practical importance have finite Fisher information; a canonical example is a uniform distribution on a convex set in $\mathbb{R}^d$.

Concerning the two different Fisher informations $\mathcal{I}_X$ and $\mathcal{J}(\pi)$, we emphasize that $\mathcal{I}_X$ depends only on the parametric family $(P_\theta)_{\theta \in \mathbb{R}^d}$, and $\mathcal{J}(\pi)$ depends only on the prior distribution π. The two definitions coincide when θ is a location parameter.

Example 15 (Fisher information for location parameter)**.** *Let $\mathcal{X} = \mathbb{R}^d$ and λ be Lebesgue measure. If there is a density $g : \mathbb{R}^d \to \mathbb{R}$ such that*

$$f(x;\theta) = g(x-\theta), \quad x, \theta \in \mathbb{R}^d,$$

then

$$\mathcal{I}_X(\theta) = \int_{\mathcal{X}} \frac{|\nabla_\theta f(x;\theta)|^2}{f(x;\theta)} dx = \int_{\mathcal{X}} \frac{|\nabla g(x)|^2}{g(x)} dx = \mathcal{J}(g), \quad \forall \theta \in \mathbb{R}^d.$$

Relative Fisher information

Finally, if $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, with $\nu \ll \mu$ and $d\nu = h d\mu$, define the **relative Fisher information** as

$$I(\nu\|\mu) := \int_{\mathbb{R}^d} \frac{|\nabla h|^2}{h} d\mu,$$

when the density h is differentiable. When h is not differentiable or if $\nu \not\ll \mu$, we adopt the convention that $I(\nu\|\mu) = +\infty$. Thus, like relative entropy, the relative Fisher information is always well-defined as a number in $[0, +\infty]$. Both play a role in defining logarithmic Sobolev inequalities, which we come to next.

1.4 Logarithmic Sobolev inequalities

Logarithmic Sobolev inequalities provide one way to link entropy and Fisher information. A probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$ is said to satisfy a **log-Sobolev inequality** with constant C if

$$D(\nu\|\mu) \le \frac{C}{2} I(\nu\|\mu), \quad \forall \nu \in \mathcal{P}(\mathbb{R}^d). \tag{1.11}$$

We abbreviate this by saying μ satisfies $\mathrm{LSI}(C)$. We define the **LSI constant** for μ to be the best such constant C; that is, with a slight abuse of notation, we define

$$\mathrm{LSI}(\mu) := \inf\{C \ge 0 \ : \ \mu \text{ satisfies } \mathrm{LSI}(C)\}.$$

The classical example of a measure that satisfies an LSI is the **standard Gaussian measure**

$$d\gamma(x) := \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}|x|^2}, \quad x \in \mathbb{R}^d,$$

for which $\mathrm{LSI}(\gamma) = 1$. LSI constants are well-behaved under various transformations, so oftentimes one begins with a measure known to satisfy an LSI, such as γ, and then perturbs it to obtain a new measure in a way that retains control on the LSI constant. We list a few examples below.

Proposition 3. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfy $\mathrm{LSI}(C)$.*

(i) (Tensorization) If $\nu \in \mathcal{P}(\mathbb{R}^{d'})$ satisfies $\mathrm{LSI}(C')$, then the product measure $\mu \times \nu$ on $\mathbb{R}^d \times \mathbb{R}^{d'}$ satisfies $\mathrm{LSI}(C \vee C')$.

(ii) If $T : \mathbb{R}^d \to \mathbb{R}^d$ is L-Lipschitz, then the pushforward $T\#\mu$ satisfies $\mathrm{LSI}(L^2 C)$.

(iii) ([Holley and Stroock, 1988]) For a function $\vartheta : \mathbb{R}^d \to \mathbb{R}$, define its **oscillation** *as*

$$\mathrm{osc}(\vartheta) := \sup \vartheta - \inf \vartheta.$$

If $\tilde{\mu} \in \mathcal{P}(\mathbb{R}^d)$ is defined by the perturbation $d\tilde{\mu} = e^{-\vartheta} d\mu$, then $\tilde{\mu}$ satisfies $\mathrm{LSI}(e^{\mathrm{osc}(\vartheta)} C)$.

(iv) ([Chen et al., 2021]) If $\nu \in \mathcal{P}(\mathbb{R}^d)$ is supported in a Euclidean ball of radius R and $\gamma_t = \mathcal{N}(0, t)$, then the convolution $\nu * \gamma_t$ satisfies $\mathrm{LSI}(C)$ for $C \le 6(4R^2 + t)e^{4R^2/t}$.

The study of logarithmic Sobolev inequalities has blossomed in recent decades, and was initiated by Gross [1975], who was the first to state the Gaussian LSI and consider its applications. However, Stam had actually discovered an equivalent form of the Gaussian LSI a decade earlier in his proof of the entropy power inequality [Stam, 1959]; the equivalence was noted by Carlen [1991]. Roughly speaking, LSIs encode a strong form of measure concentration, which can be equivalently formulated as rapid convergence of an associated diffusion process to equilibrium, or as hypercontractivity of the associated Markov semigroup. These interpretations are not of significant importance to us here, since our applications will be purely at the level of the functional inequality itself.

Chapter 2

Cramér–Rao-type bounds and log-Sobolev inequalities

2.1 A family of Bayesian Cramér–Rao-type bounds

Arguably the most common application of LSIs in statistics has been to establish concentration inequalities; see, e.g., [Raginsky and Sason, 2013, Boucheron et al., 2013, Ledoux, 2001, McDiarmid, 1998]. Indeed, the LSI constant $\mathrm{LSI}(\mu)$ encodes strong concentration properties of a probability measure $\mu \in \mathcal{P}(\mathbb{R}^d)$. More precisely, an argument due to Herbst[1] gives the following: if $F : \mathbb{R}^d \to \mathbb{R}$ has Lipschitz constant $\|F\|_{\mathrm{Lip}} \leq L$, then

$$\mu\left\{F \geq \int F d\mu + t\right\} \leq \exp\left(-\frac{t^2}{2\,\mathrm{LSI}(\mu)L^2}\right), \qquad \forall t > 0.$$

Here, we introduce a new perspective, which reveals a connection between LSIs and information quantities of significance in parametric statistics, namely, we connect the mutual information $I(\pi; P_\theta)$ to the Fisher information $\mathcal{I}_X$. The following is the main result of this Chapter.

Theorem 2. *Fix a parametric model $(P_\theta)_{\theta\in\mathbb{R}^d}$ and prior $\pi \in \mathcal{P}(\mathbb{R}^d)$. For any $\mu \in \mathcal{P}(\mathbb{R}^d)$,*

$$I(\pi; P_\theta) + D(\pi\|\mu) \leq \frac{\mathrm{LSI}(\mu)}{2}\left(I(\pi\|\mu) + \mathbb{E}_\pi[\mathcal{I}_X]\right). \tag{2.1}$$

Remark 4. *Observe that, for a fixed $\mu \in \mathcal{P}(\mathbb{R}^d)$, the LSI (1.11) can be recovered from Theorem 2 by considering a parametric model where X is independent of θ. In this case, $I(\pi; P_\theta) = \mathbb{E}_\pi[\mathcal{I}_X] = 0$.*

We emphasize here an important philosophical difference between (2.1) and (1.11). In (2.1), it is the parametric model that is fixed, and we may optimize over all reference measures

[1] The original argument of Herbst was given in a personal communication letter written in response to Gross [1975]; see, e.g., Gross and Rothaus [1998] for a brief mention of the history.

μ to obtain a suitable upper bound on $I(\pi; P_\theta)$, which can then be used to, for example, establish lower bounds on Bayes risk of estimating θ. In contrast, typical application of (1.11) consider μ as a given, fixed probability measure. Following the proof, we shall interpret (2.1) as a family of Cramér–Rao-type bounds, indexed by measures μ satisfying an LSI.

Proof of Theorem 2. We may assume that the RHS of equation (2.1) is finite; else the claim is trivially true. In particular, this means that we can restrict attention to $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying LSI(C) such that $\pi \ll \mu$. With this in mind, suppose $d\pi = h d\mu$, and note that $h(\theta) f(x;\theta)$ is the joint density of the random vector (θ, X) with respect to $\mu \times \lambda$. Let us define the marginal density of X as

$$f(x) := \int_{\mathbb{R}^d} f(x;\theta) d\pi(\theta) \quad x \in \mathcal{X},$$

and we can write the conditional density of θ given $\{X = x\}$ by

$$h_x(\theta) = \frac{h(\theta) f(x;\theta)}{f(x)} \quad \theta \in \mathbb{R}^d,\ x \in \mathcal{X}. \tag{2.2}$$

The functions $f(\cdot)$ and $h_\cdot(\cdot)$ are well-defined $(\pi \times \lambda)$-a.e. Now, since μ satisfies LSI(C), we have for λ-a.e. x,

$$\int_{\mathbb{R}^d} h_x(\theta) \log h_x(\theta) d\mu(\theta) \le \frac{C}{2} \int_{\mathbb{R}^d} \frac{|\nabla_\theta h_x(\theta)|^2}{h_x(\theta)} d\mu(\theta),$$

which follows by definition of the LSI (1.11). Integrating both sides with respect to $f d\lambda$, we have

$$\int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^d} h_x(\theta) \log h_x(\theta) d\mu(\theta) \right) d\lambda(x) \tag{2.3}$$

$$\le \frac{C}{2} \int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^d} \frac{|\nabla_\theta h_x(\theta)|^2}{h_x(\theta)} d\mu(\theta) \right) d\lambda(x). \tag{2.4}$$

Let us start by working on (2.4). Observe that

$$\begin{aligned}
&\int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^d} \frac{|\nabla_\theta h_x(\theta)|^2}{h_x(\theta)} d\mu(\theta) \right) d\lambda(x) \\
&= \int_{\mathcal{X}} \int_{\mathbb{R}^d} \frac{|\nabla_\theta (f(x) h_x(\theta))|^2}{f(x) h_x(\theta)} d\mu(\theta) d\lambda(x) \\
&= \int_{\mathcal{X}} \int_{\mathbb{R}^d} \frac{|\nabla_\theta (f(x;\theta) h(\theta))|^2}{f(x;\theta) h(\theta)} d\mu(\theta) d\lambda(x) \\
&= \int_{\mathcal{X}} \int_{\mathbb{R}^d} \left(f(x;\theta) \frac{|\nabla_\theta h(\theta)|^2}{h(\theta)} + 2 \nabla_\theta h(\theta) \cdot \nabla_\theta f(x;\theta) + h(\theta) \frac{|\nabla_\theta f(x;\theta)|^2}{f(x;\theta)} \right) d\mu(\theta) d\lambda(x) \\
&= I(\pi \| \mu) + \int_{\mathbb{R}^d} \mathcal{I}_X(\theta) d\pi(\theta) + 2 \int_{\mathcal{X}} \int_{\mathbb{R}^d} \nabla_\theta h(\theta) \cdot \nabla_\theta f(x;\theta) d\mu(\theta) d\lambda(x),
\end{aligned}$$

where the penultimate identity follows by the product rule for derivatives and expanding the square. The final cross term is integrable; indeed, Cauchy-Schwarz yields

$$\begin{aligned}
&\int_{\mathcal{X}}\int_{\mathbb{R}^d} |\nabla h(\theta)\cdot\nabla f(x;\theta)| d\mu(\theta)d\lambda(x)\\
&\le \sum_{i=1}^{d}\int_{\mathcal{X}}\int_{\mathbb{R}^d} |\frac{\partial}{\partial\theta_i}h(\theta)\frac{\partial}{\partial\theta_i}f(x;\theta)d\mu(\theta)|d\lambda(x)\\
&\le \sum_{i=1}^{d}\left(\int_{\mathcal{X}}\int_{\mathbb{R}^d}\frac{|\frac{\partial}{\partial\theta_i}h(\theta)|^2}{h(\theta)}f(x;\theta)d\mu(\theta)d\lambda(x)\right)^{1/2}\left(\int_{\mathcal{X}}\int_{\mathbb{R}^d}\frac{|\frac{\partial}{\partial\theta_i}f(x;\theta)|^2}{f(x;\theta)}h(\theta)d\mu(\theta)d\lambda(x)\right)^{1/2}\\
&\le \sqrt{I(\pi\|\mu)\int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta)}.
\end{aligned}$$

The last term is finite by our assumption that the RHS of (2.1) is finite. The exchange of integrals to obtain the last line is justified by Tonelli's theorem. Therefore, by Fubini's theorem,

$$\int_{\mathcal{X}}\int_{\mathbb{R}^d}\nabla h(\theta)\cdot\nabla f(x;\theta)d\mu(\theta)d\lambda(x) = \int_{\mathbb{R}^d}\nabla h(\theta)\cdot\left(\int_{\mathcal{X}}\nabla f(x;\theta)d\lambda(x)\right)d\mu(\theta) = 0,$$

where the last equality follows by the prevailing Assumption 1. Summarizing, we have

$$\int_{\mathcal{X}} f(x)\left(\int_{\mathbb{R}^d}\frac{|\nabla h_x(\theta)|^2}{h_x(\theta)}d\mu(\theta)\right)d\lambda(x) = I(\pi\|\mu) + \int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta). \tag{2.5}$$

Next, let us work on (2.3), for which we write

$$\begin{aligned}
&\int_{\mathcal{X}} f(x)\left(\int_{\mathbb{R}^d} h_x(\theta)\log h_x(\theta)d\mu(\theta)\right)d\lambda(x)\\
&= \int_{\mathcal{X}}\int_{\mathbb{R}^d} f(x)h_x(\theta)\log h_x(\theta)d\mu(\theta)d\lambda(x)\\
&= \int_{\mathcal{X}}\int_{\mathbb{R}^d} f(x;\theta)h(\theta)\log h_x(\theta)d\mu(\theta)d\lambda(x)\\
&= \int_{\mathcal{X}}\int_{\mathbb{R}^d} f(x;\theta)h(\theta)\log\frac{h_x(\theta)}{h(\theta)}d\mu(\theta)d\lambda(x) + \int_{\mathcal{X}}\int_{\mathbb{R}^d} f(x;\theta)h(\theta)\log h(\theta)d\mu(\theta)d\lambda(x)\\
&= \int_{\mathcal{X}}\int_{\mathbb{R}^d} f(x;\theta)h(\theta)\log\frac{f(x;\theta)}{f(x)}d\mu(\theta)d\lambda(x) + \int_{\mathbb{R}^d} h(\theta)\log h(\theta)d\mu(\theta)\\
&= I(\pi;P_\theta) + D(\pi\|\mu).
\end{aligned} \tag{2.6}$$

Finally, we can combine (2.5) and (2.6) to yield

$$I(\pi;P_\theta) + D(\pi\|\mu) \le \frac{C}{2}\left(I(\pi\|\mu) + \int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta)\right).$$

Letting C approach LSI(μ) from above concludes the proof of Theorem 2. □

2.2 The Efroimovich and van Trees inequalities

As a first application of Theorem 2, we consider the special case where μ is taken to be equal to the standard Gaussian measure γ in (2.1). By particularizing in this way, we obtain the following.

Corollary 3 (Efroimovich's inequality)**.** *Fix a parametric model $(P_\theta)_{\theta\in\mathbb{R}^d}$ and prior $\pi \in \mathcal{P}(\mathbb{R}^d)$ with finite second moments and finite entropy. It holds that*

$$\frac{1}{2\pi e} e^{\frac{2}{d}h(\theta|X)} \geq \frac{1}{\det\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}}, \tag{2.7}$$

Proof. Fix an invertible matrix $V \in \mathbb{R}^{d\times d}$, and consider the reparametrization $\eta = V\theta$. If we let $\pi_V = V\#\pi$ denote the law of η, then taking $\mu = \gamma$ in (2.1) gives

$$I(\theta; X) + D(\pi_V\|\gamma) = I(\eta; X) + D(\pi_V\|\gamma) \leq \frac{1}{2}\left(I(\pi_V\|\gamma) + \mathbb{E}_{\pi_V}[\mathcal{I}_{X;\eta}]\right), \tag{2.8}$$

where we used the fact that mutual information is invariant to invertible transformations of its arguments. Since π has finite second moments, so does π_V, which justifies the expansions

$$D(\pi_V\|\gamma) = -h(\eta) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\int_{\mathbb{R}^d} |\eta|^2 d\pi_V(\eta)$$

and

$$I(\pi_V\|\gamma) = \mathcal{J}(\pi_V) + \int_{\mathbb{R}^d} |\eta|^2 d\pi_V(\eta) - 2d.$$

On substitution into (2.8), we obtain

$$\begin{aligned}
&2I(\theta; X) + d\log(2\pi e)\\
&\leq 2h(V\theta) + \mathcal{J}(\pi_V) + \int_{\mathbb{R}^d} \mathcal{I}_X(\eta) d\pi_V(\eta) - d\\
&= 2h(\theta) + \log\det(V^TV) + \mathrm{Tr}\left(V^{-T}\left(\bar{\mathcal{J}}(\pi) + \int_{\mathbb{R}^d} \bar{\mathcal{I}}_X(\theta) d\pi(\theta)\right)V^{-1}\right) - d\\
&= 2h(\theta) + \log\det(V^TV) + \mathrm{Tr}\left((V^TV)^{-1}\left(\bar{\mathcal{J}}(\pi) + \int_{\mathbb{R}^d} \bar{\mathcal{I}}_X(\theta) d\pi(\theta)\right)\right) - d.
\end{aligned}$$

Observe that the left-hand side of this inequality does not depend on V, so we may optimize over our choice of V on the right-hand side. This task is made easy by recalling the Legendre duality for $\log\det$ on the cone of positive semidefinite $d \times d$ matrices, which can be written as:

$$d + \log\det(A) = \inf_{B>0}\{\mathrm{Tr}(AB^{-1}) + \log\det(B)\}.$$

Hence, we conclude

$$2I(\theta; X) + d\log(2\pi e) \leq 2h(\theta) + \log\det\left(\bar{\mathcal{J}}(\pi) + \int_{\mathbb{R}^d} \bar{\mathcal{I}}_X(\theta) d\pi(\theta)\right)$$

which is a rearrangement of what we aimed to prove. □

Inequality (2.7) was first discovered by Efroimovich [1979] using a different method of proof. The point to be made here is that Efroimovich's inequality follows as a special case of Theorem 2. In turn, a special case of Efroimovich's inequality is the **van Trees inequality**, which asserts that, for any estimator $\hat{\theta}$,

$$\mathbb{E}|\hat{\theta} - \theta|^2 \geq \frac{d}{\det\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}}. \tag{2.9}$$

This follows from translation invariance of Shannon entropy, together with the classical fact that entropy is maximized by a Gaussian distribution subject to a covariance constraint. In particular, if we let Σ denote the covariance of the error $(\hat{\theta} - \theta)$, then

$$\frac{2}{d}h(\theta|X) \leq \frac{2}{d}h(\hat{\theta} - \theta) \leq \log\left((2\pi e)\det(\Sigma)^{1/d}\right) \leq \log\left((2\pi e)\frac{1}{d}\mathbb{E}|\hat{\theta} - \theta|^2\right),$$

where the final step is the AM-GM inequality. This together with (2.7) immediately gives (2.9). The van Trees inequality was originally stated in 1968 by van Trees [1968], and it was subsequently popularized by Gill and Levit [1995], who interpreted it as a **Bayesian Cramér–Rao bound**. Since then, the van Trees inequality has enjoyed widespread use in the statistics literature. Despite the fact that Efroimovich's inequality is a strengthening of the van Trees inequality, it seems to be largely overlooked since it was never translated from its original Russian-language publication [Efroimovich, 1979]. To give one anecdote, Tsybakov [2008] notes several advantages of the van Trees inequality, such as its relative simplicity in application and ability to establish sharp bounds, but states a "*limitation is that the van Trees inequality applies only to the squared loss function*". Efroimovich's inequality stands in contradiction to the latter statement, in the sense that the entropic strengthening it affords over the ℓ_2 van Trees inequality can be leveraged to establish lower bounds under any loss function, via rate distortion theory.

Remark 5. *One advantage that the van Trees inequality provides over the classical Cramér–Rao bound (3.2) is that the van Trees inequality applies to any estimator $\hat{\theta}$, without restriction to unbiased estimators. Of course, it is well-known that the classical Cramér–Rao bound has extensions to biased estimators, but these bounds are not universal in the sense that they depend on fine-grained information about the bias of the estimator (e.g., the bias and its derivatives with respect to the parameter). See, for instance, [Scharf and Demeure, 1991] for a discussion.*

Asymptotic relation to the Cramér–Rao bound

Let us briefly explain how the van Trees (or, Efroimovich's) inequality is connected to the classical Cramér–Rao bound, which is based on a description given by Gill and Levit [1995]. We'll work in dimension $d = 1$ for convenience. Suppose $X := (X_1, X_2, \ldots, X_n)$ consists of

n i.i.d. samples from a density $g(\cdot;\theta)$ defined on $\mathbb{R}$. In other words,

$$f(x;\theta) = \prod_{i=1}^{n} g(x_i;\theta) \quad x \in \mathbb{R}^n.$$

Again, we consider θ having a prior π. Suppose that the expected Fisher information of $g(\cdot;\theta)$ is finite, and let us write it as $\mathbb{E}_\pi[\mathcal{I}_{X_1}]$; recall that $X_1, \ldots, X_n$ are i.i.d. and they therefore have the same expected Fisher information. Then, by additivity of Fisher information, we have

$$\mathbb{E}_\pi[\mathcal{I}_X] = n\mathbb{E}_\pi[\mathcal{I}_{X_1}]. \tag{2.10}$$

Here, we used the fact that $\mathbb{E}\log g(x_i;\theta) = 0$. With (2.10), one can write from (2.9)

$$\mathbb{E}(\hat{\theta}_n - \theta)^2 \geq \frac{1}{\mathcal{J}(\pi) + n\mathbb{E}_\pi[\mathcal{I}_{X_1}]},$$

where $\hat{\theta}_n$ is any estimator depending on the samples $X_1, \ldots, X_n$.

Assuming the prior π has finite Fisher information, we may take $n \to \infty$ to obtain the asymptotic bound

$$\liminf_{n\to\infty} \mathbb{E}\left(\sqrt{n}(\hat{\theta}_n - \theta)\right)^2 \geq \frac{1}{\mathbb{E}_\pi[\mathcal{I}_{X_1}]}.$$

The above can be related to the Cramér–Rao bound by choosing the prior π as a density (satisfying the smoothness assumptions) located in $(\theta_0 - n^{-1/2}t, \theta_0 + n^{-1/2}t)$ for some fixed point θ_0; in this case, first taking $n \to \infty$ and then taking $t \to \infty$ recovers

$$\liminf_{n\to\infty,\, t\to\infty} \mathbb{E}\left(\sqrt{n}(\hat{\theta}_n - \theta)\right)^2 \geq \frac{1}{\mathcal{I}_{X_1}(\theta_0)},$$

which is known as the asymptotic information bound [Gill and Levit, 1995], and recovers asymptotically a Cramér–Rao-type bound for any sequence of estimators $(\hat{\theta}_n)_{n\geq 1}$. We remark that, unlike the classical non-asymptotic Cramér–Rao bound, no assumption of unbiasedness is required.

The van Trees inequality can be applied to establish minimax risk under non-parametric models, oftentimes via guessing a difficult parametric sub-problem for which the van Trees inequality can be applied [Gill and Levit, 1995].

2.3 Remarks on Efroimovich's inequality

In this section, we briefly remark on different forms of Efroimovich's inequality, which are shown to be equivalent via a simple rescaling argument. The motivation is to provide a prototype for similar rescaling arguments to appear later in this dissertation.

We observe that, as a consequence of the AM-GM inequality applied to the eigenvalues of the Fisher information matrices $\bar{\mathcal{J}}(\pi)$ and $\mathbb{E}[\bar{\mathcal{I}}_X]$, Efroimovich's inequality implies

$$\frac{1}{2\pi e} e^{\frac{2}{d}h(\theta|X)} \geq \frac{d}{\mathcal{J}(\pi) + \mathbb{E}_\pi[\mathcal{I}_X]}. \tag{2.11}$$

Although (2.11) is evidently weaker than (2.9), they are formally equivalent. Indeed, by considering the reparametrization $\eta = V\theta$ for an invertible $V \in \mathbb{R}^{d\times d}$, it follows from (2.11) that

$$\begin{aligned}\frac{1}{2\pi e} e^{-\frac{2}{d}I(\theta;X)} = \frac{1}{2\pi e} e^{-\frac{2}{d}I(\eta;X)} &\geq \frac{d e^{-\frac{2}{d}h(\eta)}}{\operatorname{Tr}\left(V^{-T}\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)V^{-1}\right)} \\ &= \frac{d \det(V^TV)^{-1/d} e^{-\frac{2}{d}h(\theta)}}{\operatorname{Tr}\left((V^TV)^{-1}\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\right)}.\end{aligned}$$

Choosing V such that $V^TV = \left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right) + \epsilon \mathrm{I}_d$ and letting $\epsilon \downarrow 0$ recovers (2.9). It is always possible to choose such a V, since the Fisher information term is a positive semidefinite matrix, and the perturbation ensures strict positive definiteness. Thus, when comparing Efroimovich's inequality against later results, it will sometimes be more convenient to compare against (2.11) in the linearized form

$$I(\pi; P_\theta) \leq \frac{d}{2}\log\left(\frac{1}{2\pi e}\right) + h(\theta) + \frac{d}{2}\log\left(\frac{1}{d}\left(\mathcal{J}(\pi) + \mathbb{E}_\pi[\mathcal{I}_X]\right)\right), \tag{2.12}$$

which comes without any loss of generality.

The formal equivalence between (2.7) and (2.12) is of the same spirit as the known equivalence between the Gaussian LSI and the so-called Stam inequality [Carlen, 1991]. Indeed, if X is independent of θ, then we recover from (2.11) the uncertainty principle due to Stam [1959]

$$\frac{1}{2\pi e} e^{\frac{2}{d}h(\pi)} \geq \frac{d}{\mathcal{J}(\pi)}, \tag{2.13}$$

which coincides with an entropic improvement of the classical Cramér–Rao bound in the case where $\theta \sim \pi$ is a location parameter. Using the inequality $\log(x) \leq x - 1$, it is an easy exercise to see that Stam's inequality (2.13) implies the Gaussian log-Sobolev inequality

$$D(\pi\|\gamma) \leq \frac{1}{2} I(\pi\|\gamma). \tag{2.14}$$

However, optimizing over dilations of π in the LSI (2.14) returns us to (2.13).

Chapter 3

Cramér–Rao-type inequalities for log-concave priors

3.1 Motivation

Chapter 2 established a general interpretation of LSIs in the context of parametric models. In particular, by optimizing over measures μ in (2.1), we obtain an abstract method for bounding the mutual information $I(\pi; P_\theta)$. This provides an alternative to the traditional Fano method, which generally dispenses with all information about the prior π when the capacity of the channel $\theta \mapsto X \sim P_\theta$ is invoked. As one example of its utility, we took μ equal to the standard Gaussian distribution in (2.1) to obtain Efroimovich's inequality, which is itself a strengthening of the van Trees inequality. The former can be equivalently rewritten as

$$e^{-\frac{2}{d}I(\theta;X)} \geq \frac{\exp\left(\frac{2}{d}\mathcal{D}_g(\pi)\right)}{\det\left(\Sigma_\pi^{1/2}\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\Sigma_\pi^{1/2}\right)^{1/d}} \tag{3.1}$$

when Σ_π is nonsingular; this can be assumed without any essential loss of generality, since deterministic components of $\theta \sim \pi$ can be disregarded. Here, $\mathcal{D}_g(\pi)$ denotes the "non-Gaussianness" of π, defined as the relative entropy between π and the Gaussian distribution with the same mean and covariance; i.e.,

$$\mathcal{D}_g(\pi) := D(\pi \| \mathcal{N}(\mathbb{E}[\theta], \Sigma_\pi)).$$

The advantage of Efroimovich's inequality (3.1) is that it provides a simple upper bound on the mutual information $I(\pi; P_\theta)$ in terms of standard Fisher information quantities. A significant drawback is that the bound is degenerate whenever $\mathcal{J}(\pi) = +\infty$, which is the case for non-smooth priors, such as uniform distributions on compact sets. Hence, it is natural to ask: *to what extent can Efroimovich's inequality be improved to circumvent this difficulty?*

To begin to answer this, let us look at the RHS of (3.1). To this end, considering a parametric model with observation X independent of the parameter $\theta \sim \pi$ in (2.7) gives the

Stam-type uncertainty principle

$$e^{\frac{2}{d}h(\theta)} \det(\bar{\mathcal{J}}(\pi))^{1/d} \geq 2\pi e \quad \Rightarrow \quad \exp\left(\frac{2}{d}\mathcal{D}_g(\pi)\right) \leq \det(\Sigma_\pi)^{1/d} \det(\bar{\mathcal{J}}(\pi))^{1/d}.$$

Likewise, the classical Cramér–Rao bound for a location parameter (in which the observation noise has distribution π) gives the related matrix inequality

$$\Sigma_\pi^{1/2} \bar{\mathcal{J}}(\pi) \Sigma_\pi^{1/2} \geq \mathrm{I}_d. \tag{3.2}$$

If we assume the prior is well-behaved in some sense, then we might expect approximate equality in each of these two bounds (note: neither of these inequalities depend on the family $(P_\theta)_{\theta \in \mathbb{R}^d}$). Following this logic, substitution into the RHS of (3.1) would give

$$\begin{aligned}
e^{-\frac{2}{d}I(\theta;X)} &\geq \frac{\exp\left(\frac{2}{d}\mathcal{D}_g(\pi)\right)}{\det\left(\Sigma_\pi^{1/2}\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\Sigma_\pi^{1/2}\right)^{1/d}} \\
&\approx \frac{\det(\Sigma_\pi)^{1/d}\det(\bar{\mathcal{J}}(\pi))^{1/d}}{\det\left(\Sigma_\pi^{1/2}\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\Sigma_\pi^{1/2}\right)^{1/d}} \\
&= \frac{\det(\Sigma_\pi^{1/2}\bar{\mathcal{J}}(\pi)\Sigma_\pi^{1/2})^{1/d}}{\det\left(\Sigma_\pi^{1/2}\left(\bar{\mathcal{J}}(\pi) + \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\Sigma_\pi^{1/2}\right)^{1/d}} \\
&\approx \frac{1}{\det\left(\mathrm{I}_d + \Sigma_\pi^{1/2}\mathbb{E}_\pi[\bar{\mathcal{I}}_X]\Sigma_\pi^{1/2}\right)^{1/d}}.
\end{aligned}$$

This suggests we might hope to eliminate dependence on the Fisher information $\mathcal{J}(\pi)$ in Efroimovich's inequality (3.1) by establishing practically relevant and easily verifiable conditions on the prior π under which we can expect

$$e^{-\frac{2}{d}I(\theta;X)} \gtrsim \frac{1}{\det\left(\mathrm{I}_d + \Sigma_\pi \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}}. \tag{3.3}$$

Indeed, this coincides with Efroimovich's inequality in the special case of a Gaussian prior π, which should be regarded as a prototypical example.

As far as practically relevant and easily verifiable conditions on the prior go, log-concavity is a first condition that comes to mind. For log-concave distributions π, it is conjectured that $\frac{1}{d}\mathcal{D}_g(\pi) \leq C$ for some universal constant C. In fact, this is equivalent to the hyperplane conjecture due to Ball [1988], as shown by Bobkov and Madiman [2010]. Of note, it is well-known that the hyperplane conjecture is implied by the KLS conjecture, which states that the isoperimetric constant of any isotropic log-concave distribution is lower bounded by a universal constant. Recent research supports the validity of the (stronger) KLS conjecture, in

the sense that Chen [2020] and more recently Klartag and Lehec [2022] established an "almost constant" lower bound (i.e., a lower bound that decays subpolynomially in dimension). Thus, assuming the hyperplane (or KLS) conjecture to be true, the sequence of approximations above can be turned into a sequence of quantitative inequalities for log-concave distributions. Namely,

$$\begin{aligned}
&\frac{\exp\left(\frac{2}{d}\mathcal{D}_g(\pi)\right)}{\det\left(\Sigma_\pi^{1/2}\left(\bar{\mathcal{J}}(\pi)+\mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\Sigma_\pi^{1/2}\right)^{1/d}} \\
&\leq \frac{C}{\det\left(\Sigma_\pi^{1/2}\left(\bar{\mathcal{J}}(\pi)+\mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)\Sigma_\pi^{1/2}\right)^{1/d}} \\
&\leq \frac{C}{\det\left(\mathrm{I}_d+\Sigma_\pi\mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}}.
\end{aligned}$$

Therefore, if the hyperplane (or KLS) conjecture is true, inequality (3.3) would represent a uniform improvement — up to absolute constants — of Efroimovich's inequality (3.1) on the class of log-concave distributions, which are immensely important in practice. We summarize with a proposition.

Proposition 4. *Fix a parametric model $(P_\theta)_{\theta\in\mathbb{R}^d}$ and a log-concave prior $\pi\in\mathcal{P}(\mathbb{R}^d)$. If the hyperplane conjecture is true, then an estimate of the form*

$$e^{-\frac{2}{d}I(\theta;X)} \gtrsim \frac{1}{\det\left(\mathrm{I_d}+\Sigma_\pi\mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}} \tag{3.4}$$

represents a uniform improvement of Efroimovich's inequality, up to absolute constants.

Remark 6. *The key point is that, unlike Efroimovich's inequality, (3.4) does not suffer from degeneracy when $\mathcal{J}(\pi)=+\infty$.*

3.2 Main result

With the context of the previous section in mind, our goal this Chapter is to show that (3.3) holds for all log-concave priors π. Of note, this includes non-smooth priors, such as uniform distributions on convex sets, where Efroimovich's inequality (3.1) would be degenerate.

Before stating our main result, let us make precise the definition of a log-concave distribution. For a positive semidefinite matrix $K\in\mathbb{R}^{d\times d}$, we say that μ is **K-uniformly log-concave** if there exists a convex function $V:\mathbb{R}^d\to\mathbb{R}$ such that $d\mu(x)=e^{-V(x)-\frac{1}{2}x^TKx}dx$. In case $K=0$, we simply say that μ is **log-concave**.

As promised, the main result of this Chapter is the following.

Theorem 4. *Fix a parametric model $(P_\theta)_{\theta\in\mathbb{R}^d}$ and a K-uniformly log-concave prior $\pi \in \mathcal{P}(\mathbb{R}^d)$. It holds that*

$$e^{-\frac{2}{d}I(\pi;P_\theta)} \geq \frac{e^{-2\alpha(\pi,P_\theta)}}{\det\left(\mathrm{I_d} + \Sigma_\pi \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}}, \tag{3.5}$$

where the exponent α is defined as

$$\alpha(\pi, P_\theta) := \frac{1}{d}\operatorname{Tr}\left((\mathrm{I_d} + K^{1/2}(\mathbb{E}_\pi[\bar{\mathcal{I}}_X])^{-1}K^{1/2})^{-1}(\mathrm{I_d} - K^{1/2}\Sigma_\pi K^{1/2})\right).$$

Remark 7. *The dependence of (3.5) on K is only through the exponent $\alpha(\pi, P_\theta)$. If $\pi \in \mathcal{P}(\mathbb{R}^d)$ is K-uniformly log-concave, then it holds by the Brascamp–Lieb variance inequality [Brascamp and Lieb, 2002] that $K^{1/2}\Sigma_\pi K^{1/2} \leq \mathrm{I_d}$, with equality if and only if π is Gaussian with covariance K^{-1}. Hence, we always have $0 \leq \alpha(\pi, P_\theta) \leq 1$, with $\alpha(\pi, P_\theta) = 0$ if and only if π is Gaussian. If it is only known that the prior π is log-concave, then we may take $K = 0$, giving $\alpha(\pi, P_\theta) = 1$.*

Remark 8. *Both $\alpha(\pi, P_\theta)$ and (3.5) more generally are invariant to affine transformations of the parameter. Moreover, they both enjoy optimal dependence on dimension. Indeed, since $\alpha(\pi, P_\theta) \leq 1$, (3.5) can be written as*

$$e^{-\frac{2}{d}I(\pi;P_\theta)} \gtrsim \frac{1}{\det\left(\mathrm{I_d} + \Sigma_\pi \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right)^{1/d}}.$$

In view of Remark 7, when the prior π is Gaussian, then we have $\alpha(\pi, P_\theta) = 0$ and $\mathcal{J}(\pi) = \Sigma_\pi^{-1}$. In this case, we recover Efroimovich's inequality (2.7) exactly. More generally, if we admit the hyperplane conjecture, then Proposition 4 asserts that (3.5) can be considered as a uniform improvement up to a constant factor of Efroimovich's inequality, where dependence on the Fisher information $\mathcal{J}(\pi)$ is completely eliminated, in favor of dependence on the (inverse) second moments Σ_π^{-1}. The latter is obviously more favorable, as a consequence of the classical Cramér–Rao bound (3.2), which can be restated as the matrix inequality $\Sigma_\pi^{-1} \leq \mathcal{J}(\pi)$.

In fact, there are many special cases of practical importance for which the hyperplane conjecture is already known to hold true. Such cases include, for example, *unconditional* log-concave distributions, which are log-concave distributions that are symmetric with respect to reflection about any coordinate axis [Bobkov and Nazarov, 2003] (i.e., a random vector Y has unconditional distribution if its density f satisfies $f(y_1, \ldots, y_d) = f(s_1 y_1, \ldots, s_d y_d)$ for all choices of $y_i \in \mathbb{R}$ and signs $s_i \in \{-1, +1\}$). Such distributions include the archetypical examples of uniform distributions on ℓ_p balls, $p \geq 1$. Efroimovich's inequality is degenerate on such priors, but Theorem 4 shows that degeneracy can be avoided by replacing $\mathcal{J}(\pi)$ with Σ_π^{-1} (the latter is proportional to I_d in the unconditional case). More examples of distributions for which the hyperplane conjecture holds true can be found in the survey [Brazitikos et al., 2014, Chapter 4].

For the case of a K-uniformly log-concave non-Gaussian prior, it is possible to use known results from convex geometry [Ball and Nguyen, 2012, Bakry and Émery, 1985] to crudely bound

$$e^{\frac{2}{d}\mathcal{D}_g(\pi)} \leq C/\lambda_{\min}(K^{1/2}\Sigma_\pi K^{1/2})^2,$$

where C is an absolute constant (recall: the hyperplane conjecture speculates that the RHS above can be replaced by an absolute constant). Better estimates are possible, but this crude estimate further helps to illustrate the point that (3.5) should be thought of as an analogue to Efroimovich's inequality, where dependence on the Fisher information $\mathcal{J}(\pi)$ can be upgraded to dependence on second moments.

Proof of Theorem 4

We make a few preparations before starting the proof of Theorem 4 in earnest. First, for a random variable X having density f with respect to Lebesgue measure on $\mathbb{R}^d$, we define the **Rényi entropy of order** ∞ as

$$h_\infty(X) := -\log(\operatorname{ess\,sup} f).$$

It is immediate from definitions that $h_\infty(X) \leq h(X)$. In case X has a distribution that is uniformly log-concave, we can do slightly better than this.

Lemma 5. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be K-uniformly log-concave, with $K > 0$. If $Z \sim \mathcal{N}(0, I_d)$ is independent of $\eta \sim \mu$, then for all nonsingular $A \in \mathbb{R}^{n\times n}$,*

$$h_\infty(\eta + A^{-1}Z) + \frac{1}{2}\operatorname{Tr}\left((AK^{-1}A^T + I_d)^{-1}(A\Sigma_\mu A^T + I_d)\right) \leq h(\eta + A^{-1}Z).$$

Proof. Define the matrix $\Delta := A^TA > 0$ for convenience, let $Z_\Delta = A^{-1}Z \sim \mathcal{N}(0, \Delta^{-1})$, and consider the density of $\eta + Z_\Delta$, expressed as

$$f(\cdot) \equiv e^{-V(\cdot)} := \frac{\det(\Delta)^{1/d}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}|A(\cdot - x)|^2} d\mu(x).$$

By preservation of uniform log-concavity under convolution (see, e.g., Saumard and Wellner [2014, Theorem 3.7]), we know that μ is K_f-uniformly log-concave with

$$K_f := (K^{-1} + \Delta^{-1})^{-1}.$$

By log-concavity, the point of maximum-likelihood is well defined as $m_\Delta := \arg\max_x f(x)$. In particular, the potential V is K_f-strongly convex, so that

$$V(x) \geq V(m_\Delta) + \frac{1}{2}(x - m_\Delta)^T K_f (x - m_\Delta), \qquad \text{for all } x \in \mathbb{R}^d.$$

With this, we can lower-bound the entropy of $\eta + Z_\Delta$ by

$$\begin{aligned} h(\eta + Z_\Delta) &= -\int_{\mathbb{R}^d} f(x) \log f(x) dx \\ &= \int_{\mathbb{R}^d} f(x) V_f(x) dx \\ &\geq V(m_\Delta) + \frac{1}{2} \int_{\mathbb{R}^d} f(x)(x - m_\Delta)^T K_f (x - m_\Delta) dx \\ &\geq V(m_\Delta) + \frac{1}{2} \operatorname{Tr}\left(K_f \inf_{m \in \mathbb{R}^d} \int_{\mathbb{R}^d} f(x)(x-m)(x-m)^T dx \right) \\ &= h_\infty(\eta + Z_\Delta) + \frac{1}{2} \operatorname{Tr}\left(K_f (\Sigma_\mu + \Delta^{-1}) \right). \end{aligned}$$

Simplification yields the desired conclusion. □

Next, we recall a well-known connection between uniform log-concavity and LSI constants due to Bakry and Émery [1985].

Lemma 6 (The Bakry–Émery Theorem on $\mathbb{R}^d$)**.** *If $\mu \in \mathcal{P}(\mathbb{R}^d)$ is K-uniformly log-concave, then μ satisfies* $\mathrm{LSI}(1/\lambda_{\min}(K))$.

Now we can prove Theorem 4.

Proof of Theorem 4. We first prove the assertion under the assumption that $K > 0$. Consider a K_0-uniformly log-concave prior $\pi_0 \in \mathcal{P}(\mathbb{R}^d)$, and parameter $\eta \sim \pi_0$. Let $Z \sim \mathcal{N}(0, \mathrm{I}_d)$ be independent of η. Fix any positive definite $\Delta \in \mathbb{R}^{d \times d}$ satisfying

$$\lambda_{\max}(\Delta) \leq \lambda_{\min}(\Delta + K_0), \tag{3.6}$$

and choose a symmetric $A = A^T \in \mathbb{R}^d$ such that $\Delta = A^T A$ (e.g., $A = \Delta^{1/2}$). The density of $\eta + A^{-1}Z$ is therefore given by the convolution

$$f(x) := \frac{\det(\Delta)^{1/2}}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\frac{1}{2}|A(\eta - x)|^2} d\pi_0(\eta).$$

Now, define the point of maximum-likelihood

$$m_\Delta := \arg\max_x f(x),$$

which is well-defined by log-concavity. Define a new probability measure $\mu_\Delta \in \mathcal{P}(\mathbb{R}^d)$ by its density

$$d\mu_\Delta(\eta) := C_\Delta^{-1} e^{-\frac{1}{2}|A(\eta - m_\Delta)|^2} d\pi_0(\eta), \quad \eta \in \mathbb{R}^d,$$

where

$$C_\Delta := \int_{\mathbb{R}^d} e^{-\frac{1}{2}|A(\eta - m_\Delta)|^2} d\pi_0(\eta) \tag{3.7}$$

is a normalizing constant. We observe that μ_Δ is $(\Delta + K_0)$-uniformly log-concave by construction, so it satisfies LSI$(1/\lambda_{\min}(\Delta + K_0))$. Note that π_0 has density $C_\Delta e^{\frac{1}{2}|A(\cdot - m_\Delta)|^2}$ with respect to μ_Δ. Therefore, we may readily compute

$$\begin{aligned} D(\pi_0 \| \mu_\Delta) &= \frac{1}{2} \int_{\mathbb{R}^d} |A(\eta - m_\Delta)|^2 d\pi_0(\eta) + \log C_\Delta \\ &= \frac{1}{2} \int_{\mathbb{R}^d} (\eta - m_\Delta)^T \Delta (\eta - m_\Delta) d\pi_0(\eta) + \log C_\Delta \end{aligned}$$

and

$$I(\pi_0 \| \mu_\Delta) = \int_{\mathbb{R}^d} (\eta - m_\Delta)^T \Delta^2 (\eta - m_\Delta)^T d\pi_0(\eta).$$

Now, for $\eta \sim \pi_0$ and $X \sim P_\eta$ given η, it follows from Theorem 2 with reference measure $\mu = \mu_\Delta$, that

$$I(\eta; X) \leq -D(\pi_0 \| \mu_\Delta) + \frac{1}{2\lambda_{\min}(\Delta + K_0)} I(\pi_0 \| \mu_\Delta) + \frac{1}{2\lambda_{\min}(\Delta + K_0)} \mathbb{E}_{\pi_0}[\mathcal{I}_{X;\eta}]. \tag{3.8}$$

To bound the above, observe that

$$\begin{aligned} &\log C_\Delta - D(\pi_0 \| \mu_\Delta) + \frac{1}{2\lambda_{\min}(\Delta + K_0)} I(\pi_0 \| \mu_\Delta) \\ &= -\frac{1}{2} \int_{\mathbb{R}^d} (\eta - m_\Delta)^T \Delta (\eta - m_\Delta) d\pi_0(\eta) + \frac{1}{2\lambda_{\min}(\Delta + K_0)} \int_{\mathbb{R}^d} (\eta - m_\Delta)^T \Delta^2 (\eta - m_\Delta)^T d\pi_0(\eta) \\ &= -\frac{1}{2} \int_{\mathbb{R}^d} (\eta - m_\Delta)^T A^T (\mathrm{I}_d - \frac{1}{\lambda_{\min}(\Delta + K_0)} \Delta) A (\eta - m_\Delta) d\pi_0(\eta) \\ &\leq -\frac{1}{2} \operatorname{Tr} \left(A^T \left(\mathrm{I}_d - \frac{1}{\lambda_{\min}(\Delta + K_0)} \Delta \right) A \Sigma_{\pi_0} \right) \\ &= -\frac{1}{2} \operatorname{Tr} \left(\left(\mathrm{I}_d - \frac{1}{\lambda_{\min}(\Delta + K_0)} \Delta \right) A \Sigma_{\pi_0} A^T \right), \end{aligned}$$

where the inequality follows since (i) we chose Δ to satisfy (3.6), so that

$$\mathrm{I}_d \geq \frac{1}{\lambda_{\min}(\Delta + K_0)} \Delta$$

and, (ii) we always have the matrix inequality

$$\int_{\mathbb{R}^d} (\eta - m_\Delta)(\eta - m_\Delta)^T d\pi_0(\eta) \geq \Sigma_{\pi_0}.$$

Next, with the help of Lemma 5, write

$$
\begin{aligned}
&-\log C_\Delta \\
&= \log\left(\frac{\det(\Delta)^{1/2}}{(2\pi)^{d/2}}\right) - \log\left(\frac{\det(\Delta)^{1/2}}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{-\frac{1}{2}|A(\eta-m_\Delta)|^2} d\pi_0(\eta)\right) \\
&= \log\left(\frac{\det(\Delta)^{1/2}}{(2\pi)^{d/2}}\right) + h_\infty(\eta + A^{-1}Z) \\
&\le \log\left(\frac{\det(\Delta)^{1/2}}{(2\pi)^{d/2}}\right) - \frac{1}{2}\operatorname{Tr}\left((AK_0^{-1}A^T + \mathrm{I}_d)^{-1}(A\Sigma_{\pi_0}A^T + \mathrm{I}_d)\right) + h(\eta + A^{-1}Z) \\
&\le \frac{1}{2}\log\det(\Delta) + \frac{d}{2} - \frac{1}{2}\operatorname{Tr}\left((AK_0^{-1}A^T + \mathrm{I}_d)^{-1}(A\Sigma_{\pi_0}A^T + \mathrm{I}_d)\right) + \frac{1}{2}\log\det(\Sigma_{\pi_0} + \Delta^{-1}) \\
&= \frac{d}{2} - \frac{1}{2}\operatorname{Tr}\left((AK_0^{-1}A^T + \mathrm{I}_d)^{-1}(A\Sigma_{\pi_0}A^T + \mathrm{I}_d)\right) + \frac{1}{2}\log\det(A\Sigma_{\pi_0}A^T + \mathrm{I}_d).
\end{aligned}
$$

Combining with (3.8), we have

$$
\begin{aligned}
I(\eta; X) \le{}& \frac{d}{2} - \frac{1}{2}\operatorname{Tr}\left(A\Sigma_{\pi_0}A^T\right) + \frac{1}{2\lambda_{\min}(\Delta + K_0)}\operatorname{Tr}\left(\Delta\Sigma_{\pi_0}\Delta + \mathbb{E}_{\pi_0}[\bar{\mathcal{I}}_{X;\eta}]\right) \\
&- \frac{1}{2}\operatorname{Tr}\left((AK_0^{-1}A^T + \mathrm{I}_d)^{-1}(A\Sigma_{\pi_0}A^T + \mathrm{I}_d)\right) + \frac{1}{2}\log\det(A\Sigma_{\pi_0}A^T + \mathrm{I}_d) \\
={}& -\operatorname{Tr}\left(A\Sigma_{\pi_0}A^T\right) + \frac{1}{2\lambda_{\min}(\Delta + K_0)}\operatorname{Tr}\left(\Delta\Sigma_{\pi_0}\Delta + \mathbb{E}_{\pi_0}[\bar{\mathcal{I}}_{X;\eta}]\right) \\
&+ \frac{1}{2}\operatorname{Tr}\left((A^{-T}K_0A^{-1} + \mathrm{I}_d)^{-1}(A\Sigma_{\pi_0}A^T + \mathrm{I}_d)\right) + \frac{1}{2}\log\det(A\Sigma_{\pi_0}A^T + \mathrm{I}_d),
\end{aligned}
$$

where we made use of the Woodbury identity $(AK_0^{-1}A^T + \mathrm{I}_d)^{-1} = \mathrm{I}_d - (\mathrm{I}_d + A^{-T}K_0A^{-1})^{-1}$ [Woodbury, 1950] to get to the last line.

Now, we consider the reparametrization $\eta = A^{-1}V\theta$ for some invertible matrix V. This introduces the change of variables

$$
\begin{aligned}
\Sigma_{\pi_0} &\leftarrow A^{-1}V\Sigma_\pi V^T A^{-T} \\
\mathbb{E}_{\pi_0}[\bar{\mathcal{I}}_{X;\eta}] &\leftarrow A^T V^{-T}\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}]V^{-1}A \\
K_0 &\leftarrow A^T V^{-T} K V^{-1} A.
\end{aligned}
$$

Making these substitutions, we have

$$
\begin{aligned}
I(\pi; P_\theta) \le{}& -\operatorname{Tr}\left(V\Sigma_\pi V^T\right) \\
&+ \frac{1}{2\lambda_{\min}(A^T(\mathrm{I}_d + V^{-T}KV^{-1})A)}\operatorname{Tr}\left(A^TV\Sigma_\pi V^TA + A^TV^{-T}\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}]V^{-1}A\right) \\
&+ \frac{1}{2}\operatorname{Tr}\left((V^{-T}KV^{-1} + \mathrm{I}_d)^{-1}(V\Sigma_\pi V^T + \mathrm{I}_d)\right) + \frac{1}{2}\log\det(V\Sigma_\pi V^T + \mathrm{I}_d),
\end{aligned}
$$

for any symmetric matrix A satisfying

$$\lambda_{\max}(A^T A) \leq \lambda_{\min}(A^T(\mathrm{I}_d + V^{-T}KV^{-1})A).$$

Choosing $A = (\mathrm{I}_d + V^{-T}KV^{-1})^{-1/2}$ is a valid choice, forcing the maximum eigenvalue to be no greater than unity, which is the precise value of the minimum eigenvalue on the right. This leaves us with the unconditional inequality, for any invertible matrix V:

$$\begin{aligned} I(\pi; P_\theta) \leq &- \operatorname{Tr}\left(V\Sigma_\pi V^T\right) \\ &+ \frac{1}{2}\operatorname{Tr}\left((V^{-T}KV^{-1} + \mathrm{I}_d)^{-1}(V\Sigma_\pi V^T + V^{-T}\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}]V^{-1})\right) \\ &+ \frac{1}{2}\operatorname{Tr}\left((V^{-T}KV^{-1} + \mathrm{I}_d)^{-1}(V\Sigma_\pi V^T + \mathrm{I}_d)\right) + \frac{1}{2}\log\det(V\Sigma_\pi V^T + \mathrm{I}_d) \\ = &\frac{1}{2}\operatorname{Tr}\left((V^TV + K)^{-1}(\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}])\right) + \frac{1}{2}\operatorname{Tr}\left((V^TV + K)^{-1}(-2K\Sigma_\pi + \mathrm{I}_d)V^TV\right) \\ &+ \frac{1}{2}\log\det\left(\Sigma_\pi V^TV + \mathrm{I}_d\right). \end{aligned}$$

Now, assume $\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}]$ is nonsingular for convenience[1]. We may then further specialize by choosing $V = (\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}])^{1/2}$ to find, after simplification, that

$$\begin{aligned} I(\pi; P_\theta) \leq &\frac{1}{2}\log\det\left(\Sigma_\pi\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}] + \mathrm{I}_d\right) \\ &+ \operatorname{Tr}\left((\mathrm{I}_d + K^{1/2}(\mathbb{E}[\bar{\mathcal{I}}_{X;\theta}])^{-1}K^{1/2})^{-1}(\mathrm{I}_d - K^{1/2}\Sigma_\pi K^{1/2})\right). \end{aligned}$$

The above is the desired result under the assumption K is positive definite. In case it is not, it suffices to consider the perturbed measure π_ϵ defined by

$$d\pi_\epsilon(x) \propto e^{\epsilon|x|^2/2}d\pi(x),$$

which is $(K + \epsilon\mathrm{I}_d)$-uniformly log-concave. It is straightforward to see that letting $\epsilon \downarrow 0$ yields the desired result in the general case of $K \geq 0$. □

3.3 Examples

We present several examples to illustrate tightness of Theorem 4.

Gaussian priors and Gaussian observations

Given a Gaussian prior $\pi \sim \mathcal{N}(0_d, \Sigma_\pi)$, the corresponding K satisfies

$$K\Sigma_\pi = \mathrm{I}_d,$$

[1] If this is not the case, we can consider $V := \mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}] + \epsilon\mathrm{I}_d$, and let $\epsilon \downarrow 0$ to reach the conclusion.

and therefore Theorem 4 gives

$$e^{-\frac{2}{d}I(\pi;P_\theta)} \geq \frac{1}{\det\left(\mathrm{I}_d + \Sigma_\pi \mathbb{E}[\bar{\mathcal{I}}_X]\right)^{\frac{1}{d}}}. \tag{3.9}$$

Consider the Gaussian observation $X = \theta + Z$ with $Z \sim \mathcal{N}(0, \Sigma_Z)$ independent from θ. Then, note that $\mathbb{E}[\bar{\mathcal{I}}_X] = \Sigma_Z^{-1}$ and

$$I(\pi; P_\theta) = \frac{1}{2} \log\det\left(\mathrm{I}_d + \Sigma_\pi \Sigma_Z^{-1}\right).$$

It follows that (3.9) is satisfied with equality.

Uniformly log-concave priors with vanishing $\mathcal{I}_X$

Given any uniformly log-concave prior with $K > 0$, Theorem 4 implies

$$\lim_{\lambda_{\max}\left(\mathbb{E}[\bar{\mathcal{I}}_X]\right)\to 0} e^{-\frac{2}{d}I(\pi;P_\theta)} \geq 1.$$

Note that $I(\pi; P_\theta) \to 0$ as $\lambda_{\max}(\mathbb{E}[\bar{\mathcal{I}}_X]) \to 0$, and therefore the above is satisfied with equality.

Applications to situations where Jeffrey's prior is log-concave

Consider i.i.d. observations $X = (X_1, \ldots, X_n)$ generated according to $(P_\theta)_{\theta\in B}$. Jeffrey's prior, which has density $\psi(\cdot) \propto \sqrt{\mathcal{I}_{X_1}}$, is asymptotically least favorable under KL-divergence as $n \to \infty$ [Clarke and Barron, 1994]. Theorem 4 provides a non-asymptotic upper bound on $I(\pi_J; P_\theta)$ given π_J is a log-concave Jeffrey's prior.

Example 16 (Jeffrey's prior of a Gaussian observation model over a convex set)**.** *Suppose $\Theta \subset \mathbb{R}^d$ is any convex set with finite volume. Consider the Gaussian observation model $P_\theta = \mathcal{N}(\theta, \Sigma)$. Then, since $\mathcal{I}_X(\theta) = \Sigma^{-1}$ for all $\theta \in \Theta$ (see Example 9), it follows that the Jeffrey's prior is the uniform measure over Θ.*

Example 17 (Jeffrey's prior of a Poisson observation model)**.** *Consider the poisson distribution with parameter $e^{-\theta}$. A straightforward calculation yields $\mathcal{I}_X(\theta) = e^{-\theta}$. Therefore, the Jeffrey's prior is* Exponential$(1/2)$.

While the Jeffrey's priors in the above two examples are log-concave, Efroimovich's inequality cannot be applied to give a meaningful bound to $I(\pi; P_\theta)$ since $\mathcal{J}(\pi) = +\infty$ in both scenarios. It should be noted that Jeffrey's prior is not necessarily log-concave. An example is the Jeffrey's prior under a binomial observation model.

Example 18 (Jeffrey's prior of a binomial observation model)**.** *Suppose $X = (X_1, \ldots, X_n) \sim$* binomial$(n, \theta)$ *with $\Theta = [0, 1]$. Recall from Example 13 that the Fisher information is $\mathcal{I}_X(\theta) = n\theta^{-1}(1-\theta)^{-1}$. It follows that the Jeffrey's prior having density ψ satisfies $\psi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$, which corresponds to a* Beta$(1/2, 1/2)$ *distribution. Calculations yield $(\log\psi(\theta))'' \propto \theta^{-2} + (1-\theta)^{-2} \geq 0$, showing that π_J is not log-concave.*

3.4 Variation on a theme

Suppose a given prior $\pi \in \mathcal{P}(\mathbb{R}^d)$ is only known to be log-concave. In this case, Theorem 4 yields the upper bound

$$I(\pi; P_\theta) \leq d + \frac{1}{2} \log \det(\mathrm{I}_d + \Sigma_\pi \mathbb{E}_\pi[\bar{\mathcal{I}}_X]).$$

This is somewhat unsatisfactory in the regime where $\lambda_{\max}(\mathbb{E}_\pi[\bar{\mathcal{I}}_X]) \to 0$, since in the extreme case of $\mathbb{E}_\pi[\bar{\mathcal{I}}_X] = 0$, we have $I(\pi; P_\theta) = 0$.

The proof of Theorem 4 hints at points of possible improvements. It's possible one could find a better choice of the matrices A and V, but this seems incapable of addressing the issue noted above. The greater point of potential improvement seems to be sharpening the bound on the normalizing constant C_Δ in (3.7) by alternate means. The goal of this section is to explore this possibility. Ultimately, we obtain a result similar in spirit, but not directly comparable, to Theorem 4. It satisfies the desideratum that the upper bound on $I(\pi; P_\theta)$ vanishes as $\lambda_{\max}(\mathbb{E}_\pi[\bar{\mathcal{I}}_X]) \to 0$, and is stated as follows.

Theorem 7. *Fix a parametric model $(P_\theta)_{\theta \in \mathbb{R}^d}$ and prior $\pi \in \mathcal{P}(\mathbb{R}^d)$. Define $P := \frac{1}{d} \operatorname{Tr}(\Sigma_\pi)$ and $J := \frac{1}{d} \mathbb{E}_\pi[\mathcal{I}_X]$. If π is K-uniformly log-concave, then*

$$I(\pi; P_\theta) \leq \phi(\lambda_{\min}(K)P, JP) \tag{3.10}$$

where

$$\phi(a, b) := \begin{cases} d\left(\sqrt{a^2 + b} - a\right) & \text{if } a^2 + b < 1 \\ d\,(1 - a) + \frac{d}{2} \log(a^2 + b) & \text{if } a^2 + b \geq 1. \end{cases}$$

Let's briefly compare Theorems 4 and 7. The two are similar, in that they each provide a concise upper bound on the mutual information $I(\pi; P_\theta)$ when the prior is log-concave. Moreover, the upper bounds depend primarily on the covariance Σ_π and the Fisher information matrix $\bar{\mathcal{I}}_X$, averaged over the prior. However, the bound in Theorem 4 depends on the determinants of these matrices, whereas (3.10) involves traces. It is unclear whether the latter can be upgraded to involve only determinants like the former. The difficulty lies in the observation that this objective competes with our technique for improving the bound on the normalizing constant C_Δ.

It is straightforward to see that neither result subsumes the other. Under the assumption that the prior is simply log-concave and $\frac{1}{d} \operatorname{Tr}(\Sigma_\pi) \frac{1}{d} \mathbb{E}_\pi[\mathcal{I}_X] \gg 1$, Theorem 4 will generally be favorable compared to Theorem 7, since the former essentially depends on the geometric mean of the eigenvalues of the matrix $\Sigma_\pi^{1/2} \mathbb{E}_\pi[\mathcal{I}_X] \Sigma_\pi^{1/2}$, whereas the latter depends on the arithmetic mean. Hence, if the eigenvalues of the matrix $\Sigma_\pi^{1/2} \mathbb{E}_\pi[\mathcal{I}_X] \Sigma_\pi^{1/2}$ are not highly concentrated, the determinantal inequality will generally be superior to the trace inequality.

However, in dimension $d = 1$, there is no distinction between the trace and determinant, and the following example shows how Theorem 7 can provide uniform improvement upon Theorem 4.

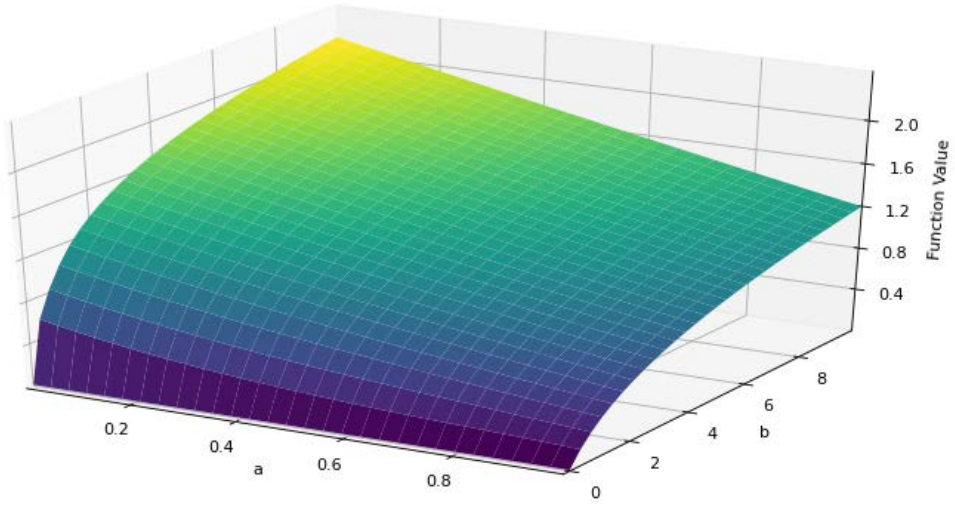

Figure 3.1: The landscape of $\frac{1}{d}\phi(a,b)$ defined in Theorem 7. The values of a are taken between $[0,1]$ by virtue of the Brascamp–Lieb inequality [Brascamp and Lieb, 2002]. Note that the slope of $\frac{1}{d}\phi(0,b)$ at $b \to 0$ approaches $+\infty$, which is not completely captured here.

Example 19. *Let $\pi \in \mathbb{P}(\mathbb{R})$ be log-concave. For $\theta \sim \pi$, Theorem 7 gives the estimate*

$$I(\pi; P_\theta) \leq \begin{cases} \sqrt{\mathrm{Var}(\theta)\mathbb{E}_\pi[\mathcal{I}_X]} & \text{if } \mathrm{Var}(\theta)\mathbb{E}_\pi[\mathcal{I}_X] < 1 \\ 1 + \frac{1}{2}\log(\mathrm{Var}(\theta)\mathbb{E}_\pi[\mathcal{I}_X]) & \text{if } \mathrm{Var}(\theta)\mathbb{E}_\pi[\mathcal{I}_X] \geq 1, \end{cases}$$

while Theorem 4 gives the uniformly inferior bound

$$I(\pi; P_\theta) \leq 1 + \frac{1}{2}\log(1 + \mathrm{Var}(\theta)\mathbb{E}_\pi[\mathcal{I}_X]).$$

All this is to say that the story here may not yet be complete, and future work could consider whether Theorems 4 and 7 can be reconciled into a single result.

Proof of Theorem 7

The proof of Theorem 7 proceeds roughly along the same lines as that of Theorem 4, with steps taken to address the questions posed at the beginning of this section. The proof hinges on the following technical proposition, which relates closely to our selection of the maximum likelihood point m_Δ in the proof of Theorem 4. It can be explained intuitively as follows: if we convolve a log-concave probability measure μ with a Gaussian of covariance $\delta^{-1}\mathrm{I}_d$, then the point of maximum likelihood of the resulting density (call it m_δ) is unique, and changes smoothly as we adjust $\delta > 0$. The last part of the proposition gives a lower bound on the likelihood at m_δ, as we require. The only real surprise is the fact that m_δ is also the

barycenter of the density proportional to $e^{-\delta|x-m_\delta|^2/2}d\mu(x)$, which is part (i) of the claim, and is also helpful in the proof.

Proposition 5. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be K-uniformly log-concave , with $k := \lambda_{\min}(K)$.*

(i) For each $\delta \geq 0$, there exists a unique $m_\delta \in \mathbb{R}^d$ such that

$$\int_{\mathbb{R}^d} x e^{-\delta|x-m_\delta|^2/2} d\mu(x) = m_\delta \int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2} d\mu(x).$$

(ii) For each $\delta > 0$, the map

$$m \longmapsto \int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2} d\mu(x)$$

has a unique global maximum at m_δ.

(iii) The map $\delta \longmapsto m_\delta$ is continuous on $\delta \in (0, \infty)$. In particular, for each $\delta > 0$, there is a neighborhood U_δ of δ and $L_\delta < \infty$ such that $|m_{\delta'} - m_\delta| \leq L_\delta |\delta' - \delta|$ for all $\delta' \in U_\delta$.

(iv) Let $\sigma^2 = \frac{1}{d}\int_{\mathbb{R}^d} |x - m_0|^2 d\mu(x)$. For m_δ as in part (i), and each $\delta \geq 0$,

$$-\log\left(\int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2} d\mu(x)\right) \leq \begin{cases} \frac{d}{2}\delta\sigma^2 & \text{if } 0 \leq \delta < \frac{1}{\sigma^2} - k, \\ \frac{d}{2}\big(1 - k\sigma^2 + \log\left((\delta + k)\sigma^2\right)\big) & \text{if } \delta \geq \frac{1}{\sigma^2} - k. \end{cases}$$

Remark 9. *The inequality in Proposition 5 translates to an upper bound on the Rényi entropy $h_\infty(Y + Z_\delta)$, where $Y \sim \mu$ and $Z_\delta \sim \mathcal{N}(0, \delta^{-1} I_d)$ are independent. In general, tools like the entropy power inequality in conjunction with estimates in [Bobkov and Madiman, 2011] can be used to obtain lower bounds on $h_\infty(Y + Z_\delta)$ in terms of the marginal entropies of Y and Z_δ. However, obtaining nontrivial upper bounds is a difficult problem, as the entropy power inequality cannot be reversed in general. Nevertheless, for log-concave random vectors (as is the case here), reversal is possible under a suitable affine transformation of Y to put it in so-called "M-position" introduced by Milman [1986.]; see, e.g., [Pisier, 1999, Bobkov and Madiman, 2012]. Unfortunately, this transformation is not explicit, and so the result is not obviously applicable to our setting.*

Before the proof of Proposition 5, we collect a few preparatory lemmas. The first is the well-known Banach fixed-point theorem [Banach, 1922].

Lemma 8 (Banach fixed-point theorem)**.** *Let $(\mathcal{X}, \rho)$ be a complete metric space, and let $T : \mathcal{X} \to \mathcal{X}$ satisfy $\rho(T(x), T(y)) \leq \lambda\rho(x, y)$ for all $x, y \in \mathcal{X}$, where $\lambda < 1$. Then T has a unique fixed point $x^* \in \mathcal{X}$. Moreover, if $x_0 \in \mathcal{X}$ and $x_{n+1} := T(x_n)$, $n \geq 0$, then*

$$\rho(x_n, x^*) \leq \frac{\lambda^n}{1 - \lambda}\rho(T(x_0), x_0), \qquad n \geq 0. \tag{3.11}$$

The next is effectively a stability result for the Poincáre constant for k-uniformly log-concave distributions [Courtade and Fathi, 2020].

Lemma 9. *Let $\nu \in \mathcal{P}(\mathbb{R}^d)$ be K-uniformly log-concave with $K \geq kI_d$, $k > 0$. If $Y \sim \nu$ and $u \in \mathbb{S}^{d-1}$, then*

$$\mathrm{Var}(\langle u, Y\rangle) \leq \frac{1}{k},$$

with equality only if $\langle u, Y\rangle$ is Gaussian with variance $1/k$, independent of $Y - \langle u, Y\rangle u$.

Proof. By rescaling, we can assume without loss of generality that $k = 1$. Thus, by the Bakry–Émery theorem, it follows that ν satisfies LSI(1). It is well-known that an LSI implies a Poincaré inequality with the same constant by Rothaus's linearization argument [Rothaus, 1981], and therefore we conclude that

$$\mathrm{Var}(\langle u, Y\rangle) \leq \mathbb{E}[|u|^2] = 1.$$

By stability results for the Poincaré inequality [Courtade and Fathi, 2020], we know that equality is attained if and only if $\langle u, Y\rangle$ is Gaussian with variance $1/k$, independent of $Y - \langle u, Y\rangle u$. □

Lemma 10 (Herbst's argument [Bakry et al., 2014, Prop. 5.4.1.])**.** *Let $\nu \in \mathcal{P}(\mathbb{R}^d)$ satisfy* LSI(C). *For every 1-Lipschitz $F : \mathbb{R}^d \to \mathbb{R}$ and $t^2 \leq 1/C$, it holds that*

$$\int_{\mathbb{R}^d} e^{t^2 F^2/2} d\nu < \infty.$$

Proof of Proposition 5. **Proof of (i):** We start by showing that the map $T_\delta : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ defined by

$$T_\delta : m \longmapsto \frac{\int_{\mathbb{R}^d} x e^{-\delta|x-m|^2/2} d\mu(x)}{\int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2} d\mu(x)}$$

is a contraction with respect to the usual Euclidean metric. Then, the claims follow from the Banach fixed-point theorem.

So, begin by fixing $\delta > 0$, and let $\mu_{m,\delta}$ denote the probability measure with density proportional to the density $e^{-\delta|x-m|^2/2} d\mu(x)$. That is,

$$d\mu_{m,\delta}(x) = Z^{-1} e^{-\delta|x-m|^2/2} d\mu(x),$$

where Z is a normalizing constant. Observe that $\mu_{m,\delta}$ is δ-uniformly log-concave.

Now, we claim that $\mu_{m,\delta}$ cannot split off an independent Gaussian factor with variance $1/\delta$. Indeed, if this were the case, then after suitable change of coordinates, we could assume $\mu_{m,\delta}$ splits off an independent Gaussian factor of variance $1/\delta$ in the first coordinate, so that

$$d\mu_{m,\delta}(x) = e^{-W(x_2,\ldots,x_d) - \delta|x_1 - c|^2/2} dx = Z^{-1} e^{-\delta|x-m|^2/2} d\mu(x),$$

for some $c \in \mathbb{R}$. Rearranging, this yields

$$\frac{d\mu}{dx}(x) = Ze^{\delta|x-m|^2/2 - W(x_2,\ldots,x_d) - \delta|x_1-c|^2/2},$$

which is not integrable in coordinate x_1, a contradiction. Hence, it follows from Lemma 9 that

$$\lambda_{\max}(\Sigma_{m,\delta}) < \frac{1}{\delta} \tag{3.12}$$

where $\Sigma_{m,\delta}$ denotes the covariance matrix associated to $\mu_{m,\delta}$.

By differentiating the ith coordinate of T_δ at m, we see that

$$\begin{aligned}
&\nabla[T_\delta]_i(m) \\
&= \delta \frac{\int_{\mathbb{R}^d} x_i(x-m)e^{-\delta|x-m|^2/2}d\mu(x)}{\int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2}d\mu(x)} - \delta \frac{\left(\int_{\mathbb{R}^d} x_i e^{-\delta|x-m|^2/2}d\mu(x)\right)\left(\int_{\mathbb{R}^d}(x-m)e^{-\delta|x-m|^2/2}d\mu(x)\right)}{\left(\int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2}d\mu(x)\right)^2} \\
&= \delta \frac{\int_{\mathbb{R}^d} x_i x e^{-\delta|x-m|^2/2}d\mu(x)}{\int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2}d\mu(x)} - \delta \frac{\left(\int_{\mathbb{R}^d} x_i e^{-\delta|x-m|^2/2}d\mu(x)\right)\left(\int_{\mathbb{R}^d} x e^{-\delta|x-m|^2/2}d\mu(x)\right)}{\left(\int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2}d\mu(x)\right)^2} \\
&= \delta \int_{\mathbb{R}^d} x_i x d\mu_{m,\delta}(x) - \delta \left(\int_{\mathbb{R}^d} x_i d\mu_{m,\delta}(x)\right)\left(\int_{\mathbb{R}^d} x d\mu_{m,\delta}(x)\right).
\end{aligned}$$

Hence, the Jacobian DT_δ of T_δ has entries $[DT_\delta(m)]_{ij} = \delta[\Sigma_{m,\delta}]_{i,j}$. Recalling (3.12), we find

$$\|T_\delta\|_{\text{Lip}} = \lambda_{\max}(DT_\delta(m)) = \delta\lambda_{\max}(\Sigma_{m,\delta}) < 1,$$

so that T_δ is a contraction as claimed. Hence, the desired existence and uniqueness of m_δ follows from the Banach fixed-point theorem.

Proof of (ii): To prove the second claim, note that for any $m \neq m_\delta$ and $t \in [0,1)$,

$$\begin{aligned}
&\frac{d}{dt}\int_{\mathbb{R}^d} e^{-\delta|x-((1-t)m+tm_\delta)|^2/2}d\mu(x) \\
&\quad= \delta\int_{\mathbb{R}^d}\langle x-((1-t)m+tm_\delta), m_\delta - m\rangle e^{-\delta|x-tm_\delta|^2/2}d\mu(x) \\
&\quad\propto \langle T_\delta((1-t)m+tm_\delta) - ((1-t)m+tm_\delta), m_\delta - m\rangle \\
&\quad= \langle T_\delta((1-t)m+tm_\delta) - T_\delta(m_\delta), m_\delta - m\rangle + (1-t)|m_\delta - m|^2 \\
&\quad\geq -|T_\delta((1-t)m+tm_\delta) - T_\delta(m_\delta)||m_\delta - m| + (1-t)|m_\delta - m|^2 \\
&\quad> -|(1-t)m - (1-t)m_\delta||m_\delta - m| + (1-t)|m_\delta - m|^2 \\
&\quad= 0.
\end{aligned}$$

The strict inequality holds since T_δ is a contraction and $(1-t)m + tm_\delta \neq m_\delta$ for $t \in [0,1)$. Thus, for any $m \in \mathbb{R}^d$ not equal to m_δ, the map $t \mapsto \int_{\mathbb{R}^d} e^{-\delta|x-((1-t)m+tm_\delta)|^2/2}d\mu(x)$ is strictly

increasing on $[0,1)$, so that $m \mapsto \int_{\mathbb{R}^d} e^{-\delta|x-m|^2/2} d\mu(x)$ achieves a unique global maximum at m_δ as claimed.

Proof of (iii): We first note that (ii) proved above yields a uniform bound on $|m_\delta|$ for all $\delta > 0$. In particular,

$$\begin{aligned}\int_{\mathbb{R}^d} |x| d\mu(x) \geq \int_{\mathbb{R}^d} |x| e^{-\delta|x-m_\delta|^2/2} d\mu(x) &\geq \left| \int_{\mathbb{R}^d} x e^{-\delta|x-m_\delta|^2/2} d\mu(x) \right| \\ &= |m_\delta| \int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2} d\mu(x) \\ &\geq |m_\delta| \int_{\mathbb{R}^d} e^{-\delta|x|^2/2} d\mu(x) \\ &\geq |m_\delta| \exp\left(-\frac{\delta}{2} \int_{\mathbb{R}^d} |x|^2 d\mu(x) \right).\end{aligned}$$

Since μ is log-concave, it has finite moments of all orders, and we conclude

$$|m_\delta| \leq \exp\left(\frac{\delta}{2} \int_{\mathbb{R}^d} |x|^2 d\mu(x) \right) \int_{\mathbb{R}^d} |x| d\mu(x) < \infty. \tag{3.13}$$

For each $\delta > 0$, we introduce the more convenient notation $\mu_\delta := \mu_{m_\delta, \delta}$, where m_δ is defined as in part (i). By Taylor's theorem

$$|e^{\epsilon f} - (1 + \epsilon f)| \leq \frac{\epsilon^2 f^2}{2} e^{|\epsilon f|},$$

so it follows that

$$\begin{aligned}&\frac{\int_{\mathbb{R}^d} x e^{-\delta|x-m_{\delta+\epsilon}|^2/2} d\mu(x)}{\int_{\mathbb{R}^d} e^{-(\delta+\epsilon)|x-m_{\delta+\epsilon}|^2/2} d\mu(x)} \\ &= \int_{\mathbb{R}^d} x e^{\epsilon|x-m_{\delta+\epsilon}|^2/2} d\mu_{\delta+\epsilon}(x) \\ &= \int_{\mathbb{R}^d} x \Big(1 + \frac{\epsilon}{2}|x - m_{\delta+\epsilon}|^2 + O(\epsilon^2 |x - m_{\delta+\epsilon}|^4 e^{|\epsilon||x-m_{\delta+\epsilon}|^2/2}) \Big) d\mu_{\delta+\epsilon}(x),\end{aligned}$$

where the big-O term hides only numerical constants. To show that the error term remains small after integration, note that

$$\begin{aligned}&\left| \int_{\mathbb{R}^d} \left(x\epsilon^2 |x - m_{\delta+\epsilon}|^4 e^{|\epsilon||x-m_{\delta+\epsilon}|^2/2} \right) d\mu_{\delta+\epsilon}(x) \right| \\ &\leq \epsilon^2 \int_{\mathbb{R}^d} \left(|x||x - m_{\delta+\epsilon}|^4 e^{|\epsilon||x-m_{\delta+\epsilon}|^2/2} \right) d\mu_{\delta+\epsilon}(x) \\ &\leq \epsilon^2 \left(\int_{\mathbb{R}^d} |x|^2 |x - m_{\delta+\epsilon}|^8 d\mu_{\delta+\epsilon}(x) \right)^{1/2} \left(\int_{\mathbb{R}^d} e^{|\epsilon||x-m_{\delta+\epsilon}|^2} d\mu_{\delta+\epsilon}(x) \right)^{1/2}.\end{aligned} \tag{3.14}$$

Since $\mu_{\delta+\epsilon}$ is log-concave, a well-known inequality of Borell (e.g., Borell [1974], Latała and Wojtaszczyk [2008]) ensures that

$$\left(\int_{\mathbb{R}^d} |x-m|^p d\mu_{\delta+\epsilon}(x)\right)^{1/p} \leq C\frac{p}{q}\left(\int_{\mathbb{R}^d} |x-m|^q d\mu_{\delta+\epsilon}(x)\right)^{1/q}$$

for all $1 \leq q \leq p < \infty$ and $m \in \mathbb{R}^d$, where C is an absolute constant. Thus, since $\int_{\mathbb{R}^d} |x-m_{\delta+\epsilon}|^2 d\mu_{\delta+\epsilon}(x) \leq \frac{d}{\delta+\epsilon}$ by (i) and Lemma 9, and $\delta \mapsto |m_\delta|$ is bounded via (3.13), the first term in (3.14) involving polynomial moments is finite and uniformly bounded (in terms of δ) for all ϵ sufficiently small. Additionally, since $\mu_{\delta+\epsilon}$ is uniformly log-concave by construction, it satisfies LSI($\delta/2$) for all $|\epsilon| < \delta/2$ by the Bakry–Émery theorem. Hence, $\int_{\mathbb{R}^d} e^{|\epsilon||x-m_{\delta+\epsilon}|^2} d\mu_{\delta+\epsilon}(x)$ is finite by Lemma 10, and uniformly bounded in terms of d, δ, for all ϵ sufficiently small.

Summarizing, we have

$$\begin{aligned}\frac{\int_{\mathbb{R}^d} x e^{-\delta|x-m_{\delta+\epsilon}|^2/2} d\mu(x)}{\int_{\mathbb{R}^d} e^{-(\delta+\epsilon)|x-m_{\delta+\epsilon}|^2/2} d\mu(x)} &= \int_{\mathbb{R}^d} x\left(1+\frac{\epsilon}{2}|x-m_{\delta+\epsilon}|^2\right) d\mu_{\delta+\epsilon}(x) + O(\epsilon^2) \\ &= m_{\delta+\epsilon} + \frac{\epsilon}{2}\int x|x-m_{\delta+\epsilon}|^2 d\mu_{\delta+\epsilon}(x) + O(\epsilon^2)\end{aligned}$$

and, by similar arguments,

$$\begin{aligned}\frac{\int_{\mathbb{R}^d} e^{-\delta|x-m_{\delta+\epsilon}|^2/2} d\mu(x)}{\int_{\mathbb{R}^d} e^{-(\delta+\epsilon)|x-m_{\delta+\epsilon}|^2/2} d\mu(x)} &= \int_{\mathbb{R}^d} \left(1+\frac{\epsilon}{2}|x-m_{\delta+\epsilon}|^2\right) d\mu_{\delta+\epsilon}(x) + O(\epsilon^2) \\ &= 1 + \frac{\epsilon}{2}\int |x-m_{\delta+\epsilon}|^2 d\mu_{\delta+\epsilon}(x) + O(\epsilon^2),\end{aligned}$$

where the big-O terms hide finite constants that depend on δ, but not on ϵ. In particular, since

$$T_\delta(m_{\delta+\epsilon}) - m_{\delta+\epsilon} = \frac{\int_{\mathbb{R}^d} x e^{-\delta|x-m_{\delta+\epsilon}|^2/2} d\mu(x)}{\int_{\mathbb{R}^d} e^{-\delta|x-m_{\delta+\epsilon}|^2/2} d\mu(x)} - m_{\delta+\epsilon},$$

we can conclude using the above estimates and uniform boundedness of $\delta \mapsto |m_\delta|$ that, for ϵ sufficiently small,

$$|T_\delta(m_{\delta+\epsilon}) - m_{\delta+\epsilon}| \leq |\epsilon| C_\delta,$$

where $C_\delta < \infty$ depends on δ, but not ϵ.

Now, applying the second part of the Banach fixed-point theorem, we find for all ϵ sufficiently small

$$|m_{\delta+\epsilon} - m_\delta| \leq \frac{1}{1-\|T_\delta\|_{\mathrm{Lip}}} |T_\delta(m_{\delta+\epsilon}) - m_{\delta+\epsilon}| \leq \frac{|\epsilon| C_\delta}{1-\|T_\delta\|_{\mathrm{Lip}}},$$

where we used the fact that m_δ is the fixed point of T_δ. Since $\|T_\delta\|_{\mathrm{Lip}} < 1$ from the proof of (i), the proof of (iii) is complete.

Proof of (iv): For convenience, define for $\delta \geq 0$

$$g(\delta) := -\log\left(\int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2} d\mu(x)\right).$$

Since μ is a probability measure, we have $g(0) = 0$, so we focus henceforth on $\delta > 0$. For $\delta', \delta > 0$ the bound $|x - m_{\delta'}|^2 \leq |x - m_\delta|^2 + 2\langle x - m_{\delta'}, m_\delta - m_{\delta'}\rangle$ applies to give

$$\begin{aligned}
g(\delta') - g(\delta) &= -\log\left(\frac{\int_{\mathbb{R}^d} e^{-\delta'|x-m_{\delta'}|^2/2} d\mu(x)}{\int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2} d\mu(x)}\right) \\
&\leq -\log\left(\int_{\mathbb{R}^d} e^{-(\delta'-\delta)|x-m_\delta|^2/2 - \delta'\langle x-m_{\delta'}, m_\delta - m_{\delta'}\rangle} d\mu_\delta(x)\right) \\
&\leq \int_{\mathbb{R}^d} \Big((\delta'-\delta)|x-m_\delta|^2/2 + \delta'\langle x - m_{\delta'}, m_\delta - m_{\delta'}\rangle\Big) d\mu_\delta(x) \\
&= (\delta'-\delta)\left(\frac{1}{2}\int_{\mathbb{R}^d} |x-m_\delta|^2 d\mu_\delta(x) + \frac{\delta'}{\delta'-\delta}|m_\delta - m_{\delta'}|^2\right),
\end{aligned}$$

where we used convexity of $t \mapsto -\log(t)$ in the second inequality, and the final equality used $\int_{\mathbb{R}^d} x d\mu_\delta = m_\delta$. Switching the roles of δ, δ', we have the reverse inequality

$$g(\delta') - g(\delta) \geq (\delta'-\delta)\left(\frac{1}{2}\int_{\mathbb{R}^d} |x-m_{\delta'}|^2 d\mu_{\delta'}(x) - \frac{\delta}{\delta'-\delta}|m_\delta - m_{\delta'}|^2\right).$$

By (iii), it holds that $|m_\delta - m_{\delta'}|^2 \leq L_\delta^2 |\delta - \delta'|^2$ for δ' sufficiently close to δ. Additionally, $\int_{\mathbb{R}^d} |x - m_\delta|^2 d\mu_\delta \leq \frac{d}{k+\delta}$ for each $\delta > 0$ by Lemma 9. Thus, for $\delta', \delta > 0$, with $|\delta - \delta'|$ sufficiently small,

$$|g(\delta') - g(\delta)| \leq |\delta' - \delta|\left(\frac{d}{2(k+\delta)} + \delta' L_\delta^2 |\delta - \delta'|\right).$$

In particular, g is continuous on $(0, \infty)$ with upper Dini derivative bounded by

$$g'_+(\delta) := \limsup_{\epsilon \to 0^+} \frac{g(\delta+\epsilon) - g(\delta)}{\epsilon} \leq \frac{d}{2(k+\delta)}.$$

Hence, we have for $\delta \geq \delta_0 > 0$

$$g(\delta) \leq g(\delta_0) + \int_{\delta_0}^{\delta} g'_+(s) ds \leq g(\delta_0) + \frac{d}{2}\int_{\delta_0}^{\delta} \frac{1}{k+s} ds = g(\delta_0) + \frac{d}{2}\log\frac{k+\delta}{k+\delta_0}. \tag{3.15}$$

By (ii) and convexity of $t \mapsto -\log t$, we have

$$\begin{aligned}
g(\delta) &= -\log\left(\int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2} d\mu(x)\right) \\
&\leq -\log\left(\int_{\mathbb{R}^d} e^{-\delta|x-m_0|^2/2} d\mu(x)\right) \\
&\leq \frac{\delta}{2}\int_{\mathbb{R}^d} |x-m_0|^2 d\mu(x) = \frac{d}{2}\delta\sigma^2.
\end{aligned}$$

Thus, we conclude from this and (3.15) that

$$-\log\left(\int_{\mathbb{R}^d} e^{-\delta|x-m_\delta|^2/2}d\mu(x)\right) \le \frac{d}{2}\inf_{\delta_0\in(0,\delta)}\left\{\delta_0 + \log\frac{k+\delta}{k+\delta_0}\right\}.$$

Optimizing over $\delta_0 > 0$, we conclude the desired upper bound. □

With Proposition 5 in hand, we can finally turn our attention to the proof of Theorem 7.

Proof of Theorem 7. For $\delta > 0$, let μ_δ denote the probability measure with density

$$d\mu_\delta(x) = C_\delta^{-1} e^{-\delta|x-m_\delta|^2/2}d\pi(x),$$

where $C_\delta = \int e^{-\delta|x-m_\delta|^2/2}d\pi(x)$ is a normalizing constant and $m_\delta \in \mathbb{R}^d$ is as described in Proposition 5(i), for the choice of $\mu = \pi$. Note that π has density $C_\delta e^{\delta|x-m_\delta|^2/2}$ with respect to μ_δ. Therefore, we may compute

$$D(\pi\|\mu_\delta) = \frac{\delta}{2}\int_{\mathbb{R}^d}|x-m_\delta|^2 d\pi(x) + \log C_\delta = \frac{1}{2\delta}I(\pi\|\mu_\delta) + \log C_\delta.$$

By the Bakry–Émery theorem and the assumed k-uniform log-concavity of π, μ_δ satisfies LSI$(1/(k+\delta))$, so it follows from Theorem 2 that

$$\begin{aligned}
I(\pi; P_\theta) &\le -D(\pi\|\mu_\delta) + \frac{1}{2(k+\delta)}I(\pi\|\mu_\delta) + \frac{1}{2(k+\delta)}\int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta) \\
&= -\frac{k}{2\delta(k+\delta)}I(\pi\|\mu_\delta) + \frac{1}{2(k+\delta)}\int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta) - \log C_\delta \\
&= -\frac{k\delta}{2(k+\delta)}\int_{\mathbb{R}^d}|x-m_\delta|^2 d\pi(x) + \frac{1}{2(k+\delta)}\int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta) - \log C_\delta \\
&\le -\frac{k\delta}{2(k+\delta)}\operatorname{Tr}(\Sigma_\pi) + \frac{1}{2(k+\delta)}\int_{\mathbb{R}^d}\mathcal{I}_X(\theta)d\pi(\theta) - \log C_\delta.
\end{aligned}$$

Multiplying through by $2/d$ and invoking Proposition 5(iv), we conclude

$$\frac{2}{d}I(\pi; P_\theta) \le -\frac{k\delta}{(k+\delta)}P + \frac{1}{(k+\delta)}J + \begin{cases}\delta P & \text{if } 0 \le \delta < \frac{1}{P} - k,\\ 1 - kP + \log\left((\delta+k)P\right) & \text{if } \delta \ge \frac{1}{P} - k,\end{cases}$$

where J, P are as defined in the statement of the theorem. Now, define $a := kP$ and $b := JP$ for convenience, and we note that $a \le 1$ due to Lemma 9. Reparametrizing the above in terms of $t := \delta P$, we may state the inequality as

$$\frac{2}{d}I(\pi; P_\theta) \le \frac{b-at}{(a+t)} + \begin{cases}t & \text{if } 0 \le t < 1-a,\\ 1 - a + \log(a+t) & \text{if } t \ge 1-a,\end{cases} \tag{3.16}$$

Now, we optimize over the choice of $t \geq 0$. In particular, elementary calculus reveals that the best choice of t is given by

$$t = \begin{cases} \sqrt{a^2 + b} - a & \text{if } a^2 + b < 1 \\ b + a^2 - a & \text{if } a^2 + b \geq 1, \end{cases}$$

which gives

$$\frac{2}{d} I(\pi; P_\theta) \leq \begin{cases} 2(\sqrt{a^2 + b} - a) & \text{if } a^2 + b < 1, \\ 2(1 - a) + \log(a^2 + b) & \text{if } a^2 + b \geq 1. \end{cases}$$

Rearranging yields (3.10). □

On self-improvement of Theorem 7

We recall here that the linearized Efroimovich inequality (2.12) can be self-improved to the nonlinear form (2.7) by a rescaling argument. Unlike Theorem 4, we note that the RHS of the inequality stated in Theorem 7 is not invariant to linear transformations, while the LHS is. Hence, it can be self-improved similar to our remarks on Efroimovich's inequality. In particular, we have the following.

Theorem 11. *Fix a parametric model $(P_\theta)_{\theta \in \mathbb{R}^d}$ and prior $\pi \in \mathcal{P}(\mathbb{R}^d)$. If π is K-uniformly log-concave, then*

$$I(\pi; P_\theta) \leq \phi\left(\lambda_{\min}(B^{-1}(\Sigma_\pi^{1/2} K \Sigma_\pi^{1/2}))\tfrac{1}{d}\operatorname{Tr}(B), \tfrac{1}{d^2}\operatorname{Tr}(B)^2\right), \tag{3.17}$$

where $B := \left(\Sigma_\pi^{1/2} \mathbb{E}_\pi[\bar{\mathcal{I}}_X] \Sigma_\pi^{1/2}\right)^{1/2}$ and ϕ is as defined in Theorem 7.

Proof. Consider a nonsingular matrix V, let $\eta = V\theta$, and let π_0 denote the law of η. Note that η is $V^{-T}KV^{-1}$-uniformly log-concave and $\Sigma_{\pi_0} = V\Sigma_\pi V^T$. Recalling the reparametrization identity

$$\mathbb{E}_{\pi_0}[\bar{\mathcal{I}}_{X;\eta}] = V^{-T} \mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}] V^{-1},$$

an application of Theorem 7 gives, for $M := V^T V$,

$$\begin{aligned} &I(\pi; P_\theta) \\ &= I(\pi_0; P_\eta) \\ &\leq \phi\left(\lambda_{\min}(V^{-T}KV^{-1})\tfrac{1}{d}\operatorname{Tr}(V\Sigma_\pi V^T), \tfrac{1}{d}\operatorname{Tr}(V\Sigma_\pi V^T) d^{-1} \operatorname{Tr}(V^{-T}\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}]V^{-1})\right) \\ &= \phi\left(\lambda_{\min}(M^{-1}K)\tfrac{1}{d}\operatorname{Tr}(M\Sigma_\pi), \tfrac{1}{d}\operatorname{Tr}(M\Sigma_\pi)\tfrac{1}{d}\operatorname{Tr}(M^{-1}\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}])\right), \end{aligned}$$

where we used the cyclic properties of trace and $\lambda_{\min}$ (the latter for square matrices). Now, choose $M = \Sigma_\pi^{-1/2}\left(\Sigma_\pi^{1/2}\mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}]\Sigma_\pi^{1/2}\right)^{1/2}\Sigma_\pi^{-1/2}$ which gives the desired result. □

With B as defined in (3.17), it is relatively straightforward to show that

$$\operatorname{Tr}(B)^2 \leq \operatorname{Tr}(\Sigma_\pi) \operatorname{Tr}(\mathbb{E}_\pi[\bar{\mathcal{I}}_X]).$$

Moreover, in the case where $K = k\mathrm{I}_d$ (which is typical in applications), we have

$$\lambda_{\min}(B^{-1}(\Sigma_\pi^{1/2} K \Sigma_\pi^{1/2})) \operatorname{Tr}(B) \leq \operatorname{Tr}(\Sigma_\pi^{1/2} K \Sigma_\pi^{1/2}) = \lambda_{\min}(K) \operatorname{Tr}(\Sigma_\pi).$$

Hence, we may generally regard Theorem 11 as a self-improvement over Theorem 7.

We remark that the matrix B is invariant to affine transformations of the parameter θ. Indeed, if θ_0 is normalized so that $\operatorname{Cov}(\theta_0) = \mathrm{I}_d$ and we let π_0 denote its law, then for $\theta = \Sigma_\pi^{1/2}\theta_0$, we have

$$\mathbb{E}_{\pi_0}[\bar{\mathcal{I}}_{X;\theta_0}] = \Sigma_\pi^{1/2} \mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}] \Sigma_\pi^{1/2}$$

Hence, if $\eta \sim \nu$ is an affine transformation of θ, then we conclude

$$\Sigma_\eta^{1/2} \mathbb{E}_\nu[\bar{\mathcal{I}}_{X;\eta}] \Sigma_\eta^{1/2} = \Sigma_\pi^{1/2} \mathbb{E}_\pi[\bar{\mathcal{I}}_{X;\theta}] \Sigma_\pi^{1/2}.$$

Likewise, the matrix $\Sigma_\pi^{1/2} K \Sigma_\pi^{1/2}$ is invariant to affine transformations of the parameter in the following sense: π_0 is K_0-uniformly log-concave with $K_0 := \Sigma_\pi^{1/2} K \Sigma_\pi^{1/2}$ if and only if π is K-uniformly log-concave.

Chapter 4

Information-theoretic bounds on risk

In the previous chapters, we developed upper bounds on the mutual information between a parameter and an observation in terms of Fisher information terms and second moments. Chapter 2 presented an abstract framework that indexes such inequalities by measures that satisfy an LSI; the Efroimovich and van Trees inequalities are special cases when the reference measure is taken to be standard Gaussian. In Chapter 3, we showed how the abstract machinery can be applied to settings involving log-concave priors, resulting in upper bounds that maintain the spirit of the Efroimovich inequality, but have the advantage of avoiding degeneracy by eliminating dependence on the Fisher information of the prior.

We now turn our attention to statistical applications; namely, bounding Bayes and minimax risk. This chapter briefly introduces important definitions and concepts, and reviews standard methods for how bounds on mutual information can be employed to provide bounds on risk. For further background, readers are referred to the references [Wainwright, 2019, Cover and Thomas, 2012, Berger, 2003].

4.1 Definition of Bayes and minimax risk

As remarked in Chapter 2, the ultimate goal of parameter estimation is to recover the latent parameter θ given the observation X. We can write any implementable function, algorithm or procedure that takes in X and gives an estimate as the estimator $\hat{\theta} : \mathcal{X} \to \mathbb{R}^d$. The central question to quantitatively address is: *how well does $\hat{\theta}$ estimate θ?* A standard way of answering this question is by introducing the **risk function** R, defined in terms of a loss function $\mathcal{L} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the estimator $\hat{\theta}$, and a given parameter θ via

$$R(\hat{\theta}, \theta) := \mathbb{E}_{P_\theta}[\mathcal{L}(\hat{\theta}(X), \theta)]. \tag{4.1}$$

If we have prior knowledge $\theta \sim \pi$, then we have the corresponding **Bayes risk for the estimator** $\hat{\theta}$ given by

$$R_\pi(\hat{\theta}) := \mathbb{E}_\pi[R(\hat{\theta}, \theta)] = \mathbb{E}[\mathcal{L}(\hat{\theta}(X), \theta)]. \tag{4.2}$$

Most commonly, the loss function is defined to satisfy $\mathcal{L}(\theta', \theta) \geq 0$, with equality when $\theta' = \theta$; i.e., the loss is minimized when the estimator $\hat{\theta}$ recovers θ correctly. It is easy to see why the risk defined in (4.1) is suitable for a wide variety of situations. For one, it models our intent precisely; the loss function $\mathcal{L}(\cdot, \cdot)$ can be thought of as a penalty function, and the risk $R(\hat{\theta}, \theta)$ measures the expected penalty incurred using $\hat{\theta}$ to estimate θ based on the specified loss $\mathcal{L}(\cdot, \cdot)$. Given that the loss function is left unspecified, it allows the statistician or the engineer to choose a suitable one for their purpose.

Given that the risk measures the average penalty incurred, a second question that follows immediately is: *how do we know if the estimator $\hat{\theta}$ is good relative to other possible choices?* This is where lower bounds on risk comes into play. From a philosophical standpoint, if we can show that *any* estimator $\hat{\theta}$ satisfies

$$R_\pi(\hat{\theta}) \geq A \quad \text{for all } \hat{\theta} : \mathcal{X} \to \mathbb{R}^d, \tag{4.3}$$

for some number $A \geq 0$, then given a particular estimator $\hat{\theta}$, upon calculating the corresponding risk $R_\pi(\hat{\theta})$, we can compare it with the value A to get an understanding of how well $\hat{\theta}$ might be. If the risk is precisely equal to A, then we know that the estimator $\hat{\theta}$ is the best possible, and there is little we can hope do to improve it (at least, in terms of the risk associated with $\mathcal{L}$). On the contrary, if the risk is much larger than A, then it is either one of two reasons: (i) a better estimator is possible; or, (ii) the lower bound is too loose.

Now, let us define the **Bayes risk** of estimating $\theta \sim \pi$ from $X \sim P_\theta$, as the smallest risk achievable over all estimators

$$R^*_\pi := \inf_{\hat{\theta}} R_\pi(\hat{\theta}).$$

Naturally, when the risk of an estimator matches a given lower bound on the Bayes risk lower bound (say, a lower bound of A as in (4.3)) with equality, it immediately follows that the Bayes risk corresponding to the prior π is exactly equal to A.

By definition, Bayes risk only measures the penalty incurred on average. A related concept is the **minimax risk**, which is defined as

$$R^* := \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_{P_\theta}[\mathcal{L}(\hat{\theta}(X), \theta)],$$

where $\Theta \subseteq \mathbb{R}^d$ is a given set of admissible parameters. By definition, the minimax risk quantifies the performance of the estimator that minimizes the worst-case penalty incurred under $\mathcal{L}$. Bayes risk and minimax risk are well-known to be related as follows.

Theorem 12. *If $\pi \in \mathcal{P}(\mathbb{R}^d)$ is a prior with support contained in $\Theta \subseteq \mathbb{R}^d$, then*

$$R^* \geq R^*_\pi. \tag{4.4}$$

Thus, any lower bound on Bayes risk implies the same lower bound on minimax risk.

Proof. The proof is immediate from definitions:

$$R^*_\pi = \inf_{\hat{\theta}} \mathbb{E}_\pi[R(\hat{\theta}, \theta)] \leq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R(\hat{\theta}, \theta) = R^*,$$

where the inequality follows since π is supported on Θ. □

In other words, the above simply states that lower bounds on minimax risk bounds can be derived from lower bounds on Bayes risk bounds by optimizing over choices of the prior π.

In the non-asymptotic regime, i.e., where d and the dimension of $\mathcal{X}$ are not necessarily taken to infinity, it is of significant interest to understand Bayes risk and minimax risk in terms of the model parameters. In these cases, the primary focus is on characterizing the dependence of risk on model parameters, and universal pre-constants are generally disregarded. We remark that in the asymptotic regime, it is possible to derive sharp bounds on minimax risk using the theory of asymptotic minimax, see, e.g., [Hájek, 1972, Van der Vaart, 2000].

In our applications, we will focus on applying bounds given in Chapter 3 to give lower bounds on Bayes and minimax risk lower bounds in the non-asymptotic regime. This Chapter and the next two are organized in the following fashion:

1. In this Chapter, we will present general procedures of deriving lower bounds on Bayes risk given upper bounds on mutual information $I(\pi; P_\theta)$. Tools from rate distortion theory will be extensively utilized. We will examine applications to the Gaussian location model and the spiked covariance model.

2. In Chapter 5 we will present applications to Generalized Linear Models (GLMs). An array of lower bounds on Bayes and minimax risk under different settings will be discussed.

3. Applications to pairwise comparison models are found in Chapter 6. There, we will introduce a general pairwise comparison framework called the General Pairwise Model (GPM), which includes as special cases popular models such as the Thurstone (Case V) Model, Bradley–Terry–Luce (BTL) model and the Ordinal model. We will show how our techniques can be applied to yield uniform minimax lower bounds under the GPM that hold uniformly over the broad class of pairwise comparison models we consider.

4.2 Rate distortion theory

For a random variable $\theta \sim \pi \in \mathcal{P}(\mathbb{R}^d)$ and a loss function $\mathcal{L} : \hat{\Theta} \times \mathbb{R}^d \to [0, +\infty)$, the **rate-distortion function** from the theory of lossy compression is defined as

$$\mathbf{R}(D; \pi, \mathcal{L}) := \inf_{Q_{\hat{\theta}|\theta} : \mathbb{E}[\mathcal{L}(\hat{\theta}, \theta)] \leq D} I(\hat{\theta}; \theta).$$

In the optimization problem above, the infimum is over all random variables $\hat{\theta} \in \hat{\Theta}$ jointly distributed with $\theta \sim \pi$ such that $\mathbb{E}[\mathcal{L}(\hat{\theta}, \theta)] \leq D$, where expectation is with respect to the

joint law. It is well known that the function $D \mapsto \mathbf{R}(D; \pi, \mathcal{L})$ is convex and decreasing. Moreover, the optimization problem defining $\mathbf{R}(D; \pi, \mathcal{L})$ is a convex program, and can therefore be solved in principle; see, e.g., [Blahut, 1972, Arimoto, 1972, Chiang and Boyd, 2004] for more background and discussions. Closed-form results are known in a few special cases.

Example 20 (Gaussian source and squared loss)**.** *When $d = 1$, $\pi = N(0, \sigma^2)$, $\hat{\Theta} = \mathbb{R}$ and $\mathcal{L}(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, then*

$$\mathbf{R}(D; \pi, \mathcal{L}) = \frac{1}{2} \log \left(\frac{\sigma^2}{D} \vee 1 \right), \qquad D \geq 0.$$

Example 21 (Exponential source and one-sided loss)**.** *When $d = 1$, $\pi = \text{Exponential}(\lambda)$, $\hat{\Theta} = \mathbb{R}$ and*

$$\mathcal{L}(\hat{\theta}, \theta) = \begin{cases} \hat{\theta} - \theta & \text{if } \hat{\theta} \geq \theta \\ \infty & \text{otherwise,} \end{cases}$$

then

$$\mathbf{R}(D; \pi, \mathcal{L}) = \log \left(\frac{1}{\lambda D} \vee 1 \right), \qquad D \geq 0.$$

There are many other loss functions one can consider. For example, Adler et al. [2021] considers $\mathcal{L}(\hat{\theta}, \theta) = \theta \log(\theta/\hat{\theta})$ and presents lower bounds on the rate distortion function with applications to quantization of random distributions. Tan and Yao [1975] provides explicit evaluations of the rate distortion function under $\mathcal{L}(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ for a large class of sources. Courtade and Wesel [2011] consider a entropy-based distortion measure and Courtade and Weissman [2013] consider a log-loss.

Example 22 (Entropy loss)**.** *For a density p on $\mathbb{R}^d$, we can define the entropy loss as $\mathcal{L}(p, \theta) = -\log p(\theta)$. Given a prior π admitting a density with respect to Lebesgue measure, we have*

$$\mathbf{R}(D; \pi, \mathcal{L}) = h(\theta) - (D \wedge h(\theta)), \qquad D \in \mathbb{R}.$$

Since the rate distortion function is always a decreasing function of the argument D, and for any estimator $\hat{\theta}$ we have the Markov relation

$$\theta \to X \to \hat{\theta}(X),$$

definitions and the data processing inequality for mutual information directly imply

$$\mathbf{R}(R^*_\pi; \pi, \mathcal{L}) \leq \mathbf{R}(R_\pi(\hat{\theta}); \pi, \mathcal{L}) \leq I(\theta; \hat{\theta}(X)) \leq I(\theta; X)$$

for any parametric model with parameter $\theta \sim \pi$ and observation $X \sim P_\theta$. Since $\mathbf{R}(\cdot; \pi, \mathcal{L})$ is decreasing in its argument, we can combine the above with any upper bound on the mutual information to obtain a lower bound on the Bayes risk R^*_π. For example, we have the following immediate corollary of Theorem 4.

Proposition 6. *Fix a parametric model $(P_\theta)_{\theta\in\mathbb{R}^d}$ and a K-uniformly log-concave prior $\pi \in \mathcal{P}(\mathbb{R}^d)$. For a given loss function $\mathcal{L} : \hat{\Theta} \times \mathbb{R}^d \to [0,\infty]$, we have*

$$\mathbf{R}(R^*_\pi; \pi, \mathcal{L}) \leq \frac{1}{2} \log\det\left(I_d + \Sigma_\pi \mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right) + \mathrm{Tr}\left((I_d + K^{1/2}(\mathbb{E}_\pi[\bar{\mathcal{I}}_X])^{-1}K^{1/2})^{-1}(I_d - K^{1/2}\Sigma_\pi K^{1/2})\right).$$

Making use of the rate distortion function for a Gaussian source under squared loss, a simple illustration of the above is the following:

Example 23. *Let $d = 1$, $\pi = N(0,\sigma^2)$ and $\mathcal{L}(\hat{\theta},\theta) = (\hat{\theta}-\theta)^2$. For an observation model $(P_\theta)_{\theta\in\mathbb{R}}$, we have*

$$R^*_\pi \geq \frac{1}{\sigma^{-2} + \mathbb{E}_\pi[\mathcal{I}_X]}.$$

The above is an immediate consequence of the van Trees inequality. However, with Proposition 6, we can easily go beyond where the van Trees inequality can take us, provided we assume a log-concave prior. Indeed, the exponential distribution has infinite Fisher information. Nevertheless, we can immediately deduce the following.

Example 24. *Let $d = 1$, $\pi = \mathrm{Exponential}(\lambda)$, $\hat{\Theta} = \mathbb{R}$ and*

$$\mathcal{L}(\hat{\theta},\theta) = \begin{cases} \hat{\theta} - \theta & \text{if } \hat{\theta} \geq \theta \\ \infty & \text{otherwise.} \end{cases}$$

For an observation model $(P_\theta)_{\theta\in\mathbb{R}^d}$, we have

$$R^*_\pi \geq \frac{e^{-1}}{(\lambda^2 + \mathbb{E}_\pi[\mathcal{I}_X])^{1/2}}.$$

Remark 10. *The bound in the above example can be improved slightly by invoking Theorem 7 instead of Theorem 4, which was used in stating Proposition 6. Indeed, in this case we would find*

$$R^*_\pi \geq \begin{cases} \lambda^{-1}\exp\left(-\lambda^{-1}\sqrt{\mathbb{E}_\pi[\mathcal{I}_X]}\right) & \text{if } \mathbb{E}_\pi[\mathcal{I}_X] < \lambda^2 \\ \frac{e^{-1}}{(\mathbb{E}_\pi[\mathcal{I}_X])^{1/2}} & \text{if } \mathbb{E}_\pi[\mathcal{I}_X] \geq \lambda^2. \end{cases}$$

Additional remarks on entropy risk and universal source coding

The relative entropy $\int_{\mathbb{R}^d} D(P_\theta\|q)d\pi(\theta)$ associated with a density $q \in \mathcal{P}(\mathcal{X})$ is called the entropy risk [Clarke and Barron, 1994], and is minimized by

$$q^* = \int_{\mathbb{R}^d} P_\theta d\pi(\theta), \tag{4.5}$$

which is known as the Bayes strategy [Aitchison, 1975, Clarke and Barron, 1994]. Interestingly, the entropy risk of the Bayes strategy is equal to the mutual information $I(\pi; P_\theta)$, i.e.,

$$I(\pi; P_\theta) = \int_{\mathbb{R}^d} D(P_\theta \| q^*) d\pi(\theta).$$

This therefore relates entropy risk to the entropy loss $\mathcal{L}$ defined in Example 22 by noting that

$$\lim_{D\to\infty} \inf_{Q_{\hat{\theta}|X}: \mathbb{E}[\mathcal{L}(\hat{\theta},\theta)] \le D} I(\hat{\theta}; \theta) = I(\pi; P_\theta).$$

Entropy risk determines the Bayes redundancy of universal noiseless source codes. In particular, given a code $\{\Phi\}$ with codelengths $\ell(\Phi(X))$, the Bayes redundancy of the code is $\int_{\mathbb{R}^d} \left(\mathbb{E}\ell(\Phi(X)) - h(P_\theta)\right) d\pi(\theta) = \int_{\mathbb{R}^d} D(P_\theta \| q) d\pi(\theta)$, where q is the density with $q(X) = 2^{-\ell(\Phi(X))}$ [Clarke and Barron, 1994]. Therefore, the choice q^* in (4.5) characterizes the code with minimal Bayes redundancy, which is precisely $I(\pi; P_\theta)$, for which Theorems 4 and 7 present useful upper bounds. See also [Davisson, 1973, Willems et al., 1995, Welch, 1984, Ziv and Lempel, 1978] for more discussions of universal noiseless source coding and algorithms.

4.3 The Shannon lower bound

As mentioned earlier, rate distortion functions typically do not have a closed form solution except in some special cases. Luckily, for our purpose of deriving lower bounds on risk, the Shannon lower bound [Shannon, 1959] provides a lower bound on the rate distortion function that can be translated into a lower bound on risk. The following formulation of the Shannon lower bound can be found in Wu [2017, Theorem 13.1]. The proof is elementary, following from basic information theory identities and a maximum entropy result due to Yamada et al. [1980].

Theorem 13 (Shannon lower bound)**.** *Suppose $\theta \in \mathbb{R}^d$ has a prior $\pi \in \mathcal{P}(\mathbb{R}^d)$, and suppose $X \sim P_\theta$. Suppose $\|\cdot\|$ is any norm defined on $\mathbb{R}^d$ and $r > 0$ is a given fixed parameter. Then, for any $D \ge 0$, the following holds under the loss $\mathcal{L}(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^r$.*

$$\mathbf{R}(D; \pi, \mathcal{L}) \ge h(\theta) - \log\left(V \left(\frac{Dre}{d}\right)^{\frac{d}{r}} \Gamma\left(1 + \frac{d}{r}\right)\right),$$

where V is the volume of the unit ball with respect to the norm $\|\cdot\|$.

Here, $\Gamma(\cdot) : \mathbb{R} \longrightarrow \mathbb{R}$ is the Gamma function defined by Legendre (see [Davis, 1959] for an interesting description of the history of the Gamma function), defined as

$$\Gamma(u) = \int_0^\infty x^{u-1} e^{-x} dx \quad u \in \mathbb{R}^+.$$

When u is an integer, $\Gamma(u) = (u-1)!$ corresponds to the factorial function of $u - 1$.

Proof of Theorem 13. The proof begins with the elementary operations (see [Wu, 2017])

$$\begin{aligned}\mathbf{R}(D;\pi,\mathcal{L}) &= \inf_{Q_{\hat{\theta}|\theta}:\mathbb{E}\|\hat{\theta}-\theta\|^r\le D} I(\hat{\theta};\theta)\\ &= h(\theta) - \sup_{Q_{\hat{\theta}|\theta}:\mathbb{E}\|\hat{\theta}-\theta\|^r\le D} h(\theta|\hat{\theta})\\ &\ge h(\theta) - \sup_{Q_{\hat{\theta}|\theta}:\mathbb{E}\|\hat{\theta}-\theta\|^r\le D} h(\hat{\theta}-\theta),\end{aligned} \tag{4.6}$$

where the final inequality follows from conditioning reduces entropy. The last quantity is precisely about the maximum entropy distribution under a difference distortion measure, for which Yamada et al. [1980] gives the following explicit formula.

Lemma 14 (Shannon lower bound under a difference distortion measure [Yamada et al., 1980])**.** *Consider $\|\cdot\|$ as any norm. Then,*

$$h(\theta) - \sup_{Q_{\hat{\theta}|\theta}:\mathbb{E}\|\hat{\theta}-\theta\|^r\le D} h(\hat{\theta}-\theta) = h(\theta) - \frac{d}{r}\log\frac{D}{d} - \log V - \frac{d}{r}\log\left(er\Gamma\left(1+\frac{d}{r}\right)^{\frac{r}{d}}\right)$$

where V is the volume of the unit ball under the norm $\|\cdot\|$.

Combining (4.6) with Lemma 14 and rearranging yields our desired result.

□

The Shannon lower bound is asymptotically tight as $D\to 0$ under certain conditions [Koch, 2016, Linder and Zamir, 1994, Binia et al., 1974, Gerrish and Schultheiss, 1964]. Finite-blocklength refinements for the Shannon lower bound are available in [Kostina, 2017, 2016]. The Shannon lower bound is also useful as a benchmark for quantization; see, e.g., [Gray et al., 2002, Gish and Pierce, 1968, Zador, 1964]

From Theorem 13 we derive the following theorem, which provides a systematic way of lower bounding Bayes risk under any norm.

Theorem 15. *Suppose $\theta\in\mathbb{R}^d$ has a prior $\pi\in\mathcal{P}(\mathbb{R}^d)$, and suppose $X\sim P_\theta$. Suppose $\|\cdot\|$ is any norm defined on $\mathbb{R}^d$ and $r>0$ is a given fixed parameter. Then, the following holds for any estimator $\hat{\theta}:\mathcal{X}\to\mathbb{R}^d$,*

$$\mathbb{E}\|\hat{\theta}-\theta\|^r \ge \frac{d}{re}\left(V\Gamma\left(1+\frac{d}{r}\right)\right)^{-\frac{r}{d}}\exp\left(\frac{r}{d}\left(h(\theta)-I(\pi;P_\theta)\right)\right), \tag{4.7}$$

where V is the volume of the unit ball with respect to the norm $\|\cdot\|$.

Proof. Given a particular estimator $\hat{\theta}$, Theorem 13 gives under the setting $D=\mathbb{E}\|\hat{\theta}-\theta\|^r$,

$$I(\theta;\hat{\theta}) \ge h(\theta) - \log\left(V\left(\frac{\mathbb{E}\|\hat{\theta}-\theta\|^r re}{d}\right)^{\frac{d}{r}}\Gamma\left(1+\frac{d}{r}\right)\right). \tag{4.8}$$

We can apply the data processing inequality on the Markov chain $\theta \to X \to \hat{\theta}$ and use the property $I(\pi; P_\theta) = h(\theta) - h(\theta|X)$ within (4.8) to write

$$h(\theta) - I(\pi; P_\theta) \leq \log \left(V \left(\frac{\mathbb{E}\|\hat{\theta} - \theta\|^r re}{d} \right)^{\frac{d}{r}} \Gamma\left(1 + \frac{d}{r}\right) \right).$$

Then, a rearrangement recovers (4.7) as desired. □

The lower bound on Bayes risk given in Theorem 15 is particularly useful when ℓ_p norms are considered, since the volume V_p of unit ℓ_p balls in $\mathbb{R}^d$ are well known (see, for example, the work of Wang [2005] for an interesting derivation of the volume of generalized unit balls), given by

$$V_p := \frac{\left(\Gamma\left(\frac{1}{p} + 1\right)\right)^d}{\Gamma\left(\frac{d}{p} + 1\right)} 2^d. \tag{4.9}$$

This allows us to write the following corollary, which will be useful for later developments.

Corollary 16 (Lower bound on Bayes risk for the ℓ_p norm)**.** *Suppose $\theta \in \mathbb{R}^d$ has a prior $\pi \in \mathcal{P}(\mathbb{R}^d)$, and suppose $X \sim P_\theta$. Fix $p \geq 1$ and any $r > 0$. Then, for any estimator $\hat{\theta}$, the following holds for the ℓ_p norm.*

$$\mathbb{E}\|\hat{\theta} - \theta\|_p^r \geq \frac{d}{re} \left(2\Gamma\left(\frac{1}{p} + 1\right)\right)^{-r} \exp\left(\frac{r}{d}\left(h(\theta) - I(\pi; P_\theta)\right)\right). \tag{4.10}$$

Next, we provide two examples: estimation under a Gaussian observation model and under a spiked covariance model, both with a uniform prior π.

Gaussian observation model under a uniform prior

Suppose π is a uniform prior over the ℓ_2 ball with a fixed radius $R \geq 0$, and observation X is generated according to the Gaussian $\mathcal{N}(\theta, \sigma^2 \mathrm{I}_d)$. We will establish the following lower bound on Bayes risk.

Proposition 7. *Suppose $X \sim \mathcal{N}(\theta, \sigma^2 I_d)$ and π is a uniform prior over the ℓ_2 ball in $\mathbb{R}^d$ with a fixed radius $R \geq 0$. Then, the following holds for any estimator $\hat{\theta} : \mathbb{R}^d \to \mathbb{R}^d$,*

$$\mathbb{E}|\hat{\theta} - \theta|^2 \gtrsim \min\left(R^2, \sigma^2 d\right), \tag{4.11}$$

The Bayes risk derived in (4.11) is tight up to constants, and can also be derived with a Bayes variation of Fano's inequality, see, e.g., [Chen et al., 2016], but is quite involved compared to our proof based on information measures (see, e.g., Examples 1 and 2 of [Chen et al., 2016]).

Proof of Proposition 7. We begin with a reminder that the entropy of a uniform distribution over a convex set having volume V is $\log(V)$ (e.g., Example 5). For ease of expressing our results, let us define $V_2(r)$ as the volume of the ℓ_2 ball with radius $r \geq 0$ and write

$$h(\theta) = \log(V_2(R)) = \log\left(\frac{\pi^{\frac{d}{2}} R^d}{\Gamma\left(\frac{d}{2}+1\right)}\right), \tag{4.12}$$

where the formula (4.9) for the volume of the ℓ_2 ball is invoked.

We may derive the covariance matrix Σ_π with the following identities

$$\begin{aligned}
\mathbb{E}\theta_i^2 &= \frac{1}{d}\mathbb{E}|\theta|^2 \\
&= \frac{1}{d}\int_0^R r^2 \left(\frac{\frac{d}{dr}V_2(r)}{\int_0^R \frac{d}{dt}V_2(t)dt}\right) dr \\
&= \frac{1}{d}\int_0^R r^2 \left(\frac{dr^{d-1}}{R^d}\right) dr \\
&= \frac{1}{R^d}\int_0^R r^{d+1} dr \\
&= \frac{R^2}{d+2} \qquad i = 1, 2, \ldots, d,
\end{aligned}$$

where the second equality follows by a spherical integration, and

$$\mathbb{E}\theta_i\theta_j = 0 \qquad i \neq j,$$

since the density of θ is symmetric with respect to the origin. It therefore follows that

$$\Sigma_\pi = \frac{R^2}{d+2}\mathrm{I}_d. \tag{4.13}$$

We are now in position to derive a lower bound for ℓ_2 Bayes risk by invoking Corollary 16 and Theorem 7. For any estimator $\hat{\theta} : \mathcal{X} \longrightarrow \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}|\hat{\theta} - \theta|^2 &\geq \frac{1}{2\pi e} e^{\frac{2}{d}(h(\theta) - I(\pi; P_\theta))} \\
&\geq \frac{d}{2\pi e} e^{\frac{2}{d}h(\theta) - \frac{2}{d}\phi\left(0, \frac{1}{d^2}\operatorname{Tr}(\Sigma_\pi)\mathbb{E}_\pi[\mathcal{I}_X]\right)} \\
&= \frac{d}{2e}\left(\Gamma\left(\frac{d}{2}+1\right)\right)^{-\frac{2}{d}} R^2 e^{-\frac{2}{d}\phi\left(0, \frac{R^2}{(d+2)\sigma^2}\right)}.
\end{aligned} \tag{4.14}$$

Here we used (4.12) and (4.13) in the final equality. The expected Fisher information $\mathbb{E}_\pi[\mathcal{I}_X] = d/\sigma^2$ is a straightforward calculation (see, e.g., Example 9).

Moving forward, it remains to lower bound the quantity in the right hand side of (4.14). Here, note that the function $g(\cdot) : \mathbb{Z}^+ \to \mathbb{R}$ defined over non-negative integers as

$$g(d) := d\Gamma\left(\frac{d}{2}+1\right)^{-\frac{2}{d}}$$

is increasing in d and satisfies $g(1) = \frac{4}{\pi}$, $\lim_{d\to\infty} g(d) = 2e$. Therefore, it follows that

$$\mathbb{E}|\hat{\theta}-\theta|^2 \gtrsim R^2 e^{-\frac{2}{d}\phi\left(0,\frac{R^2}{(d+2)\sigma^2}\right)} \gtrsim \min\left(R^2, \sigma^2 d\right).$$

The last inequality follows directly from the definition of $\phi(\cdot)$.

□

It should be noted that Efroimovich's inequality (and the classic van Trees inequality) cannot be applied directly to yield a Bayes risk bound in this scenario, since $\mathcal{J}(\pi)$ is infinite for a uniform prior.

Remark 11. *A Bayes risk lower bound for the setting where $X = (X_1, \ldots, X_n)$ are n i.i.d. samples of $\mathcal{N}(\theta, \sigma^2 I_d)$ can be attained with the same approach by noting $\mathbb{E}_\pi[\mathcal{I}_X] = nd/\sigma^2$, which immediately gives*

$$\mathbb{E}|\hat{\theta}-\theta|^2 \gtrsim \min\left(R^2, \frac{d\sigma^2}{n}\right).$$

Spiked covariance model under a uniform prior

Suppose π is a uniform prior over the unit ℓ_2 ball and $X = (X_1, X_2, \ldots, X_n)$ are generated i.i.d. with $X_i \sim \mathcal{N}\left(0, \mathrm{I}_d + \theta\theta^T\right), i = 1, \ldots, n$. We establish the following lower bound on Bayes risk.

Proposition 8. *Suppose π is a uniform prior over the unit ℓ_2 ball in $\mathbb{R}^d$. Suppose $X = (X_1, \ldots, X_n)$ are generated i.i.d. with $X_i \sim \mathcal{N}\left(0, I_d + \theta\theta^T\right), i = 1, \ldots, n$. Then, for any estimator $\hat{\theta} : \mathbb{R}^d \to \mathbb{R}^d$,*

$$\mathbb{E}\|\hat{\theta}-\theta\|_p^r \geq C(r) \min\left(1, \frac{d}{n}\right)^{\frac{r}{2}},$$

where $C(r)$ is a constant determined by r only.

A proof using a Bayes variation of Fano's method is available in Section C.2. of Chen et al. [2016].

Proof of Proposition 8. We begin our proof by noting that for any estimator $\hat{\theta} : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ and any fixed $r > 0$, (4.10) gives

$$\begin{aligned}\mathbb{E}\|\hat{\theta} - \theta\|_p^r &\geq \frac{d}{re}\left(V_2\Gamma\left(1+\frac{d}{r}\right)\right)^{-\frac{r}{d}} \exp\left(\frac{r}{d}\left(h(\theta) - I(\pi; P_\theta)\right)\right) \\ &= \frac{d}{re}\left(\Gamma\left(1+\frac{d}{r}\right)\right)^{-\frac{r}{d}} \exp\left(-\frac{r}{d}I(\pi; P_\theta)\right) \\ &= C_1(r)\exp\left(-\frac{r}{d}I(\pi; P_\theta)\right). \end{aligned} \tag{4.15}$$

where we recall that the entropy of θ satisfies $h(\theta) = \log V_2$. $C_1(r)$ is a constant determined by r only. The following lemma provides an upper bound on $I(\pi; P_\theta)$; the proof is deferred to a later section.

Lemma 17. *Under the same setting as Proposition 8,*

$$I(\pi; P_\theta) \leq \phi\left(0, \frac{24n}{d}\right).$$

We can then combine this with (4.15) and Theorem 7 to get

$$\begin{aligned}\mathbb{E}\|\hat{\theta} - \theta\|_p^r &\geq C_1(r)\exp\left(-\frac{r}{d}\phi\left(0, \frac{24n}{d}\right)\right) \\ &\geq C_2(r)\min\left(1, \frac{d}{n}\right)^{\frac{r}{2}}.\end{aligned}$$

where $C_2(r)$ is a constant determined by r only. Therefore, we recover a simple, straightforward proof of a lower bound on Bayes risk of the spiked covariance model under a uniform prior on the ℓ_2 ball. We again remark that a similar bound cannot be derived using either Efroimovich's inequality or the van Trees inequality since the Fisher information of the prior is infinity (or, undefined) in this case. □

Proof of Lemma 17

We bound $I(\pi; P_\theta)$ by making use of Theorem 7, for which we need two ingredients: the covariance matrix Σ_π and the expected Fisher information $\mathbb{E}_\pi[\mathcal{I}_X]$. Recall that the covariance matrix Σ_π of a uniform prior π over the unit ℓ_2 ball is given in (4.13) by

$$\Sigma_\pi = \frac{1}{d+2}\mathrm{I}_d.$$

The expected Fisher information $\mathbb{E}_\pi[\mathcal{I}_X]$ is a bit trickier. To begin, let us first assume that $n = 1$; then, since $\mathbb{E}_\pi[\mathcal{I}_X] = n\mathbb{E}_\pi[\mathcal{I}_{X_1}]$, we can directly infer the Fisher information for a

general $n > 1$. Let us start with the following calculations.

$$\begin{aligned}\nabla_\theta \log f(x;\theta) &= -\frac{1}{2}\nabla_\theta \log\det(\mathrm{I}_d + \theta\theta^T) - \frac{1}{2}\nabla_\theta x^T(\mathrm{I}_d + \theta\theta^T)^{-1}x \\ &= -\frac{1}{2}\nabla_\theta \log(1+|\theta|^2) - \frac{1}{2}\nabla_\theta x^T\left(\mathrm{I}_d - \frac{\theta\theta^T}{1+|\theta|^2}\right)x \\ &= -\frac{\theta}{1+|\theta|^2} + \frac{xx^T\theta}{1+|\theta|^2} - \frac{(x^T\theta\theta^T x)\theta}{(1+|\theta|^2)^2}.\end{aligned}$$

Here, the middle equality follows from the Sherman-Morrison formula (see, e.g., [Sherman and Morrison, 1950]), which states that

$$(\mathrm{I}_d + \theta\theta^T)^{-1} = \mathrm{I}_d - \frac{\theta\theta^T}{1+|\theta|^2}.$$

It is a short exercise to show that this is indeed true by multiplying both sides by $(\mathrm{I}_d + \theta\theta^T)$. Moving onwards, we can then compute the expected Fisher information as

$$\begin{aligned}&\mathbb{E}_\pi[\mathcal{I}_X] \\ &\quad= \mathbb{E}|\nabla_\theta \log f(X;\theta)|^2 \\ &\quad= \mathbb{E}\left|-\frac{2\theta}{1+|\theta|^2} + \frac{XX^T\theta}{1+|\theta|^2} + \frac{(X^T\theta\theta^T X)\theta}{(1+|\theta|^2)^2}\right|^2 \\ &\quad\le 3\left(\mathbb{E}\left[\left|-\frac{2\theta}{1+|\theta|^2}\right|^2\right] + \mathbb{E}\left|\frac{XX^T\theta}{1+|\theta|^2}\right|^2 + \mathbb{E}\left|\frac{(X^T\theta\theta^T X)\theta}{(1+|\theta|^2)^2}\right|^2\right). \qquad (4.16)\end{aligned}$$

We can bound the terms one by one. The first term, since $|\theta|^2 \le 1$ by recalling that θ is sampled within the unit ℓ_2 ball, satisfies

$$\mathbb{E}\left[\left|-\frac{2\theta}{1+|\theta|^2}\right|^2\right] \le 4. \qquad (4.17)$$

The second term can be bounded by

$$\mathbb{E}\left|\frac{XX^T\theta}{1+|\theta|^2}\right|^2 \le \mathbb{E}\left[\mathbb{E}\left[\left|XX^T\theta\right|^2\middle|\theta\right]\right]. \qquad (4.18)$$

Given a fixed θ, we can write $X = Z_1 + \theta Z_2$, with $Z_1 \in \mathcal{N}(0, \mathrm{I}_d)$ and $Z_2 \in \mathcal{N}(0,1)$, both independent Gaussians. Then, it follows that

$$\begin{aligned}\mathbb{E}[|XX^T\theta|^2|\theta] &= \mathbb{E}[\left|(Z_1+\theta Z_2)(Z_1^T\theta + Z_2\theta^T\theta)\right|^2|\theta] \\ &= \mathbb{E}[\theta^T Z_1 Z_1^T Z_1 Z_1^T\theta|\theta] + \mathbb{E}[|Z_1|^2 Z_2^2|\theta|^4|\theta] + \mathbb{E}[Z_2^2|Z_1^T\theta|^2|\theta|^2|\theta] \\ &\quad + \mathbb{E}[Z_2^4|\theta|^6|\theta] + 4\mathbb{E}[(Z_1^T\theta)^2|\theta|^2 Z_2^2|\theta].\end{aligned}$$

Note that in the last equation we only considered terms with even powers in Z_1 and Z_2, since Z_1 and Z_2 are independent with mean 0_d and 0 respectively. Next, since $\mathbb{E}Z_2^2 = 1$, $\mathbb{E}Z_2^4 = \frac{1}{6}$, $\mathbb{E}|Z_1|^2 = d$, and $|\theta|^2 \leq 1$, we can bound

$$\begin{aligned}
\mathbb{E}[|XX^T\theta|^2|\theta] &\leq \mathbb{E}[\theta^T Z_1 Z_1^T Z_1 Z_1^T \theta|\theta] + 7d|\theta|^2 \\
&= \mathbb{E}[\left(|Z_1|^2 (Z_1^T\theta)^2\right)|\theta] + 7d|\theta|^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d Z_{1i}^2\right)\left(\sum_{j=1}^d Z_{1j}\theta_j\right)^2 \middle| \theta\right] + 7d|\theta|^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^d Z_{1i}^2\right)\left(\sum_{j=1}^d Z_{1j}^2\theta_j^2\right) \middle| \theta\right] + 7d|\theta|^2 \\
&= \sum_{j=1}^d \left(\theta_j^2 \sum_{i=1}^d \mathbb{E}Z_{1i}^2 Z_{1j}^2\right) + 7d|\theta|^2 \\
&= \sum_{j=1}^d \theta_j^2 (\mathbb{E}Z_{1j}^4 + (d-1)(\mathbb{E}Z_{1j}^2)^2) + 7d|\theta|^2 \\
&= \sum_{j=1}^d \theta_j^2 \left(\frac{1}{6} + (d-1)\right) + 7d|\theta|^2 \\
&\leq 8d|\theta|^2.
\end{aligned}$$

Note that the constants here, although not in the tightest form possible, are sufficient for our purposes. Moving onwards, we can then combine this with (4.18) to get

$$\mathbb{E}\left|\frac{XX^T\theta}{1+|\theta|^2}\right|^2 \leq 8d\mathbb{E}|\theta|^2 \leq 8d. \tag{4.19}$$

Finally, it remains to bound the third term on the right hand side of (4.16),

$$\begin{aligned}
\mathbb{E}\left|\frac{(X^T\theta\theta^T X)\theta}{(1+|\theta|^2)^2}\right|^2 &\leq \mathbb{E}(\theta^T X)^2|\theta|^2 \\
&\leq \mathbb{E}\left[\mathbb{E}[(\theta^T X)^2|\theta]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\,\mathrm{Tr}\left(\theta^T(I+\theta\theta^T)\theta\right)\right] \\
&= \mathbb{E}|\theta|^2 + \mathbb{E}|\theta|^4 \\
&\leq 2.
\end{aligned} \tag{4.20}$$

Combining (4.16), (4.17), (4.19) and (4.20), we yield

$$\mathbb{E}_\pi[\mathcal{I}_X] \leq 3n(8d+6). \tag{4.21}$$

We can then combine this with Theorem 7 to get

$$\begin{aligned} I(\pi; P_\theta) &\leq \phi\left(0, \frac{\operatorname{Tr}(\Sigma_\pi)\mathbb{E}_\pi[\mathcal{I}_X]}{d^2}\right) \\ &\leq \phi\left(0, \frac{3n(8d+6)}{d(d+2)}\right) \\ &\leq \phi\left(0, \frac{24n}{d}\right), \end{aligned}$$

where the last inequality holds for all $n, d \geq 1$.

Additional numerical simulations

Consider a k-uniformly logconcave prior π and observations $X \sim P_\theta$. The following lower bound on Bayes ℓ_2 risk is obtained by an application of the Shannon lower bound (Theorem 13) associated with ℓ_2 loss in conjunction with Theorem 7,

$$\mathbb{E}|\hat{\theta} - \theta|^2 \geq \frac{d}{2\pi e} e^{2h(\theta)} e^{-2(1-kP) - \frac{2}{d}\phi(kP, JP)}, \tag{4.22}$$

with $P := \frac{1}{d}\operatorname{Tr}(\Sigma_\pi)$, $J := \frac{1}{d}\mathbb{E}[\mathcal{I}_X]$, and $\phi(\cdot,\cdot)$ defined as in Theorem 7.

It is of practical interest to understand whether the lower bound of (4.22) is within a constant multiple of certain estimators. Towards this end, we investigate the average error of the maximum a-posteriori (MAP) estimator under several toy examples. We remark that the inequality (4.22) holds for *any* estimator $\hat{\theta} \in \mathbb{R}^d$, so if the MAP estimator has expected error within a constant multiple of the lower bound, so does the "best" estimator.

All simulations are implemented with PyMC3 [Salvatier et al., 2016].

Gaussian observation model

In our first numerical example, we compare the error of the MAP estimator and the lower bound (4.22) under a Gaussian observation model.

Suppose $X = (X_1, \ldots, X_n)$ are sampled i.i.d. with $X_i \sim \mathcal{N}(\theta, \sigma^2 \mathrm{I}_d)$, $i = 1, \ldots, n$. We consider four priors with independent components following Uniform$(0, 2\sqrt{3})$, Exponential(1), Gaussian(1) and a truncated normal that is Gaussian(1) truncated to $[-1, 1]$, respectively.

In Figure 4.1, we plot the ratio between errors of 100 random samples under $n = d = 1$ with varying values of σ^2 and our lower bound given in (4.22). Across the board, we see that the ratio is within a constant multiple.

In Figure 4.2, we fix $\sigma^2 = 1$ and $d = 10$ and vary the values of n. In Figure 4.3, we fix $\sigma^2 = 1$ and $n = 10$ and vary the values of d. In both figures, we plot the ratio between the error and the lower bound of (4.22) over 100 random samples, and observe that the errors are also within constant multiples as desired.

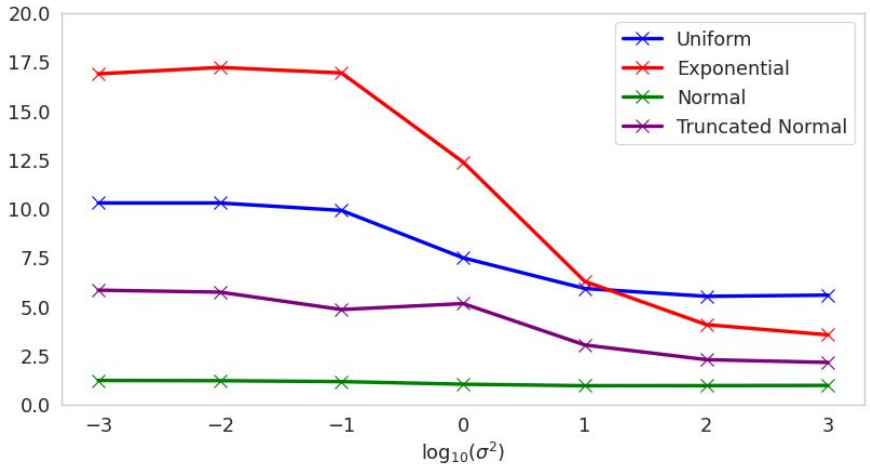

Figure 4.1: Ratio between average error of MAP estimator and lower bound of (4.22), plotted with varied σ^2 under $n = d = 1$. $X = (X_1, \ldots, X_n)$ is sampled i.i.d. with $X_i \sim \mathcal{N}(\theta, \sigma^2 \mathrm{I}_d)$, $i = 1, \ldots, n$ and different π.

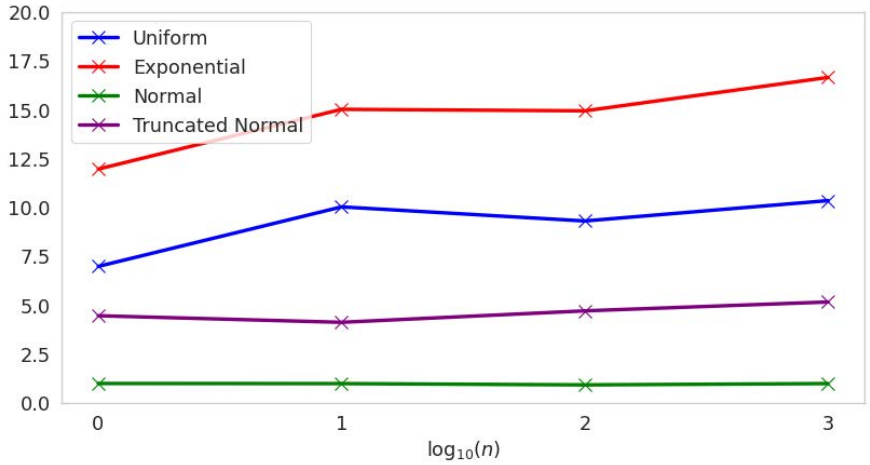

Figure 4.2: Ratio between average error of MAP estimator and lower bound of (4.22), plotted with varied n under $d = 10$ and $\sigma^2 = 1$. $X = (X_1, \ldots, X_n)$ is sampled i.i.d. with $X_i \sim \mathcal{N}(\theta, \sigma^2 \mathrm{I}_d)$, $i = 1, \ldots, n$ and different π.

Bernoulli observation model

A similar theme can be found in a Bernoulli observation model, where $X = (X_1, \ldots, X_n)$ are sampled with i.i.d. components from $\mathrm{Ber}\left(e^\theta/(1+e^\theta)\right)$. We consider three priors with independent components following $\mathrm{Uniform}(-\sqrt{3}, \sqrt{3})$, Gaussian(1) and a truncated normal that is Gaussian(1) truncated to $[-1, 1]$, respectively. The average error over 100 random

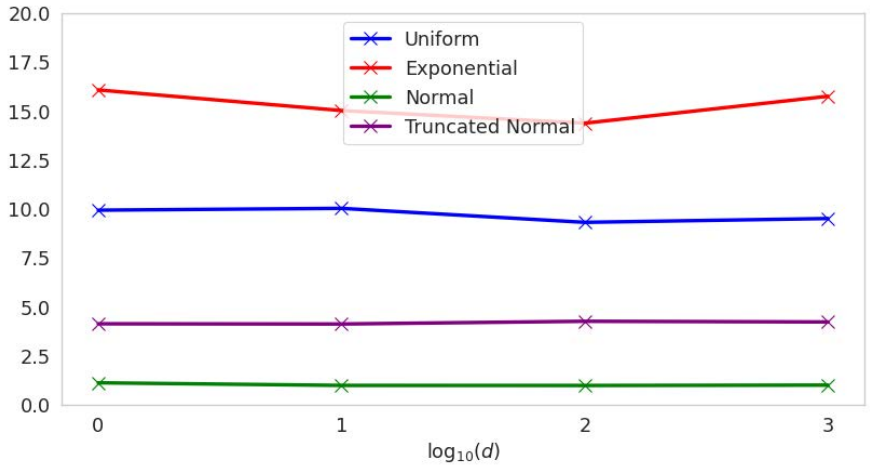

Figure 4.3: Ratio between average error of MAP estimator and lower bound of (4.22), plotted with varied d under $n = 10$ and $\sigma^2 = 1$. $X = (X_1, \ldots, X_n)$ is sampled i.i.d. with $X_i \sim \mathcal{N}(\theta, \sigma^2 \mathrm{I}_d)$, $i = 1, \ldots, n$ and different π.

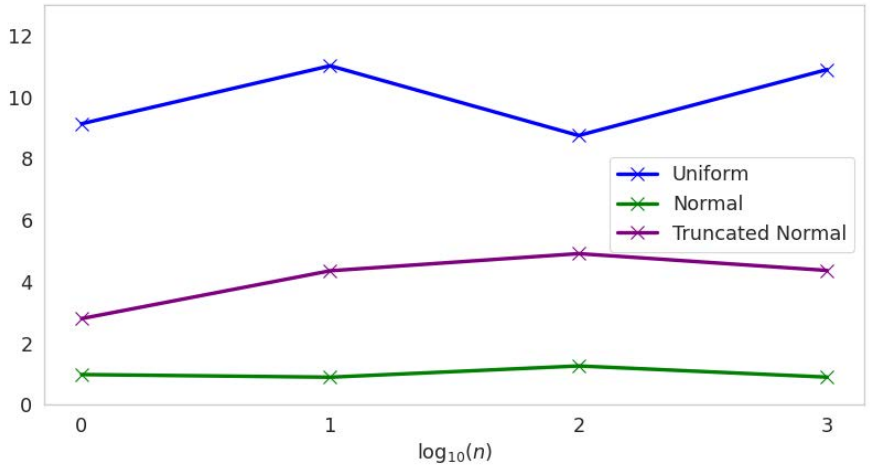

Figure 4.4: Ratio between average error of MAP and the lower bound of (4.22) under varying n. $X = (X_1, \ldots, X_n)$ are generated i.i.d. according to Bernoulli$(e^\theta/(1+e^\theta))$ and different π.

experiments with varying n and three priors (uniform, truncated normal and normal) are presented in Fig. 4.4. We find that the average error of the MAP estimator are within constant multiples of the lower bound given by Theorem 7, consistent with our statements.

Chapter 5

Lower bounds on risk under the Generalized Linear Model

In the previous chapter, we presented procedures to acquire lower bounds on risk using our new upper bounds on mutual information, Theorems 4 and 7. In this chapter, we discuss applications to the Generalized Linear Model.

5.1 Introduction

Generalized Linear Models (GLMs) are a flexible class of parametric statistical models that extend the class of linear models relating a random observation $X \in \mathbb{R}^n$ to a parameter $\theta \in \mathbb{R}^d$ via the linear relation

$$X = M\theta + Z, \tag{5.1}$$

where $M \in \mathbb{R}^{n \times d}$ is a known fixed **design matrix**, also called the **measurement matrix**, and $Z \in \mathbb{R}^n$ is a random noise vector. Given a univariate GLM in canonical form with natural parameter $\eta \in \mathbb{R}$, the density of observation $X \in \mathbb{R}$ given η is expressed as the exponential family

$$f(x;\eta) = h(x) \exp\left(\frac{\eta x - \Phi(\eta)}{s(\sigma)}\right) \quad x \in \mathcal{X}, \tag{5.2}$$

for known functions $h : \mathcal{X} \subseteq \mathbb{R} \to [0, \infty)$ (the *base measure*), $\Phi : \mathbb{R} \to \mathbb{R}$ (the *cumulant function*) and a scale parameter $s(\sigma) > 0$. This class of models captures a wide variety of parametric models such as binomial, Gaussian, Poisson, etc. As a specific example, we can take $\mathcal{X} = \{0, 1, 2, \ldots\}$ equipped with the counting measure λ. Under $h(x) = 1/x!$, $\Phi(t) = e^t$ and $s(\sigma) = 1$, the density $f(\cdot;\eta)$ is Poisson(e^η).

In this chapter, we restrict our attention to multivariate GLMs of the form

$$f(x;\theta) = \prod_{i=1}^{n} \left\{ h(x_i) \exp\left(\frac{x_i \langle m_i, \theta\rangle - \Phi(\langle m_i, \theta\rangle)}{s(\sigma)}\right)\right\} \quad x_i \in \mathcal{X} \subseteq \mathbb{R}^d, \tag{5.3}$$

for a real parameter $\theta \in \mathbb{R}^d$ and a fixed design matrix $M \in \mathbb{R}^{n\times d}$, with rows given by the vectors $\{m_i\}_{i=1}^n \subset \mathbb{R}^d$. In words, the above model assumes each X_i is drawn from the same exponential family, with respective natural parameters $\langle m_i, \theta\rangle$, $i = 1, \ldots, n$. This captures the linear model (5.1) in the usual case where the noise vector Z is assumed to be standard normal on $\mathbb{R}^n$, but is also flexible enough to capture many other models of interest. We refer the interested reader to [McCullagh, 2019, Dobson and Barnett, 2018, Nelder and Wedderburn, 1972] for more context on the history and theory of the generalized linear model.

Before we state our main results, we make the following assumption throughout:

Assumption 2. *We assume the cumulant function $\Phi : \mathbb{R} \to \mathbb{R}$ in (5.3) is twice-differentiable, with second derivative uniformly bounded as $\Phi'' \leq L$, for some $L > 0$.*

Remark 12. *This assumption is standard in the literature on minimax estimation for GLMs, and is equivalent to the map $\theta \longmapsto \mathbb{E}_{X\sim f(\cdot;\theta)}[X]$ being L-Lipschitz. See, for example, [Abramovich and Grinshtein, 2016, Loh and Wainwright, 2015, Negahban et al., 2012, Müller and Stadtmüller, 2005].*

5.2 Bayes risk

Let us start with the Bayes risk of the GLM under two different priors: a Gaussian prior and a uniform prior.

Bayes risk under a Gaussian prior

Suppose that observations $X = (X_1, \ldots, X_n)$ are generated i.i.d. according to (5.3). When the prior π is $\mathcal{N}(0, \tau^2 \mathrm{I}_d)$, the following lower bound on Bayes risk is given by Chen et al. [2016],

$$\mathbb{E}|\hat{\theta} - \theta|^r \geq C_1(r) \left[d \min\left(\frac{s(\sigma)}{L\lambda_{\max}(M^T M)}, \tau^2 \right) \right]^{\frac{r}{2}}, \tag{5.4}$$

which holds for any estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^d$ with a universal constant $C_1(r)$ only depending on r.

A better dependence on the eigenvalues of $M^T M$ can in fact be achieved, as shown in the following theorem.

Theorem 18 (Lower bound on Bayes risk of the GLM under a Gaussian prior)**.** *Suppose observations $X = (X_1, \ldots, X_n)$ are generated i.i.d. from a GLM defined in (5.3), and suppose Assumption 2 holds. Under the prior $\pi = \mathcal{N}(0, \tau^2 I_d)$, the following holds for any estimator*

$\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^d$,

$$\mathbb{E}|\hat{\theta} - \theta|^r \geq C_2(r) \left(d \min \left(\frac{ds(\sigma)}{L \sum_{i=1}^d \lambda_i(M^T M)}, \tau^2 \right) \right)^{\frac{r}{2}}. \tag{5.5}$$

Therefore, we improve the dependence on $M^T M$ from the maximum eigenvalue to an average of eigenvalues, which can be significant when extreme eigenvalues exist. This is illustrated by the relation

$$\frac{d}{\sum_{i=1}^d \lambda_i(M^T M)} \geq \frac{1}{\lambda_{\max}(M^T M)},$$

from which it is clear that (5.5) has a sharper topological dependence on M than (5.4).

Before we move on to the proof of Theorem 18, we require the following lemma, which relates diagonal terms of the Fisher information matrix $\bar{\mathcal{I}}_X$ with columns in the measurement matrix M.

Lemma 19. *Suppose observations $X = (X_1, \ldots, X_n)$ are generated i.i.d. from a GLM defined in (5.3), and suppose Assumption 2 holds. Then,*

$$\mathbb{E}_\pi [\bar{\mathcal{I}}_X]_{ii} \leq \frac{L}{s(\sigma)} \sum_{j=1}^n M_{ji}^2$$

holds for all $i = 1, \ldots, d$.

Proof of Lemma 19. Our proof relies on the following well-known identities associated with exponential families of the form we consider.

Lemma 20 ([McCullagh, 2019, Page 29])**.** *Fix m and θ, and consider a density $f(x;\theta) = h(x) \exp\left(\frac{x\langle m,\theta\rangle - \Phi(\langle m,\theta\rangle)}{s(\sigma)} \right)$ with respect to λ. A random observation $X \sim f(\cdot;\theta)$ has mean $\Phi'(\langle m,\theta\rangle)$ and variance $s(\sigma) \cdot \Phi''(\langle m,\theta\rangle)$.*

It follows that, with our assumption $\Phi'' \leq L$, we have for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
[\bar{\mathcal{I}}_X(\theta)]_{ii} &= \mathbb{E}_{X \sim f(\cdot;\theta)} \left(\frac{\partial}{\partial \theta_i} \log f(X;\theta) \right)^2 \\
&= \frac{1}{s^2(\sigma)} \mathbb{E}_{X \sim f(\cdot;\theta)} \left(\sum_{j=1}^n M_{ji} \left(X_j - \Phi'(\langle m_j, \theta \rangle) \right) \right)^2 \\
&= \frac{1}{s^2(\sigma)} \sum_{j=1}^n \left(M_{ji}^2 \operatorname{Var}(X_j) \right) \\
&\leq \frac{1}{s(\sigma)} \sum_{j=1}^n \left(M_{ji}^2 L \right) \\
&= \frac{L}{s(\sigma)} [M^T M]_{ii}.
\end{aligned} \tag{5.6}$$

Taking expectation on both sides completes the proof. □

Now we are ready to give the proof of Theorem 18.

Proof of Theorem 18. The proof begins by noting that Lemma 19 immediately gives

$$\begin{aligned}\mathbb{E}_\pi[\mathcal{I}_X] &= \mathrm{Tr}\left(\mathbb{E}_\pi[\bar{\mathcal{I}}_X]\right) \\ &= \sum_{i=1}^{d} \mathbb{E}_\pi[\bar{\mathcal{I}}_X]_{ii} \\ &\leq \frac{L}{s(\sigma)} \sum_{i=1}^{d} \lambda_i(M^T M). \end{aligned} \tag{5.7}$$

Under the Gaussian prior $\pi = \mathcal{N}(0, \tau^2 \mathrm{I}_d)$, we may invoke Theorem 4 to get

$$\begin{aligned} I(\pi; P_\theta) &\leq \frac{d}{2} \log\left(1 + \frac{\mathrm{Tr}(\Sigma_\pi)\mathbb{E}_\pi[\mathcal{I}_X]}{d^2}\right) \\ &\leq \frac{d}{2} \log\left(1 + \frac{\tau^2 L}{s(\sigma) d} \sum_{i=1}^{n} \lambda_i(M^T M)\right), \end{aligned}$$

which combined with (4.10), and recalling the entropy of a Gaussian prior in Example 6, yields the following lower bound for any estimator $\hat{\theta}$,

$$\begin{aligned} \mathbb{E}|\hat{\theta} - \theta|^r &\geq \frac{d}{re} \pi^{-\frac{r}{2}} \left(\frac{\Gamma(d/r + 1)}{\Gamma(d/2 + 1)}\right)^{-\frac{r}{d}} \exp\left(\frac{r}{d}\left(h(\theta) - I(\pi; P_\theta)\right)\right) \\ &\geq c(r) \tau^r d^{\frac{r}{2}} \left(1 + \frac{L\tau^2 \sum_{i=1}^{d} \lambda_i(M^T M)}{d s(\sigma)}\right)^{-\frac{r}{2}} \\ &\geq C_2(r) \left(d \min\left(\frac{d s(\sigma)}{L \sum_{i=1}^{d} \lambda_i(M^T M)}, \tau^2\right)\right)^{\frac{r}{2}}. \end{aligned}$$

Here, $c(r)$ and $C_2(r)$ are constants that only depend on r, and the second equation makes use of $\Gamma\left(d/2 + 1\right)^{-1/d} \asymp d^{-1/2}$, a consequence of Stirling's approximation originally due to de Moivre; see, e.g., [Dutka, 1991], for a history of Stirling's approximation.

□

Remark 13. *A tighter dependence on eigenvalues can be achieved by a component-wise treatment we will introduce for establishing minimax risk of the GLM later in the chapter (see Section 5.3). Here, our technique yields a more straightforward improvement over* (5.4)*, which is the main point.*

In the case of a Gaussian prior, one can alternatively use Efroimovich's inequality (see, e.g., (2.7)) to yield the same bound. This is not the case, however, in our next example where π is taken to be a uniform prior over the unit ℓ_2 ball.

Bayes risk under a uniform prior

Suppose that observations $X = (X_1, \ldots, X_n)$ are generated i.i.d. according to (5.3). When the prior π is a uniform distribution on the ℓ_2 ball, we obtain the following lower bound on Bayes risk.

Proposition 9 (Lower bound on Bayes risk of the GLM under a uniform prior)**.** *Suppose observations $X = (X_1, \ldots, X_n)$ are generated i.i.d. from a GLM defined in (5.3), and suppose Assumption 2 holds. When the prior π is the uniform measure over a unit ℓ_2 ball, the following holds for any estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^d$,*

$$\mathbb{E}|\hat{\theta} - \theta|^r \geq C_3(r) \left(d \min \left(\frac{ds(\sigma)}{L \sum_{i=1}^d \lambda_i(M^T M)}, 1 \right) \right)^{r/2}.$$

Therefore, we find that the lower bound exhibits a similar topological dependence on $M^T M$ as in the case where the prior π is Gaussian.

Proof of Proposition 9. Under the uniform prior π, we may invoke Theorem 7 and (5.7) to get

$$\begin{aligned} I(\pi; P_\theta) &\leq d\phi \left(0, \frac{\mathrm{Tr}(\Sigma_\pi)\mathbb{E}_\pi[\mathcal{I}_X]}{d^2} \right) \\ &\leq d\phi \left(0, \frac{\mathbb{E}_\pi[\mathcal{I}_X]}{d(d+2)} \right) \\ &\leq \frac{d}{2}\phi \left(0, \frac{L}{s(\sigma)d} \sum_{i=1}^n \lambda_i(M^T M) \right), \end{aligned} \tag{5.8}$$

where we remind ourselves that the covariance matrix of the uniform distribution over the unit ℓ_2 ball is

$$\Sigma_\pi = \frac{1}{d+2} \mathrm{I}_d$$

as given in (4.13).

Moving onwards, we can then combine (5.8) with (4.10) to yield the following lower bound for any estimator $\hat{\theta}$,

$$\begin{aligned} \mathbb{E}|\hat{\theta} - \theta|^r &\geq \frac{d}{re} \pi^{-\frac{r}{2}} \left(\frac{\Gamma(d/r + 1)}{\Gamma(d/2 + 1)} \right)^{-\frac{r}{d}} \exp\left(\frac{r}{d} \left(h(\theta) - I(\pi; P_\theta) \right) \right) \\ &\geq C(r) d^{\frac{r}{2}} \exp\left(-\frac{r}{2d} \phi \left(0, \frac{L}{s(\sigma)d} \sum_{i=1}^n \lambda_i(M^T M) \right) \right) \\ &\geq C_3(r) \left(d \min \left(\frac{ds(\sigma)}{L \sum_{i=1}^d \lambda_i(M^T M)}, 1 \right) \right)^{r/2}. \end{aligned}$$

Here, $C(r), C_3(r)$ are constants that only depend on r.

□

Finally, we remark that the procedure we used to obtain lower bounds on Bayes risk under Gaussian and uniform priors can be easily modified to suit other log-concave priors by an application of Theorem 4 or Theorem 7.

5.3 Minimax estimation risk

In terms of parameter estimation, a common figure of merit is the constrained ℓ_2 minimax risk, which corresponds to the worst-case ℓ_2 estimation error, with θ allowed to range over a constrained set Θ. For our purposes, let us consider in this section Θ equal to the Euclidean ball in $\mathbb{R}^d$, denoted $\mathbb{B}_2^d(1) := \{v : v \in \mathbb{R}^d, |v|^2 \leq 1\}$, which is a common choice in applications. A similar result for Θ taken as a Euclidean ball with radius R can be naturally extended from our results under $R = 1$. More precisely, we make the following definition for the constrained minimax risk.

Definition 1. *For the GLM (5.3), we define the associated minimax risk with $\Theta = \mathbb{B}_2^d(1)$ via*

$$R^*(h, \Phi, M, s(\sigma)) := \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}|\hat{\theta} - \theta|^2,$$

where the infimum is over all $\mathbb{R}^d$-valued estimators $\hat{\theta}$ (i.e., measurable functions of the observation $X \sim P_\theta$).

We establish the following general lower bound on the minimax risk for the class of GLMs introduced above.

Theorem 21. *The ℓ_2 minimax risk for the class of models (5.3) is lower bounded according to*

$$R^*(h, \Phi, M, s(\sigma)) \gtrsim \min\left(\frac{s(\sigma)}{L} \operatorname{Tr}\left((M^T M)^{-1}\right), 1\right), \tag{5.9}$$

where $\gtrsim$ denotes inequality, up to a universal constant.

In the case where $M^T M$ is not invertible, we adopt the convention that $\operatorname{Tr}\left((M^T M)^{-1}\right) = +\infty$. This situation occurs when M is not full rank, in which case θ is not identifiable in the null space of M and constant error is unavoidable.

Remark 14. *In fact, with minor modification, Theorem 21 holds for the more general class of GLMs with observations drawn from densities of the form*

$$f(x; \theta) = \prod_{i=1}^n \left\{ h_i(x_i) \exp\left(\frac{x_i \langle m_i, \theta\rangle - \Phi_i(\langle m_i, \theta\rangle)}{s_i(\sigma)}\right)\right\}.$$

Remark 15. *Since minimax risk is generally characterized modulo universal constants, the statement* (5.9) *in terms of* $\gtrsim$ *is sufficient for our purposes. However, a careful analysis of our arguments reveals that* $\gtrsim$ *can be replaced with* $\geq$ *at the expense of including a modest constant in the RHS of* (5.9) *(e.g.,* $1/(\pi e^3)$*).*

Most interestingly, the minimax bound (5.9) holds uniformly over the class of GLMs given by (5.3), and is of the correct order for the canonical linear model (5.1). Indeed, under the linear model $X = LM\theta + Z$, where Z is standard Gaussian with covariance $\sigma^2 L \cdot I$, the design matrix M is full rank and $L > 0$, it is well known that the maximum likelihood estimator (MLE) minimizes the ℓ_2 estimation error. The MLE estimator $\hat{\theta}_{\text{MLE}}$ is given by

$$\hat{\theta}_{\text{MLE}} = L^{-1}(M^T M)^{-1} M^T X,$$

and one can explicitly calculate the ℓ_2 error as

$$\begin{aligned} \mathbb{E}|\hat{\theta}_{\text{MLE}} - \theta|^2 &= \mathbb{E}|\theta - L^{-1}(M^T M)^{-1} M^T X|^2 \\ &= \frac{1}{L^2}\mathbb{E}|(M^T M)^{-1} M^T Z|^2 \\ &= \frac{\sigma^2}{L}\operatorname{Tr}((M^T M)^{-1}). \end{aligned} \tag{5.10}$$

The linear model in this case corresponds to $h(x) = e^{-x^2/(2L\sigma^2)}$, $s(\sigma) = \sigma^2$, and $\Phi(t) = Lt^2/2$ in (5.3).

Comparing (5.10) to Theorem 21, we find that our minimax lower bound is achieved (up to a universal constant) for linear models of the above form. To summarize, we have the following:

Corollary 22. *Fix a design matrix* M*, scale parameter* $s(\sigma)$ *and* $L > 0$*. Among those generalized linear models in* (5.3) *with* $\Phi'' \leq L$*, linear models are most favorable in terms of minimax risk. More precisely, among this class of models,*

$$R^*(h, \Phi, M, s(\sigma)) \gtrsim R^*(e^{-(\cdot)^2/(2Ls(\sigma))}, (\cdot)^2 L/2, M, s(\sigma)).$$

Roughly speaking, the above shows that linear models are most favorable among a broad class of GLMs.

Remarks

A few remarks are in order. First, we note that the argument yields the stronger entropic inequality,

$$\inf_{\hat{\theta}} \sup_{\theta \sim \pi} \sum_{i=1}^{d} e^{2h(\theta_i|\hat{\theta}_i)} \gtrsim \min\left(\frac{s(\sigma)}{L}\operatorname{Tr}\left((M^T M)^{-1}\right), 1\right)$$

which improves Theorem 21 (seen by the max-entropy property of gaussians). Here, the supremum is taken over all distributions π supported on the ℓ_2 ball $\mathbb{B}_2^d(1)$. The expression

on the left hand side can be thought of as a Bayes minimax problem on an entropic loss, which is connected to logarithmic loss in the statistical learning and information literature, see, e.g., Jiao et al. [2015], Courtade and Weissman [2013], Courtade and Wesel [2011], Cesa-Bianchi and Lugosi [2006].

Second, we remark that our analysis is flexible enough for generalizations to other forms of the GLM. For example, consider observation X drawn from the density

$$f(x;\theta) = \prod_{i=1}^{n} \left\{ h_i(x_i) \exp\left(\frac{x_i \langle m_i, \theta \rangle - \Phi_i(\langle m_i, \theta \rangle)}{s_i(\sigma)} \right) \right\}.$$

Suppose Assumption 2 holds for each cumulant function Φ_i (i.e., $\Phi_i'' \leq L$ for each $i = 1, \ldots, n$). Then, a slight modification in (5.6) yields

$$[\bar{\mathcal{I}}_X(\theta)]_{ii} \leq \frac{L}{s^*(\sigma)} [M^T M]_{ii}$$

where $s^*(\sigma) = \min_{i=1,\ldots,n} s_i(\sigma)$. Following (5.22) and the same choice of $(\epsilon_i)_{i=1,\ldots,d}$ in Section 5.6 with the argument $s(\sigma)$ replaced by $s^*(\sigma)$, we obtain minimax lower bound

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}|\hat{\theta} - \theta|^2 \gtrsim \min\left(\frac{s^*(\sigma)}{L} \operatorname{Tr}\left((M^T M)^{-1} \right), 1 \right).$$

5.4 Minimax prediction risk

In this section, we investigate another figure of merit, the minimax prediction risk, defined as

$$R_p^*(h, \Phi, M, s(\sigma)) := \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \frac{1}{n} \mathbb{E}|M\hat{\theta} - M\theta|^2. \tag{5.11}$$

We establish the following lower bound on the minimax prediction risk.

Theorem 23. *Suppose observations $X \in \mathbb{R}^n$ are generated via the GLM (5.3) satisfying $\Phi'' \leq L$ for some constant $L > 0$. With a fixed design matrix $M \in \mathbb{R}^{n \times d}$, the following lower bound on minimax prediction risk holds.*

$$R_p^*(h, \Phi, M, s(\sigma)) \geq \frac{s(\sigma)}{Ln} \operatorname{rank}(M). \tag{5.12}$$

We remark that (5.12) holds uniformly across all generalized linear models, provided they satisfy $\Phi''(\cdot) \leq L$ for some $L > 0$. The bound is tight for the Gaussian linear model (see, e.g., [Wainwright, 2019]). Indeed, under the linear model $X = LM\theta + Z$, where Z is Gaussian with covariance $\sigma^2 L \mathrm{I}_n$, the maximum likelihood estimator (MLE) $\hat{\theta}_{\mathrm{MLE}}$ is given by

$$\hat{\theta}_{\mathrm{MLE}} = L^{-1}(M^T M)^{\dagger} M^T X.$$

One can explicitly calculate the ℓ_2 prediction error of the MLE as

$$\begin{aligned}\frac{1}{n}\mathbb{E}|M\hat{\theta}_{\text{MLE}} - M\theta|^2 &= \frac{1}{n}\mathbb{E}|ML^{-1}(M^TM)^\dagger M^TX - M\theta|^2 \\ &= \frac{1}{L^2n}\mathbb{E}|M(M^TM)^\dagger M^TZ|^2 \\ &= \frac{\sigma^2}{Ln}\text{rank}(M). \end{aligned} \tag{5.13}$$

The linear model in this case corresponds to $h(x) = e^{-x^2/(2L\sigma^2)}$, $s(\sigma) = \sigma^2$, and $\Phi(t) = Lt^2/2$ in (5.3). Comparing (5.13) to (5.12), we find that our minimax lower bound is achieved with an exact constant for linear models of the above form.

5.5 Bibilographical remarks

A closely related work is that of Abramovich and Grinshtein [2016], who consider the generalized linear model albeit with a slightly different setup. Their paper provides minimax lower bounds for the Kullback-Leibler divergence between the vector $M\theta$ and any estimator $\widehat{M\theta}$ under a k-sparse setting $\|\theta\|_0 \le k$, with the parameter θ constrained to have at most k non-zero entries. When the cumulant function Φ is strongly convex with $0 < R \le \Phi'' \le L$ for some fixed constants R, L, the arguments of Abramovich and Grinshtein [2016] can be adapted to obtain the following minimax lower bound

$$\inf_{\widehat{M\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}|\widehat{M\theta} - M\theta|^2 \gtrsim \frac{ds(\sigma)R}{L^2} \cdot \frac{\lambda_{\min}(M^TM)}{\lambda_{\max}(M^TM)}, \tag{5.14}$$

where M is assumed to be full rank and $\lambda_{\min}$ and $\lambda_{\max}$ denote smallest and largest eigenvalues, respectively. The bound is weaker than our bound given in (5.12) in both topological dependence on M and assumptions (e.g., we do not require a lower bound $\Phi'' \ge R$).

Although (5.14) is on minimax prediction error, it can be translated into a bound on minimax estimation error using the operator norm inequality $|M(\theta-\hat{\theta})|^2 \le \lambda_{\max}(M^TM)|\theta-\hat{\theta}|^2$, giving

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}|\hat{\theta} - \theta|^2 \gtrsim \frac{ds(\sigma)R}{L^2} \cdot \frac{\lambda_{\min}(M^TM)}{\lambda^2_{\max}(M^TM)}.$$

A direct computation shows that (5.9) is sharper than the ℓ_2 minimax lower bound given above, as evidenced by the relation

$$\frac{d\,\lambda_{\min}(M^TM)}{\lambda^2_{\max}(M^TM)} \le \frac{d}{\lambda_{\max}(M^TM)} \le \text{Tr}\left(\left(M^TM\right)^{-1}\right).$$

As for a general theory, apart from the gaussian linear model, the minimax estimator for the GLM typically does not have a closed form, but the Maximum Likelihood Estimator

(MLE) can be approximated by iterative weighted linear regression [Nelder and Wedderburn, 1972]. A variety of estimators such as aggregate estimators [Rigollet, 2012], robust estimators [Cantoni and Ronchetti, 2001] and GLM with Lasso [Van de Geer, 2008] have been proposed to solve different settings of the GLM. We refer interested readers to [McCullagh, 2019] for the theory of GLMs.

Another line of related work explores models with stochastic design matrix M. Duchi et al. [2018] consider inference of a parameter θ under privacy constraints. Negahban et al. [2012] and Loh and Wainwright [2015] provide consistency and convergence rates for M-estimators in GLMs with low-dimensional structure under high-dimensional scaling.

Separate from the minimax problems considered here, model selection is another line of popular work. Model selection in linear regression dates back to the seventies and has regained popularity over the past decade, due to the increase in need of data exploration for high dimensional data; see [Verzelen, 2012, Birgé and Massart, 2007, Akaike, 1998] and many other works for the history. More recently, tools in model selection for linear regression have been adapted for the GLM; see [Abramovich and Grinshtein, 2016] for a brief discussion.

There is a large body of work that establish minimax lower bounds on prediction error for specific models of the generalized linear model. Typically, these analyses depend on methods involving metric entropy (see, for example, [Wainwright, 2019, Abramovich and Grinshtein, 2016, Cai et al., 2016, Raskutti et al., 2011, Candes and Plan, 2011, Cai et al., 2010]). A popular minimax result is due to Raskutti et al. [2011], who consider the sparse Gaussian linear model, where for a fixed design matrix M with an additional sparsity constraint $\|\theta\|_0 \le k$,

$$\sigma^2 \frac{\Phi_{2k,-}(M)}{\Phi_{2k,+}(M)} \frac{k}{n} \log\left(\frac{ed}{k}\right) \lesssim \inf_{\hat{\theta}} \sup_{\|\theta\|_0 \le k} \frac{1}{n} \mathbb{E}|M\hat{\theta} - M\theta|^2 \lesssim \sigma^2 \min\left(\frac{k}{n} \log\left(\frac{ed}{k}\right), 1\right). \quad (5.15)$$

Here the terms $\Phi_{r,-}(M)$ and $\Phi_{r,+}(M)$ correspond to the *constrained eigenvalues*,

$$\Phi_{r,-}(M) := \inf_{0 \ne \|\theta\|_0 \le r} \frac{|M\theta|^2}{|\theta|^2}, \qquad \Phi_{r,+}(M) := \sup_{0 \ne \|\theta\|_0 \le r} \frac{|M\theta|^2}{|\theta|^2}. \quad (5.16)$$

The upper bound of (5.15) is achieved by classical methods such as aggregation [Verzelen, 2012, Bunea et al., 2007, Birgé and Massart, 2007, 2001].

There are also lines of work on specific settings of the generalized linear model. For example, Candes and Plan [2011] discusses low-rank matrix recovery, and Cai et al. [2016] considers phase retrieval. In a sparse setting, Abramovich and Grinshtein [2016] also provides similar bounds for sparse estimation under the generalized linear model that depend on the ratio between restricted eigenvalues. We remark that our minimax lower bound of (5.12) can naturally be extended to the sparse setting of generalized linear models, yielding

$$\frac{1}{n} \inf_{\hat{\theta}} \sup_{\|\theta\|_0 \le k} \mathbb{E}|M\hat{\theta} - M\theta|^2 \ge \frac{s(\sigma)}{L} \min(k, \operatorname{rank}(M)).$$

Therefore, in cases where the ratio of restricted eigenvalues are large, our bound is tighter despite lacking a logarithmic factor.

5.6 Additional proofs

Proof of Theorem 21

Recall that the design matrix M has as its rows $\{m_i\}_{i=1}^n \subset \mathbb{R}^d$. Writing the matrix M in terms of its SVD $M = U\Sigma V^T$ and defining u_i as the i-th column of the matrix U^T, we have

$$\langle m_i, \theta \rangle = \langle \underbrace{\Sigma u_i}_{\bar{m}_i}, \underbrace{V^T\theta}_{\bar{\theta}} \rangle = \langle \bar{m}_i, \bar{\theta} \rangle, \tag{5.17}$$

where we defined the variables $\bar{m}_i := \Sigma u_i$ and $\bar{\theta} := V^T\theta$. Since V is an orthogonal matrix by definition, it follows by rotation invariance of the ℓ_2 ball $\mathbb{B}_2^d(1)$ that the estimation problem can be equivalently formulated under the reparametrization $(\theta, M) \longrightarrow (\bar{\theta}, \bar{M})$, where $\bar{M} := MV = U\Sigma$. More specifically, the minimax risk for θ over the set of estimators for estimating $\theta \in \mathbb{B}_2^d(1)$ is equal to the minimax risk for estimating $\bar{\theta} \in \mathbb{B}_2^d(1)$,

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{B}_2^d(1)} \mathbb{E}|\hat{\theta} - \theta|^2 = \inf_{\hat{\bar{\theta}}} \sup_{\bar{\theta} \in \mathbb{B}_2^d(1)} \mathbb{E}|\hat{\bar{\theta}} - \bar{\theta}|^2.$$

As a result, we may assume without loss of generality that M^TM is a diagonal matrix.

By Theorem 12, minimax risk is lower bounded by the Bayes risk when θ is assumed to be distributed according to a prior π defined on the ℓ_2 ball $\mathbb{B}_2^d(1)$. Hence, our task is to judiciously select a prior π that yields the desired lower bound. Toward this end, we will let π be the uniform measure on the rectangle $\prod_{i=1}^d [-\epsilon_i/2, \epsilon_i/2]$ for values $(\epsilon_i)_{i=1}^d$ to be determined below satisfying

$$\sum_{i=1}^d \epsilon_i^2 \leq 4. \tag{5.18}$$

In other words, our construction implies θ has independent components, with the i-th coordinate θ_i uniform on the interval $[-\epsilon_i/2, \epsilon_i/2]$. The interval lengths will, in general, be chosen to exploit the structure of the design matrix M.

We now describe our construction of the sequence $(\epsilon_i)_{i=1}^d$. We start with the simple case, in which the matrix M does not have full (column) rank. In this case, there exists an eigenvalue $\lambda_k(M^TM) = 0$. For this index k, we set $\epsilon_i = 2\delta_{ik}$, $i = 1, \ldots, d$, where δ_{ij} is the

Kronecker delta function. Now, we may bound

$$\begin{aligned}
\mathbb{E}|\hat{\theta} - \theta|^2 &\geq \mathrm{Var}(\hat{\theta}_k - \theta_k) \\
&\overset{(a)}{\geq} \frac{1}{2\pi e} e^{2h(\hat{\theta}_k - \theta_k)} \\
&\overset{(b)}{\geq} \frac{1}{2\pi e} e^{2h(\theta_k|\hat{\theta}_k)} \\
&= \frac{1}{2\pi e} e^{2h(\theta_k) - 2I(\theta_k;\hat{\theta}_k)} \\
&\overset{(c)}{\geq} \frac{1}{2\pi e} e^{2h(\theta_k) - 2I(\theta_k;X)} \\
&\overset{(d)}{=} \frac{2}{\pi e},
\end{aligned}$$

where (a) follows from the max-entropy property of gaussians; (b) follows since conditioning reduces entropy: $h(\hat{\theta}_k - \theta_k) \geq h(\hat{\theta}_k - \theta_k|\hat{\theta}_k) = h(\theta_k|\hat{\theta}_k)$; (c) follows from the data processing inequality since $\theta_k \to X \to \hat{\theta}_k$ forms a Markov chain; and (d) follows since $\theta_k \sim \mathrm{Unif}(-1, 1)$ and $I(\theta_k; X) = 0$, since π is supported in the kernel of M by construction.

Having shown the minimax risk is lower bounded by a constant when M does not have full (column) rank, we assume henceforth that M has full rank.

Note that under our assumptions, the pair (X, θ) has a joint distribution, and therefore so does the pair (X, θ_i). Consistent with the previously introduced notation, we write $\mathcal{I}_X(\theta_i)$ to denote the Fisher information of X drawn according to the conditional law of X given θ_i. With this notation in hand, the next lemma provides a comparison between the expected Fisher information conditioned on a single component θ_i of the parameter θ and the i-th diagonal entry of the expected Fisher information matrix conditioned for parameter θ.

Lemma 24. *When the components of parameter $\theta \sim \pi \in \mathcal{P}(\mathbb{R}^d)$ are independent and X is generated by the GLM (5.3), we have*

$$\mathbb{E}_\pi[\bar{\mathcal{I}}_X]_{ii} \geq \mathbb{E}_{\pi_i}[\mathcal{I}_X(\theta_i)] \quad i = 1, \ldots, d.$$

Here, π_i denotes the marginal law of θ_i.

Proof. Begin by noting that $x \longmapsto \mathbb{E}\left[\, f(X;\theta) \middle|\, \theta_i, X = x\right]$ is the density of X given θ_i with

respect to λ. Now, the proof is by straightforward computation:

$$\begin{aligned}
\mathbb{E}_\pi[\bar{\mathcal{I}}_X]_{ii} &= \mathbb{E}\left[\frac{\left(\frac{\partial}{\partial\theta_i}f(X;\theta)\right)^2}{f^2(X;\theta)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\left.\frac{\left(\frac{\partial}{\partial\theta_i}f(X;\theta)\right)^2}{f^2(X;\theta)}\right|\theta_i, X\right]\right] \\
&\overset{(a)}{\geq} \mathbb{E}\left[\frac{\left(\mathbb{E}\left[\left.\frac{\partial}{\partial\theta_i}f(X;\theta)\right|\theta_i, X\right]\right)^2}{\left(\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}\right] \\
&\overset{(b)}{=} \mathbb{E}\left[\frac{\left(\frac{\partial}{\partial\theta_i}\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}{\left(\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}\right] \\
&= \mathbb{E}_{\pi_i}[\mathcal{I}_X(\theta_i)].
\end{aligned}$$

In the above, (a) follows from Cauchy-Schwarz. Indeed, let π_i and $\pi^{(i)}$ denote the marginal laws of θ_i and $\theta^{(i)} := (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$ (i.e., the leave-one-out vector), respectively. Using independence of θ_i and $\theta^{(i)}$, note that

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(\frac{\partial}{\partial\theta_i}f(X;\theta)\right)^2}{f(X;\theta)^2}\right] &= \int_{\mathbb{R}}\int_{\mathcal{X}}\int_{\mathbb{R}^{d-1}}\frac{\left(\frac{\partial}{\partial\theta_i}f(x;\theta)\right)^2}{f(x;\theta)}d\pi^{(i)}(\theta^{(i)})d\lambda(x)d\pi_i(\theta_i) \\
&\geq \int_{\mathbb{R}}\int_{\mathcal{X}}\frac{\left(\int_{\mathbb{R}^{d-1}}\frac{\partial}{\partial\theta_i}f(x;\theta)d\pi^{(i)}(\theta^{(i)})\right)^2}{\int_{\mathbb{R}^{d-1}}f(x;\theta)d\pi^{(i)}(\theta^{(i)})}d\lambda(x)d\pi_i(\theta_i) \\
&= \mathbb{E}\left[\frac{\left(\mathbb{E}\left[\left.\frac{\partial}{\partial\theta_i}f(X;\theta)\right|\theta_i, X\right]\right)^2}{\left(\mathbb{E}\left[f(X;\theta)|\,\theta_i, X\right]\right)^2}\right],
\end{aligned}$$

where the last line follows since

$$x \longmapsto \mathbb{E}\left[f(X;\theta)|\,\theta_i, X = x\right] = \int_{\mathbb{R}^{d-1}} f(x;\theta)d\pi^{(i)}(\theta^{(i)})$$

is the density (w.r.t. λ) of X given θ_i.

Moving onwards, (b) follows from independence between θ_i and $\theta^{(i)}$ and the Leibniz integral rule. Application of the latter can be justified by the assumed regularity of Φ and compactness of $\mathbb{B}_2^d(1)$. □

Next, fix $\epsilon_i > 0$. Since $\theta_i \sim \text{Unif}(-\epsilon_i/2, \epsilon_i/2)$ has a log-concave distribution, we can apply Lemma 24 to conclude

$$\begin{aligned} e^{2h(\theta_i|\hat{\theta}_i)} &\geq e^{2h(\theta_i)-2I(\theta_i;X)} \\ &\geq e^{2h(\theta_i)-2\phi\left(\text{Var}(\theta_i)\mathbb{E}_{\pi_i}[\mathcal{I}_X]\right)} \\ &\geq \epsilon_i^2 e^{-2\phi\left(\frac{\epsilon_i^2}{12}\mathbb{E}_{\pi}[\mathcal{I}_X]_{ii}\right)} \qquad i = 1, \ldots, d. \end{aligned} \tag{5.19}$$

Here, we define the function $\phi(\cdot)$ as:

$$\phi(t) := \begin{cases} \sqrt{t} & \text{if } t < 1 \\ 1 + \frac{1}{2}\log(t) & \text{if } t \geq 1. \end{cases} \tag{5.20}$$

Remark 16. *This definition coincides with the single dimensional version of $\phi(0, t)$ defined in Theorem 7. We remark that the definition of $\phi(\cdot, \cdot)$ there depends on the dimension d, whereas in our application here we invoke the single dimensional Theorem 7 on the mutual information $I(\theta_i; X)$, and hence the notation $\phi(\cdot)$ is adopted to prevent confusion with dimensions.*

Note that the last inequality used the identities $\text{Var}(\theta_i) = \frac{\epsilon_i^2}{12}$ and $h(\theta_i) = \log(\epsilon_i)$, holding by construction.

Putting (5.19) and Lemma 19 together, we conclude for any choice of $\epsilon_i > 0$,

$$e^{2h(\theta_i|\hat{\theta}_i)} \geq \epsilon_i^2 \exp\left[-2\phi\left(\frac{\epsilon_i^2}{12}\frac{L}{s(\sigma)}[M^T M]_{ii}\right)\right]. \tag{5.21}$$

In case $\epsilon_i = 0$, we have the trivial equality $e^{2h(\theta_i|\hat{\theta}_i)} = 0$, which is consistent with the RHS of (5.21) evaluated at $\epsilon_i = 0$. Hence, the estimate (5.21) holds for all $\epsilon_i \geq 0$.

Summing (5.21) from $i = 1, \ldots, d$, for parameter $\theta \sim \pi = \prod_{i=1}^d \text{Unif}(-\epsilon_i/2, \epsilon_i/2)$, we have the following lower bound on the Bayes risk,

$$\begin{aligned} \mathbb{E}|\hat{\theta} - \theta|^2 &\geq \sum_{i=1}^d \text{Var}(\hat{\theta}_i - \theta_i) \\ &\geq \frac{1}{2\pi e}\sum_{i=1}^d e^{2h(\theta_i|\hat{\theta}_i)} \\ &\geq \frac{1}{2\pi e}\sum_{i=1}^d \epsilon_i^2 \exp\left[-2\phi\left(\frac{\epsilon_i^2}{12}\frac{L}{s(\sigma)}[M^T M]_{ii}\right)\right]. \end{aligned} \tag{5.22}$$

It remains to choose an appropriate sequence $(\epsilon_i)_{i=1,\ldots,d}$ to obtain the desired lower bound. Toward this end, we consider two cases:

Case 1: $\mathrm{Tr}((M^TM)^{-1}) \leq \frac{1}{3}\frac{L}{s(\sigma)}$

In this case, we choose $\epsilon_i^2 = 12\frac{s(\sigma)}{L}\left([M^TM]_{ii}\right)^{-1}$ for $i = 1, \ldots, d$. Note that by our assumption that M^TM is diagonal,

$$\sum_{i=1}^{d} \epsilon_i^2 = 12\frac{s(\sigma)}{L}\,\mathrm{Tr}((M^TM)^{-1}) \leq 4,$$

so that (5.18) is satisfied. By an application of (5.22), we have

$$\begin{aligned}
\mathbb{E}|\hat{\theta} - \theta|^2 &\gtrsim \sum_{i=1}^{d} \epsilon_i^2 \exp\left[-2\phi\left(\frac{\epsilon_i^2}{12}\frac{L}{s(\sigma)}[M^TM]_{ii}\right)\right] \\
&= \frac{12}{e^2}\frac{s(\sigma)}{L}\sum_{i=1}^{d}\frac{1}{[M^TM]_{ii}} \\
&\gtrsim \frac{s(\sigma)}{L}\,\mathrm{Tr}((M^TM)^{-1}).
\end{aligned}$$

Case 2: $\mathrm{Tr}((M^TM)^{-1}) > \frac{1}{3}\frac{L}{s(\sigma)}$

This case is the more difficult of the two. We shall make use of the following technical lemma.

Lemma 25. *Let $(a_i)_{i=1,\ldots,d}$ be any positive sequence satisfying $\sum_{i=1}^{d} a_i^{-1} > 4$. Then, there exists a non-negative sequence $(\epsilon_i)_{i=1,2,\ldots,d}$ such that $\sum_{i=1}^{d}\epsilon_i^2 \leq 4$ and $\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} \geq 2e^{-2}$.*

Proof. Without loss of generality, assume that $a_1 \geq a_2 \geq \cdots \geq a_d > 0$. If $a_1 \leq 1/4$, then taking $(\epsilon_1, \epsilon_2, \ldots, \epsilon_d) = (2, 0, 0, \ldots, 0)$, and noticing that $\phi(\cdot)$ is an increasing function, we conclude

$$\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_1)} = 4e^{-2\phi(4a_1)} \geq 4e^{-2\phi(1)} > 2e^{-2}.$$

Now, in the following we assume that $a_1 > 1/4$. Let t denote the largest integer $k \in \{1, \ldots, d\}$ satisfying $\sum_{i=1}^{k} a_i^{-1} \leq 4$. By the assumption that $\sum_{i=1}^{d} a_i^{-1} > 4$, we know that there always exists such a t, and t will satisfy $t < d$. We set

$$\epsilon_i = \begin{cases} a_i^{-1/2} & \text{if } 1 \leq i \leq t \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \ldots, d. \tag{5.23}$$

By definition, $\sum_{i=1}^{d}\epsilon_2^2 = \sum_{i=1}^{t} a_i^{-1} \leq 4$ satisfies (5.18). This procedure results in

$$\sum_{i=1}^{d}\epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} = e^{-2}\sum_{i=1}^{t}\frac{1}{a_i}.$$

If $\sum_{i=1}^t a_i^{-1} \geq 2$, we can immediately see from the above and (5.23) that $\sum_{i=1}^d \epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} \geq 2e^{-2}$.

On the other hand, if $\sum_{i=1}^t a_i^{-1} < 2$, this implies that $a_{t+1}^{-1} \geq 2$. In this case, we simply take $\epsilon_{t+1} = 2$, and take $\epsilon_i = 0$ for $i \neq t+1$. With this choice, we have

$$\sum_{i=1}^d \epsilon_i^2 e^{-2\phi(\epsilon_i^2 a_i)} = 4e^{-2\phi(4a_{t+1})} \geq 4e^{-2\phi(2)} = 2e^{-2}.$$

The above discussion concludes the proof of Lemma 25. □

By considering the values $a_i = \frac{L}{12s(\sigma)}[M^T M]_{ii}$, Lemma 25 ensures the existence of choices of $(\epsilon_i)_{i=1,\dots,d}$ satisfying (5.18) and, together with (5.22), gives

$$\mathbb{E}|\hat{\theta} - \theta|^2 \gtrsim 1.$$

This completes the proof of Theorem 21.

Proof of Theorem 23

Our proof follows as an extension of the proof of Theorem 21. We begin by stating the following lower bound.

Lemma 26. *Suppose observations $X \in \mathbb{R}^n$ are generated via the GLM (5.3) satisfying $\Phi'' \leq L$ for some constant $L > 0$. Let $\pi = \mathcal{N}(0, \beta^2 I_d)$. Then, the following lower bound on entropic error of the i-th component holds for any estimator $\hat{\theta}_i : \mathbb{R} \to \mathbb{R}$ and any fixed constant $t \geq 0$,*

$$\begin{aligned}\exp\left(2h(\sqrt{t}\theta_i|\hat{\theta}_i)\right) &\geq 2\pi e \frac{t\beta^2}{1 + \beta^2 \frac{L}{s(\sigma)}[M^T M]_{ii}} \\ &\to 2\pi e \frac{s(\sigma)}{L} \frac{t}{[M^T M]_{ii}} \quad \text{as } \beta^2 \to \infty \quad i = 1, \dots, d\end{aligned}$$

Proof. The inequality is trivially true when $t = 0$, and henceforth we will assume $t > 0$. For any $i = 1, \dots, d$, recall from Lemma 24 and (5.6) that

$$\mathcal{I}_X(\theta_i) \leq [\bar{\mathcal{I}}_X(\theta)]_{ii} = \frac{1}{s^2(\sigma)} \sum_{j=1}^n \left(M_{ji}^2 \operatorname{Var}(X_j)\right).$$

By applying the transformation $\theta \leftarrow \sqrt{t}\theta_i$, we get

$$\begin{aligned}\mathcal{I}_X(\sqrt{t}\theta_i) &= \frac{1}{t}\mathcal{I}_X(\theta_i) \\ &\leq \frac{1}{ts^2(\sigma)} \sum_{j=1}^n \left(M_{ji}^2 \operatorname{Var}(X_j)\right) \\ &\leq \frac{L}{ts(\sigma)}[M^T M]_{ii},\end{aligned}$$

where we recall that the variance of X_i is given as in Lemma 20. It therefore follows that the following holds for any $\hat{\theta}_i : \mathbb{R} \to \mathbb{R}$ with an application of Theorem 4

$$\begin{aligned}
\exp\left(2h(\sqrt{t}\theta_i|\hat{\theta}_i)\right) &= \exp\left(2h(\sqrt{t}\theta_i) - 2I(\sqrt{t}\theta_i;\hat{\theta}_i)\right) \\
&= 2\pi e \mathrm{Var}(\sqrt{t}\theta_i)\exp\left(-2I(\sqrt{t}\theta_i;X)\right) \\
&\geq 2\pi e \frac{t\beta^2}{1+\beta^2\frac{L}{s(\sigma)}[M^TM]_{ii}} \\
&\to 2\pi e \frac{s(\sigma)}{L}\frac{t}{[M^TM]_{ii}} \quad \text{as } \beta^2 \to \infty,
\end{aligned}$$

which concludes the proof of Lemma 26. □

Moving onwards, we will write $M = USV^T$ as the SVD of M and define $\lambda_i := \lambda_i(M^TM)$. Let us first examine what happens by taking $\theta \sim \pi = \mathcal{N}(0, \beta^2 \mathrm{I}_d)$. Given a fixed $i \in \{1, \ldots, d\}$, it follows from Lemma 26 that for any estimator $\hat{\theta}_i : \mathbb{R}^n \to \mathbb{R}$,

$$\begin{aligned}
\lambda_i \mathbb{E}(\hat{\theta}_i - \theta_i)^2 &\geq \frac{1}{2\pi e}\exp\left(2h(\sqrt{\lambda_i}\theta_i|\hat{\theta}_i)\right) \\
&\geq \frac{\lambda_i\beta^2}{1+\beta^2\frac{L}{s(\sigma)}[M^TM]_{ii}} \quad \text{if } \lambda_i > 0 \text{ and } i = 1, \ldots, d. \qquad (5.24)
\end{aligned}$$

In the case $\lambda_i = 0$, (5.24) holds trivially, and we can say that (5.24) holds for all $\lambda_i \geq 0$. This allows us to write

$$\mathbb{E}|US\hat{\theta} - US\theta|^2 = \sum_{i=1}^n \lambda_i \mathbb{E}(\hat{\theta}_i - \theta_i)^2 \geq \sum_{i=1}^n \frac{\lambda_i\beta^2}{1+\beta^2\frac{L}{s(\sigma)}[M^TM]_{ii}}. \qquad (5.25)$$

In other words, given a measurement matrix M with $M = USV^T$,

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}|US\hat{\theta}' - US\theta'|^2 \geq \sum_{i=1}^n \frac{\lambda_i\beta^2}{1+\beta^2\frac{L}{s(\sigma)}[M^TM]_{ii}}.$$

By simply noting that the minimax risk under the GLM with a measurement matrix $M = USV^T$ is equivalent to the the minimax risk under the GLM with a measurement matrix $M' = US$, it therefore follows that

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}|US\hat{\theta} - US\theta|^2 \geq \sum_{i=1}^n \frac{\lambda_i\beta^2}{1+\beta^2\frac{L}{s(\sigma)}\lambda_i}.$$

Consequently,

$$\inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}|M\hat{\theta} - M\theta|^2 = \inf_{\hat{\theta}} \sup_{\theta\in\mathbb{R}^d} \mathbb{E}|US\hat{\theta} - US\theta|^2 \geq \sum_{i=1}^n \frac{\lambda_i\beta^2}{1+\beta^2\frac{L}{s(\sigma)}\lambda_i}. \qquad (5.26)$$

Finally, take $\beta \to \infty$ and we yield

$$\frac{1}{n} \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E} \left| M\hat{\theta} - M\theta \right|^2 \geq \frac{s(\sigma)}{Ln} \text{rank}(M),$$

completing the proof.

Remark 17. *Bayes risk under prediction error for other log-concave priors with independent components can be derived with a similar procedure.*

Remarks on a weaker form of Theorem 23

As a historical remark, we note that in an earlier paper [Lee and Courtade, 2020] the following weaker result was established.

Theorem 27. *Suppose observations $X \in \mathbb{R}^n$ are generated via the GLM (5.3) satisfying $\Phi'' \leq L$ for some constant $L > 0$. With a fixed design matrix $M \in \mathbb{R}^{n \times d}$, the following holds for any estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^d$,*

$$\frac{1}{n} \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E} |M\hat{\theta} - M\theta|^2 \gtrsim \frac{s(\sigma)}{nL} \sum_{i=1}^{n} \frac{|m_i|^4}{|Mm_i|^2}. \tag{5.27}$$

The original proof of Theorem 27 is somewhat complicated and is omitted from this dissertation. We refer the interested reader to [Lee and Courtade, 2020] for more details.

Chapter 6

Lower bounds on risk under a general pairwise comparison framework

In the previous Chapter, we saw how our techniques can lead to tight Bayes and minimax risk lower bounds for the GLM. In this Chapter, we focus our attention to problems modeled in terms of pairwise comparisons, and discuss applications of our main results to a general pairwise comparison framework.

6.1 Introduction

Pairwise comparisons arise naturally: they may occur when a consumer is asked for their preference between two products, or when drivers on the road are asked to answer questions such as "*are you in a traffic jam?*", in which case answers are aggregated to improve navigation. Pairwise comparisons also are a key component in competitive sports; the wins and losses of all 30 NBA teams throughout a season can be thought of as a set of observations based on pairwise comparisons, for example.

Pairwise comparison scenarios in practice do not necessarily produce binary outcomes. For instance, comparisons among students in a classroom may consist of test scores that are in the set of non-negative half-integers, say, $\{0, 0.5, \ldots, 99.5, 100\}$. Another example is measuring the athletic abilities of two players in a game of badminton where they play to 21 points (ignoring deuces): a win with a difference of 15 points suggests more confidently that the first player is stronger than does a win with a difference of 3 points. In both examples, non-binary models are more informative than a simple binary "win-or-lose" model.

In a typical parametric framework for pairwise comparisons, items are assigned weights, and the difference $w_a - w_b$ between weights associated to two items a and b plays a key role in determining an observation; the larger the weight difference, the easier it should be to differentiate item a from item b. With this in mind, we develop our results on a general framework which we call the **General Pairwise Model (GPM)** where an observation between items a and b are generated according to a density in the exponential family. We

remark that the terminology *General Pairwise Model* is adopted to evoke an analogy to the well-known class of generalized linear models, describing the connection between linear models and exponential families; it is not meant to suggest that it is the most general parametric model one could possibly describe.

Definition 2 (General Pairwise Model)**.** *Consider two items a and b. We say that an observation $X \in \mathcal{X} \subseteq \mathbb{R}$ based on a comparison between items a and b is generated according to the General Pairwise Model if the density $f(\cdot; w)$ can be expressed as*

$$f(x; w) = h(x) \exp\left[G\left(\frac{w_a - w_b}{\sigma}\right)\frac{x}{\sigma} - \Phi\left(G\left(\frac{w_a - w_b}{\sigma}\right)\right)\right] \quad x \in \mathcal{X}. \tag{6.1}$$

In other words, comparisons between items a and b are generalized linear models with natural parameter $G\left((w_a - w_b)/\sigma\right)$ for some function $G(\cdot) : \mathbb{R} \to \mathbb{R}$ and noise parameter $\sigma \in \mathbb{R}$. The parameter σ is assumed to be fixed and the functions Φ and G may include dependence on σ, as we will see in the ordinal model later. The function $h : \mathcal{X} \to [0, \infty)$ is a base measure.

Note that the above definition is invariant to translations of weights by a constant factor. Hence, we make the standard assumption that the weights are centered; i.e., $\sum_a w_a = 0$, where the sum ranges over all items a for comparison.

Next, let us go through a list of popular pairwise comparison models that the General Pairwise Model includes.

Gaussian pairwise cardinal model

The observation X for a comparison between items a and b is generated by

$$X = w_a - w_b + Z, \tag{6.2}$$

where $Z \sim \mathcal{N}(0, \sigma^2)$ is independent Gaussian noise. The Gaussian pairwise cardinal model corresponds to the GPM with $\Phi(t) = t^2/2$ and $G(u) = u$.

Remark 18. *The above Gaussian pairwise cardinal model (and many other models within the GPM) has continuous outputs. This is a benefit of the GPM stated using the exponential family.*

At the other end of the spectrum, pairwise comparisons with binary outputs have a long history, initiated by Thurstone [1927] in the 1920's. The *Thurstone (Case V) model* has led to fruitful applications (see, for example, [Krabbe, 2008, Herbrich et al., 2007, Nosofsky, 1985, Swets, 1973]), and can be written in the following form.

Thurstone (case V) model

The observation X for a comparison between items a and b is generated by

$$X = \mathbf{1}\left(w_a - w_b + Z\right),$$

where $Z \sim \mathcal{N}(0, \sigma^2)$ corresponds to observation noise. Here, $\mathbf{1}(\cdot) : \mathbb{R} \to \mathbb{R}$ refers to an indicator function with $\mathbf{1}(t) = 1$ if $t > 0$ and $\mathbf{1}(t) = -1$ otherwise.

Another example is the *Bradley-Terry-Luce (BTL) model* due to Bradley and Terry [1952] and Luce [1959], which also has broad applications [Loewen et al., 2012, Heldsinger and Humphry, 2010, Nosofsky, 1985].

Bradley-Terry-Luce (BTL) model

The observation X for a comparison between items a and b is generated with

$$\Pr(X = 1) = \frac{1}{1 + \exp\left(-\frac{w_a - w_b}{\sigma}\right)}$$

and $\Pr(X = -1) = 1 - \Pr(X = 1)$.

Both the Thurstone and BTL models are special cases of the *ordinal model*, proposed by Shah et al. [2016].

Ordinal model

The observation X for a comparison between items a and b is generated with

$$\Pr\left(X = 1\right) = F_{\text{ord}}\left(\frac{w_a - w_b}{\sigma}\right) \tag{6.3}$$

and $\Pr\left(X = -1\right) = 1 - \Pr\left(X = 1\right)$ for some function $F_{ord} : \mathbb{R} \to [0, 1]$. The ordinal model is a special case of the GPM, corresponding to the choice $\Phi(t) = -\frac{t}{\sigma} + \log(1 + e^{\frac{2t}{\sigma}}), t \in R$ and $G(\cdot) = \frac{\sigma}{2} \log\left(F_{\text{ord}}(\cdot)/(1 - F_{\text{ord}}(\cdot))\right)$. Consequently, the Thurstone and the BTL model are special cases of the GPM.

6.2 Our setup

Suppose there are a total of d items and we are given n independent observations generated from the GPM (6.1) based on n pairs of items. It is helpful to envision which pairs are compared with the assistance of a *measurement matrix* $M \in \mathbb{R}^{n \times d}$. To be precise, for the i-th observation, suppose items a and b are compared. Then, we can represent this comparison by a *measurement vector* $m_i \in \mathbb{R}^d$ with the a-th entry equal to $+1$ and the b-th entry equal to -1 with all other entries equal to 0. Together, the collection of vectors $x_1, \ldots, x_n$ can be concatenated to form the measurement matrix M, with the i-th row corresponding to the measurement vector m_i.

The measurement matrix M can be viewed as a graph where nodes a and b connected if items a and b are compared. Our results are closely connected with the **Laplacian** L of M,

defined to be

$$L := \frac{1}{n} M^T M. \tag{6.4}$$

We let $L^\dagger$ denote the pseudoinverse of L.

Remark 19. *Our results depend only on the eigenvalues of L. Flipping the signs of any row of M will not change $M^T M$ and therefore will preserve the eigenvalues of L. Hence, for an observed pair (a, b), it is not important if item a is labeled as $+1$ or -1.*

We assume that observations are independent, i.e., we observe independent observations $x_1, x_2, \ldots, x_n$ generated from the GPM based on pairs chosen via the measurement vectors $m_1, m_2, \ldots, m_n$ respectively.

The question to answer is the following: *To what extent can we estimate $w_1, \ldots, w_d$ given observations $x_1, \ldots, x_n$?*

One way to answer this question is to establish minimax lower and upper bounds for the estimation of the true vector of weights w^*. In other words, we would like to provide quantitative analysis of the minimax ℓ_2 error defined as

$$\inf_{\hat{w}} \sup_{w^* \in \mathcal{W}} \mathbb{E}|\hat{w} - w^*|^2,$$

where $\mathcal{W}$ is the support of w^* to be considered. Here, recall that the minimax error characterizes the performance (in terms of ℓ_2 error) of the "best" estimator for the "worst-case" scenario. For ease of expressing our results, we define

$$\mathcal{W}_B := \{w : w \in \mathbb{B}_2^d(B), \sum_a w_a = 0\}$$

as the intersection of the subspace $\{w : \sum_a w_a = 0\}$ and the ℓ_2 ball with radius B, defined as

$$\mathbb{B}_2^d(B) := \{w \in \mathbb{R}^d : |w| \leq B\}. \tag{6.5}$$

Related work

There is a vast amount of work for parametric pairwise comparisons. Minimax bounds for the ordinal model were developed by Shah et al. [2016], who provided deep insights into the connection of minimax rates with the topology of the ordinal model, expressed in terms of the eigenvalues of the Laplacian L. Apart from Shah et al. [2016], Negahban et al. [2017, 2012] focused on a particular setting for the BTL model where comparisons are evenly distributed. A work by Hajek et al. [2014] considered the Plackett-Luce model [Plackett, 1975, Luce, 1959], which is an extension of the BTL model to comparing two or more items at a time: They were able to derive lower bounds in terms of the degree of vertices in the

comparison graph, which is suboptimal. Lower bounds for *unbiased* estimators that have the scaling of $\mathrm{Tr}(L^\dagger)$ were derived for the Plackett-Luce model by Hajek et al. [2014] via the Cramér-Rao bound, but they were unable to prove it for biased estimators. A recent work by Hendrickx et al. [2020] proves a lower bound for the BTL model with the same scaling but in a considerably restrictive setting where *all* pairs are compared k times (with k large enough). In our work, we present a lower bound that holds for *all* estimators, any configuration of the measurement matrix M, and has the scaling $\mathrm{Tr}(L^\dagger)$ for the entire class of GPMs with minor assumptions and almost no restrictions.

Another related line of work relates the estimation of bivariate isotronic matrices, where the goal is to model observations based on probabilities referenced by a $d \times d$ matrix and find good estimates of this matrix [Pananjady et al., 2020, Shah et al., 2019, Mao et al., 2018, Heckel et al., 2018]. There are also works regarding *ranking*, i.e., finding the ordering of items from highest weight to lowest weight; for example, Heckel et al. [2019] discusses active ranking with adaptive measurements. These settings are different from the parametric setup we consider.

6.3 Main results and discussion

Prior to introducing our main results, we state the notation that will be used throughout the section.

Notation

Consistent with notation used throughout the dissertation, we will use the symbol $\gtrsim$ (resp. $\lesssim$, $\asymp$) to express $\geq$ (resp. $\leq, =$) up to absolute constants that do not depend on any parameters. Given a vector $v \in \mathbb{R}^d$, we will write $v := (v_1, v_2, \ldots, v_d)$ where v_i is the i-th coordinate of the vector v. We will also write $v^k := (v_1, v_2, \ldots, v_k)$ as the vector corresponding to the first $k \leq d$ components of v. The $\log(\cdot)$ function is understood to be the natural logarithm.

Given two vectors u, v of the same length, we write $\langle u, v \rangle := u^T v$ as the inner product between u and v. We write 1_d (and 0_d) as the all-ones (and all-zero) vector of length d. Given a square matrix $M \in \mathbb{R}^{d \times d}$, we write the eigenvalues of M in increasing order: $\lambda_1(M) \leq \lambda_2(M) \leq \ldots \leq \lambda_d(M)$.

Minimax lower bounds

Our results are stated in terms of the following quantity $\mathcal{U}$.

Assumption 3. *The natural parameter $G(\cdot)$ and cumulant function $\Phi(\cdot)$ of the GPM (6.1) satisfies*

$$\left(G'(u)\right)^2 \Phi''(G(u)) \leq \mathcal{U} \quad \text{for all } u \in \mathbb{B}_2^d(2B/\sigma)$$

for some positive constant $\mathcal{U} > 0$.

This assumption is mild, and $\mathcal{U}$ can typically be viewed as a constant. To see some examples, the Gaussian pairwise cardinal model corresponds to $G(u) = u$ and $\Phi(t) = t^2/2$, and hence $\mathcal{U} = 1$. For the ordinal model, it can be shown that a bound on the log-concavity of $F_{\text{ord}}(\cdot)$ and $1 - F_{\text{ord}}(\cdot)$ of the form $-\frac{d^2}{dt^2} \log F_{\text{ord}}(t) \leq \mathcal{U}$ and $-\frac{d^2}{dt^2} \log(1 - F_{\text{ord}}(t)) \leq \mathcal{U}$ implies Assumption 3 holds with the same constant $\mathcal{U}$. For instance, the BTL model has $F_{\text{BTL}}(t) = 1/(1 + e^{-t})$ with

$$\max\left\{(-\log F_{\text{BTL}}(t))'', (-\log(1 - F_{\text{BTL}}(t)))''\right\} = \frac{e^{-t}}{(1 + e^{-t})^2} \leq \frac{1}{4},$$

and hence for the BTL model, Assumption 3 is satisfied with $\mathcal{U} = 1/4$. We remark that an assumption on an upper bound of Φ'' is typical for exponential family analysis; see [Loh and Wainwright, 2015, Negahban et al., 2012, Müller and Stadtmüller, 2005] for examples.

Our main result is a tight minimax lower bound for GPMs.

Theorem 28. *Given measurement matrix $M \in \mathbb{R}^{n \times d}$ and weight vector $w \in \mathcal{W}_B$, suppose we have n independent observations generated via the GPM (6.1) satisfying Assumption 3 for a constant $\mathcal{U} > 0$. The minimax ℓ_2 risk is lower bounded by*

$$\begin{aligned}\inf_{\hat{w}} \sup_{w \in \mathcal{W}_B} \mathbb{E}|\hat{w} - w|^2 &\gtrsim \min\left(B^2, \frac{\sigma^2}{\mathcal{U}n} \sum_{i=2}^{d} \frac{1}{\lambda_i(L)}\right) \\ &\gtrsim \min\left(B^2, \frac{\sigma^2}{\mathcal{U}n} \operatorname{Tr}(L^\dagger)\right).\end{aligned}$$

Here, L is the Laplacian of M defined in (6.4).

We adopt the convention that $1/\lambda_2(L) = +\infty$ if $\lambda_2(L) = 0$, which occurs when M is not connected. It should also be noted that M satisfies $L1_d = 0_d$, and hence $\lambda_1(L) = 0$. We remark that random guessing or simply guessing $\hat{w} = 0_d$ achieves the minimax rate when $\frac{\sigma^2}{\mathcal{U}n} \operatorname{Tr}(L^\dagger) > B^2$.

A special case of Theorem 28: the ordinal model

The topological dependence of Theorem 28 leads to interesting results for the ordinal model. Before proceeding, we first state known minimax bounds due to Shah et al. [2016].

Suppose $\|w\|_\infty \leq r$ and further define the auxiliary variable ζ for the function $F_{\text{ord}}(\cdot)$ corresponding to the ordinal model (see (6.3)) as

$$\zeta := \frac{\max_{x \in [0, 2r/\sigma]} F'_{\text{ord}}(x)}{F_{\text{ord}}(2r/\sigma)(1 - F_{\text{ord}}(2r/\sigma))}. \tag{6.6}$$

The following lower bound for the ordinal model is established by Shah et al. [2016].

Lemma 29. *[Shah et al., 2016, Theorem 2] Fix a constant $r > 0$. For a sample size $n \geq \frac{c_2\sigma^2 \operatorname{Tr}(L^\dagger)}{\zeta r^2}$, any estimator $\hat{w}$ based on n samples from the ordinal model satisfies*

$$\sup_{\|w\|_\infty \leq r} \mathbb{E}|\hat{w} - w|^2 \gtrsim \frac{\sigma^2}{n} \max\left\{ d^2, \max_{d' \in \{2,\ldots,d\}} \sum_{i=\lceil 0.99d' \rceil}^{d'} \frac{1}{\lambda_i(L)} \right\}. \tag{6.7}$$

As discussed by Shah et al. [2016] and Negahban et al. [2012], in situations of practical interest the parameters $B, r, \sigma, \zeta, \mathcal{U}$ are independent of n and d, and henceforth we will view them as constants.

Despite Lemma 29 being stated in terms of the ℓ_∞ norm, our Theorem 28 (stated in terms of the ℓ_2 norm) is strong enough to establish the following uniform improvement.

Corollary 30. *Any estimator $\hat{w}$ based on n samples from the ordinal model satisfies*

$$\sup_{\|w\|_\infty \leq r} \mathbb{E}|\hat{w} - w|^2 \gtrsim \frac{\sigma^2}{n} \max\left\{ d^2, \sum_{i=2}^{d} \frac{1}{\lambda_i(L)} \right\}. \tag{6.8}$$

Proof. To see this, take $B = r$, and we find that Theorem 28 leads to

$$\sup_{\|w^*\|_\infty \leq r} \mathbb{E}|\hat{w} - w^*|^2 \gtrsim \min\left\{ r^2, \frac{\sigma^2}{\mathcal{U}n} \sum_{i=2}^{d} \frac{1}{\lambda_i(L)} \right\}. \tag{6.9}$$

Now, notice there is a sample size constraint present in Lemma 29. By using the fact that $\operatorname{Tr}(L^\dagger) \geq \frac{d^2}{4}$ (see, for example, Lemma 14 of Shah et al. [2016]), we see whenever Lemma 29 holds, we would consequently have

$$r^2 \gtrsim \frac{\sigma^2}{\zeta n} \operatorname{Tr}(L^\dagger) \gtrsim \frac{\sigma^2 d^2}{n} \frac{1}{\zeta}. \tag{6.10}$$

Combining (6.10) with (6.9) we conclude the corollary. □

Note that Corollary 30 implies a uniform improvement over the result of Shah et al. [2016]. It is not hard to find situations where (6.8) is order-wise tighter than (6.7). For example, if $\lambda_k = Ck$ with $C = 2(\sum_{i=2}^{d} i)^{-1} \approx 4/d^2$, then (6.8) is of the order $\sigma^2 d^2 \log d/n$ while (6.7) is of the order $\sigma^2 d^2/n$. It should be noted that our result does not require a limitation on sample size n as required in Lemma 29. We also note that Shah et al. [2016] conjectured that the lower and upper bounds for the ordinal model should scale as $\operatorname{Tr}(L^\dagger)$; Theorem 28 confirms the lower bound part of this conjecture.

Further discussion of tightness

In this section, we will discuss performance guarantees of estimators for several settings of the GPM, and use them to support the tightness of our lower bounds. We will be assuming that the comparison graph induced by observation matrix M is connected; otherwise, there is at least one weight that cannot be estimated well.

Gaussian pairwise cardinal model

Suppose we observe n independent pairwise comparisons generated according to (6.2) with a weight vector $w \in \mathcal{W}_B$ and a measurement matrix M. The following error guarantee for the unconditional least-squares estimator is classical; we refer the interested reader to Shah et al. [2016] for a proof.

Lemma 31. *Suppose we have n observations based on the observation matrix M and a weight vector w satisfying $\langle 1, w\rangle = 0$ under the Gaussian pairwise cardinal model of (6.2). Then the unconditional least-squares estimator $\hat{w} := \frac{1}{n} L^\dagger M^T X$ satisfies*

$$\mathbb{E}|\hat{w} - w|^2 \leq \frac{\sigma^2}{n} \operatorname{Tr}(L^\dagger).$$

When $\frac{\sigma^2}{n} \operatorname{Tr}(L^T) \geq B^2$ we can use the estimator $\hat{w}' := 0_d$ to achieve an error upper bound of $\mathbb{E}|\hat{w}' - w|^2 \leq B^2$. Combined with Lemma 31, we immediately see that our lower bound of Theorem 32 is matched.

A direct consequence can be stated in terms of the optimal model selection problem, where we let $(\mathcal{G}, \mathcal{P})$ be the set of all possible functions $(G(\cdot), \Phi(\cdot))$ satisfying Assumption 3, and observations are generated under the GPM of (6.1) with a measurement matrix M. The risk associated with selections of $(G(\cdot), \Phi(\cdot))$ is therefore established to be

$$\inf_{(G,\Phi)\in\mathcal{G},\mathcal{P}} \inf_{\hat{w}} \sup_{w\in\mathcal{W}_B} \mathbb{E}|\hat{w} - w|^2 \asymp \min\left(B^2, \frac{\sigma^2}{\mathcal{U}n} \operatorname{Tr}(L^\dagger)\right),$$

with equality holding for the Gaussian pairwise cardinal model (which satisfies Assumption 3 with $\mathcal{U} = 1$). This suggests that Theorem 28 cannot be improved in general, unless finer-grained information about the model (e.g., further properties of G, Φ) is incorporated into the lower bound.

Subgaussian GPMs

Under additional constraints on subgaussianity and strong-convexity of the negative log-likelihood of the observations X (detailed in the next section), it can be shown that in many natural settings the Maximum Likelihood Estimator (MLE) has ℓ_2 error matching our lower bound of Theorem 28 up to constants, providing further evidence of tightness.

Theorem 32. *Suppose we have n observations based on the connected observation matrix M and a weight vector $w \in \mathcal{W}_B$ for some $B > 0$ under the General Pairwise model of (6.1) satisfying Assumption 3 and the two assumptions given in (6.12) and (6.13) for constants $\mathcal{U}, \mathcal{L}, R > 0$. Then, the MLE has guaranteed error performance*

$$\sup_{w\in\mathcal{W}_B} \mathbb{E}|\hat{w} - w|^2 \lesssim \sigma^2 \frac{\mathcal{U}}{\mathcal{L}^2} \frac{d}{n\lambda_2(L)}. \tag{6.11}$$

Theorem 32 recovers the performance guarantee for the ordinal model established by [Shah et al., 2016, Theorem 2] as a special case.

It is worthwhile to discuss when the lower bound of Theorem 28 and the upper bound of (6.11) matches. By observation, $\mathrm{Tr}(L^\dagger)$ has to be of the same order as $d/(n\lambda_2(L))$. In other words, the smallest $\Theta(d)$ positive eigenvalues of L will need to be of the same order. Examples of common measurement matrices that satisfy this are given below.

Random Matrices. When each measurement is sampled uniformly at random over all possible $\binom{d}{2}$ choices, all positive eigenvalues of the Laplacian matrix L will be of the same order with high probability when n is large enough compared to d; see, for example, [Wainwright, 2019].

Specific Graph Models. As discussed by Shah et al. [2016], examples of the measurement matrix M where the smallest $\Theta(d)$ positive eigenvalues of L are of the same order include the following: *complete graph* (every pair is compared once), *constant-degree expander* (see, e.g., constructions by Alon et al. [2008]), *star* (one item is compared to every other item once), or *complete bipartite* (all items are split into two sets; there is an edge between every pair of items in different sets and none between pairs in the same set). When M is designed as any of these models, our upper and lower bounds are tight.

Summary of contributions

We make the following two main contributions.

1. We discuss pairwise comparisons under the General Pairwise Model, a general parametric framework for the pairwise comparison paradigm. The GPM unifies pairwise comparison with continuous and discrete observations under the exponential family. As special cases, our framework covers classic models such as the Thurstone (Case V) model, BTL model, ordinal model and Gaussian pairwise cardinal model.

2. We establish an ℓ_2 minimax lower bound scaling as $\mathrm{Tr}(L^\dagger)$ for the GPM. We have matching lower and upper bounds for the Gaussian pairwise cardinal model, suggesting our lower bound cannot in general be improved. When applied to the ordinal model as a special case, our minimax lower bound achieves uniform improvement over the lower bound of Shah et al. [2016] and is the first proof of one direction of a conjecture set in their paper. We establish performance guarantees for the maximum likelihood estimator under the GPM with subgaussian constraints, which serves as evidence that our minimax lower bound is tight.

Finally, we remark that Theorem 28 is a uniform bound over the General Pairwise Model cleanly stated in terms of $\mathrm{Tr}(L^\dagger)$. Our lower bound technique can be extended without too much extra effort to general m-ary comparison frameworks, which is intended as future work.

6.4 Additional proofs

Additional assumptions for Theorem 32

In this section we discuss the additional assumptions required in the statement of Theorem 32:

1. **Strong-convexity of negative log-likelihood.** For any $u \in \mathbb{R}$ with $|u| \leq 2B/\sigma$, and for all $x \in \mathcal{X}$,

$$-\frac{\partial^2}{\partial u^2}\left(\frac{x}{\sigma}G(u) - \Phi\left(G(u)\right)\right) \geq \mathcal{L}. \tag{6.12}$$

   In other words, the negative log-likelihood of any observation generated by the General Pairwise Model of (6.1) is strongly-convex.

2. **Subgaussianity.** Consider observation X generated according to measurement matrix M and weight vector w from the General Pairwise Model of (6.1). Then, for all $1 \leq i \leq n$, the centered random variable $X_i - \mathbb{E}[X_i]$ is subgaussian with parameter s. We will be focusing on subgaussian X that satisfies

$$\frac{s^2R^2}{\mathcal{U}\sigma^2} \leq C, \tag{6.13}$$

   for some constant $C > 0$, and $R := \sup_{|u|\leq 2B/\sigma} |G'(u)|$.

Here, we say that a random variable $Z \in \mathbb{R}$ is r-subgaussian if $\mathbb{E}[Z] = 0$ and $\mathbb{E}\exp(tZ) \leq \exp\left(r^2t^2/2\right)$ for all $t \in \mathbb{R}$.

The first assumption is standard for analysis of MLEs. We note that (6.12) holds for a broad class of models; for example, if $G(u) = u, \forall u \in \mathbb{R}$ (satisfied by the Gaussian pairwise cardinal model), then (6.12) corresponds to a constraint of $\Phi''(\cdot) \geq \mathcal{L}$. This complements Assumption 3, which in this case can be written as an *upper bound* $\Phi''(\cdot) \leq \mathcal{U}$. For the ordinal model, (6.12) is implied by strong log-concavity of $F_{\text{ord}}(\cdot)$, i.e., $-\frac{d^2}{dt^2}\log F_{\text{ord}}(t) \geq \mathcal{L}, \forall t \in \mathbb{R}$, which is assumed by Shah et al. [2016].

On the other hand, we remark that not all models within the exponential family are subgaussian (for example, the exponential density is not subgaussian), and in this section we will be focusing our attention to subgaussian models. Examples include models with bounded outputs (such as the ordinal model), which are subgaussian by nature. We further remark that the subgaussian constraint of (6.13) is not restrictive: recall that $s, \mathcal{U}, \sigma$ and R are all parameters of the General Pairwise Model, and are typically considered to be constants. As an example, under the ordinal model where $G(u) = \sigma\log\left(F_{\text{ord}}(\cdot)/(1 - F_{\text{ord}}(\cdot))\right)$ and $s \leq 1$, the value of R^2/σ^2 is precisely ζ^2 (see (6.6)), and can be typically considered an $O(1)$ value that does not depend on d and n.

Proof of Theorem 28

Recall that the observations X are generated according to n independent observations $X_1, X_2, \ldots, X_n$ from the GPM (6.1) with measurement matrix M and weight $w \in \mathcal{W}_B$ based on the density function

$$f(x; w) := \prod_{i=1}^{n} f(x_i; w) \quad x \in \mathbb{R}^n,$$

where $f(x_i; w)$ can be written as

$$f(x_i; w) = \exp\left(G\left(\frac{\langle m_i, w\rangle}{\sigma}\right)\frac{x_i}{\sigma} - \Phi\left(\frac{\langle m_i, w\rangle}{\sigma}\right)\right) \quad x_i \in \mathbb{R},\ i = 1, 2, \ldots, n,$$

with $m_i \in \mathbb{R}^d$ corresponding to the i-th row of the measurement matrix.

The first step of the proof is to bound the minimax entropic risk in terms of the Bayes entropic risk, where

$$\inf_{\hat{w}} \sup_{w \in \mathcal{W}_B} \mathbb{E}|\hat{w} - w|^2 \geq \inf_{\hat{w}} \sup_{W \sim \pi} \mathbb{E}|\hat{w} - W|^2$$

given a prior π defined on $\mathcal{W}_B$. In order to specify the prior we choose, let us first write the SVD of M as $M = USV_0^T$. Define $V_0 := \begin{bmatrix} V & v_d \end{bmatrix}$ where $V = \begin{bmatrix} v_1 & \ldots & v_{d-1} \end{bmatrix} \in \mathbb{R}^{d \times (d-1)}$ and $v_d \in \mathbb{R}^d$. We will design prior π such that W is sampled according to

$$W = V_0 \begin{bmatrix} \theta \\ z \end{bmatrix} = V\theta + z v_d \tag{6.14}$$

for a random vector θ on $\mathbb{R}^{d-1}$, with each component θ_i sampled according to an independent uniform distribution with support $[-t_i, t_i]$ (the specific values of t_i will be specified later). The term z is a deterministic value depending on θ, and is such that $\langle 1_d, W\rangle = 0$; a straightforward calculation yields $z = -\frac{1_d^T V\theta}{1_d^T v_d}$, which can be shown to be equal to 0.

Lemma 33. *If $\lambda_2(L) > 0$, we have $V^T 1_d = 0_{d-1}$ and consequently, the term $z = -\frac{1_d^T V\theta}{1_d^T v_d}$ satisfies $z = 0$.*

Proof. Recall that the vector 1_d satisfies $M1_d = 0_d^T$, and hence we can write

$$USV_0^T 1_d = 0.$$

This implies that

$$1_d^T V_0 S^T S V_0^T 1_d = 0,$$

or equivalently,

$$\sum_{i=1}^{d}(1_d^T v_i)^2 \lambda_{d-i+1}(L)^2 = 0.$$

Now, since $\lambda_i(L) > 0$ for all $i > 1$ and $\lambda_1(L) = 0$, it follows that

$$1_d^T v_j = 0 \quad \text{for all } 1 \leq j \leq d-1.$$

This is equivalent to

$$V^T 1_d = \begin{bmatrix} 1_d^T v_1 & 1_d^T v_2 & \dots & 1_d^T v_{d-1} \end{bmatrix}^T = 0_{d-1}$$

as desired. It therefore follows that $z = -\frac{1_d^T V\theta}{1_d^T v_d} = 0$. □

Given our choice of a log-concave prior π, we may write the following sequence of inequalities for any estimator $\hat{w}$

$$\begin{aligned}
&\mathbb{E}|\hat{w} - W|^2 \\
&\quad = \mathbb{E}|V_0^T\hat{w} - V_0^T W|^2 \\
&\quad \overset{(a)}{\geq} \mathbb{E}|V^T\hat{w} - \theta^{d-1}|^2 \\
&\quad \overset{(b)}{\gtrsim} \sum_{i=1}^{d-1} \exp\left(2h(\theta_i|(V^T\hat{w})_i)\right) \\
&\quad \overset{(c)}{\geq} \sum_{i=1}^{d-1} \exp\left(2h(\theta_i) - 2I(\theta_i; X)\right) \\
&\quad \overset{(d)}{\geq} \sum_{i=1}^{d-1} \exp\left(2h(\theta_i) - 2\phi\left(\mathrm{Var}(\theta_i)\mathbb{E}_{\theta_i}\mathcal{I}_X(\theta_i)\right)\right) \\
&\quad \overset{(e)}{\geq} \sum_{i=1}^{d-1} \exp\left(2h(\theta_i) - 2\phi\left(\mathrm{Var}(\theta_i)\mathbb{E}_{\pi}[\mathcal{I}_X]_{ii}\right)\right),
\end{aligned} \tag{6.15}$$

with $\phi(\cdot)$ defined in (5.20). Here, (a) follows since

$$\begin{aligned}
\mathbb{E}|V_0^T\hat{w} - V_0^T W|^2 &= \mathbb{E}|V^T\hat{w} - \theta^{d-1}|^2 + \mathbb{E}(v_d^T\hat{w} - z)^2 \\
&\geq \mathbb{E}|V^T\hat{w} - \theta^{d-1}|^2.
\end{aligned}$$

The inequality (b) should be familiar to us, but we remark it again for completness: it follows since Gaussians maximize entropy and conditioning reduces entropy, so that

$$\mathbb{E}(\alpha - \beta)^2 \gtrsim \exp(2h(\alpha - \beta)) \gtrsim \exp(2h(\alpha|\beta))$$

for two real-valued random variables α, β. In (c), we used the data processing Inequality on the Markov chain $\theta_i \to X \to (V^T\hat{w})_i$, which implies

$$\begin{aligned} I(\theta_i; X) &\geq I(\theta_i; (V^T\hat{w})_i) \\ &= h(\theta_i) - h(\theta_i|(V^T\hat{w})_i). \end{aligned}$$

In (d), for each $i = 1, 2, \ldots, d-1$ we bound $I(\theta_i; X)$ by using Theorem 7 and writing the density of X in terms of θ_i. Note that our choice of the prior of θ_i as the uniform distribution is log-concave by default. Finally, (e) is due to Lemma 24, which can be adapted to the GPM with the regularity of $\Phi(\cdot)$ and $G(\cdot)$.

Moving on, define the variable u_j as

$$u_j := \frac{\langle x_j, V\theta + zv_d\rangle}{\sigma} = \frac{\langle x_j, V\theta\rangle}{\sigma}.$$

From direct calculation, we arrive at

$$\begin{aligned} \mathbb{E}_\pi \left[\bar{\mathcal{I}}_X\right]_{ii} &\leq \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{n} (\Psi_j)^2 \middle| \theta\right]\right] \\ &\leq \sum_{j=1}^{n} \Phi''\left(G\left(u_j\right)\right)\left(\frac{\partial}{\partial\theta_i} G\left(u_j\right)\right)^2, \end{aligned} \tag{6.16}$$

where $\Psi_j := \frac{\partial}{\partial\theta_i}\left(\frac{x_j}{\sigma} G\left(u_j\right) - \Phi\left(G\left(u_j\right)\right)\right)$. In the above, we used the fact that the exponential family of (6.1) has mean $\sigma\Phi'(\eta)$ and variance $\sigma^2\Phi''(\eta)$ and hence

$$\mathbb{E}\left[\left(x_i - \sigma\Phi'\left(G(u_j)\right)\right)^2 \middle| \theta\right] = \sigma^2\Phi''\left(G\left(u_j\right)\right).$$

Now, note that we have by chain rule

$$\frac{\partial}{\partial\theta_i} G\left(u_j\right) = G'\left(u_j\right)\frac{\langle m_j, v_i\rangle}{\sigma}.$$

Combining this with (6.16) and Assumption (3) gives us

$$\mathbb{E}_\pi[\bar{\mathcal{I}}_X]_{ii} \leq \frac{\mathcal{U}}{\sigma^2}|Mv_i|^2 = \frac{\mathcal{U}}{\sigma^2}S_{ii}^2 = \frac{\mathcal{U}n}{\sigma^2}\lambda_{d-i+1}(L),$$

for all $i = 1, 2, \ldots, d-1$. Therefore, we have successfully extracted out the eigenvalues of the Laplacian L. Combining this with (6.15) and a calculation of $h(\theta_i)$ and $\mathrm{Var}(\theta_i)$ for $\theta_i \sim \mathrm{Unif}[-t_i, t_i]$, we see that

$$\mathbb{E}|\hat{w} - w|^2 \gtrsim \sum_{i=1}^{d-1} t_i^2 \exp\left[-2\phi\left(\frac{t_i^2}{3}\frac{\mathcal{U}n}{\sigma^2}\lambda_{d-i+1}(L)\right)\right]. \tag{6.17}$$

Now, it remains to find a good choice of the supports of the uniform distributions to yield the final result of Theorem 28. We start our discussion with the number of zero eigenvalues of L. As discussed in Section 6.3, since 1_d spans the null-space of L, we have $\lambda_1(L) = 0$. If $\lambda_2(L) = 0$, we can take $t_{d-i-1} = B$ and $t_j = 0$ for all $j \neq d-i-1$, and (6.17) recovers

$$\mathbb{E}|\hat{w} - w|^2 \gtrsim B^2 = \min\left(B^2, \frac{\sigma^2}{\mathcal{U}n}\sum_{i=2}^{d}\frac{1}{\lambda_i(L)}\right),$$

where recall that we let $1/\lambda_2(L) = +\infty$ by convention if $\lambda_2(L) = 0$. Hence, we will continue our discussion for the remaining case where $\lambda_2(L) > 0$. In other words, the only zero eigenvalue is $\lambda_1(L)$ and L is connected.

It remains to judiciously choose the values of t_i. Lemma 33 allows us to rewrite our constraint of $|W|^2 \leq B^2$ as $|\theta|^2 \leq B^2$, since $z = 0$ implies

$$|W|^2 = \left|V_0\begin{bmatrix}\theta\\ z\end{bmatrix}\right|^2 = |\theta|^2 + z^2 = |\theta|^2.$$

Hence, we would require our choices of t_i, $i = 1, \ldots, d-1$ satisfy

$$\sum_{i=1}^{d-1} t_i^2 \leq B^2. \tag{6.18}$$

We will continue our discussion in terms of the relation between $\mathrm{Tr}(L^\dagger)$ and $\frac{\mathcal{U}n}{\sigma^2 B}$.

If $\frac{\sigma^2}{\mathcal{U}n}\mathrm{Tr}(L^\dagger) \leq \frac{1}{4}B^2$: We can choose

$$t_i^2 = \frac{4\sigma^2}{\mathcal{U}n}\frac{1}{\lambda_{d-i+1}(L)} \quad \text{for all} \quad i = 1, \ldots, d-1.$$

This satisfies (6.18), and continuing on from (6.17) we have

$$\mathbb{E}|\hat{w} - w|^2 \gtrsim \frac{\sigma^2}{\mathcal{U}n}\sum_{i=1}^{d-1}\frac{1}{\lambda_{d-i+1}(L)} = \frac{\sigma^2}{\mathcal{U}n}\mathrm{Tr}(L^\dagger).$$

If $\frac{\sigma^2}{\mathcal{U}n}\mathrm{Tr}(L^\dagger) > \frac{1}{4}B^2$: Let's define the auxiliary variables $a_1, \ldots, a_{d-1}$ with

$$a_i := \frac{\mathcal{U}n}{4\sigma^2}\lambda_{d-i+1}(L^\dagger) \quad \text{for all} \quad i = 1, \ldots, d-1.$$

Then, $a_1, \ldots, a_{d-1}$ is a positive sequence satisfying $\sum_{i=1}^{d-1} a_i^{-1} > B^2$. We can directly use the following lemma which is a restatement of Lemma 25.

Lemma 34 (Restatement of Lemma 25)**.** *Fix $B > 0$. Let $(a_i)_{i=1,\dots,q}$ be any positive sequence with $\sum_{i=1}^{q} a_i^{-1} > B^2$. Then, there exists a non-negative sequence $(\epsilon_i)_{i=1,\dots,d}$ such that $\sum_{i=1}^{d} \epsilon_i^2 \le B^2$ and $\sum_{i=1}^{q} \epsilon_i^2 \exp(-2\phi(\epsilon_i^2 a_i)) \ge B^2 e^{-2}/2$.*

Then, we can choose $q = d - 1$ and $t_i = \epsilon_i$ based on Lemma 34. Hence, in this setting we conclude that

$$\mathbb{E}|\hat{w} - w|^2 \gtrsim \sum_{i=1}^{q} t_i^2 \exp(-2\phi(t_i^2 a_i)) \gtrsim B^2$$

as desired. Combining the two discussions for different regimes of $\frac{\sigma^2}{Un} \operatorname{Tr}(L^\dagger)$ concludes the proof.

Proof of Theorem 32

The Maximum Likelihood Estimator for the General Pairwise Model solves the optimization problem $\hat{w} := \arg\min_{w \in \mathcal{W}_B} \ell(w)$, where the objective function $\ell(w)$ is defined as

$$\ell(w) = -\frac{1}{n} \sum_{i=1}^{n} \left(G\left(\frac{\langle m_i, w \rangle}{\sigma} \right) \frac{x_i}{\sigma} - \Phi\left(G\left(\frac{\langle m_i, w \rangle}{\sigma} \right) \right) \right).$$

The analysis of the MLE here follows a standard treatment of M-estimators tailored to the General Pairwise Model, and is essentially a modification of the proof of the guarantee for the ordinal model due to [Shah et al., 2016, Theorem 2].

The semi-norm of a vector $v \in \mathbb{R}^d$ with respect to a PSD matrix $Q \in \mathbb{R}^{d \times d}$ is written as

$$\|v\|_Q := \sqrt{v^T Q v}.$$

For ease of notation, let us write the true weight vector as w^*. The following lemma given by Shah et al. [2016] allows one to bound the error in terms of L semi-norm by the $L^\dagger$ semi-norm of gradients.

Lemma 35. [Shah et al., 2016, Lemma 9] *Consider the estimator $\hat{w} = \arg\min_{w \in \mathcal{W}_B} \ell(w)$. Then, if $\ell(\cdot)$ satisfies the following κ-strong convexity condition*

$$\ell(w^* + \Delta) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \ge \kappa \|\Delta\|_L^2 \tag{6.19}$$

for all $\Delta \in \mathbb{R}^d$ with $w^* + \Delta \in \mathcal{W}_B$, then $\|\hat{w} - w^*\|_L \le \frac{1}{\kappa} \|\nabla \ell(w^*)\|_{L^\dagger}$.

Hence, the proof consists of two parts. We will first show that $\ell(w)$ satisfies the $\frac{\mathcal{L}}{\sigma^2}$-strong convexity condition. Then, we will bound the gradient in the $L^\dagger$ semi-norm.

A straightforward calculation for the function $\ell(w)$ yields

$$\nabla^2 \ell(w) = -\frac{1}{n\sigma^2} \sum_{i=1}^{n} \frac{\partial^2}{\partial u_i^2} \left(\frac{x}{\sigma} G(u_i) - \Phi\left(G(u_i)\right) \right) m_i m_i^T,$$

where u_i is defined to be $u_i = \frac{\langle m_i, w\rangle}{\sigma}$. By our assumption of (6.12), it follows that $\ell(w)$ satisfies

$$\begin{aligned} v^T \nabla^2 \ell(w) v &\geq \frac{\mathcal{L}}{n\sigma^2} |Mv|^2 \\ &= \frac{\mathcal{L}}{\sigma^2} \|v\|_L^2 \quad \text{for all } v, w \in \mathcal{W}_B. \end{aligned}$$

This implies that $\nabla^2 \ell(w)$ is strongly convex in the L semi-norm with parameter $\frac{\mathcal{L}}{\sigma^2}$, and we may use the following property of strong convexity:

$$\ell(w^* + \Delta) - \ell(w^*) - \langle \nabla \ell(w^*), \Delta \rangle \geq \frac{\mathcal{L}}{\sigma^2} \|\Delta\|_L^2.$$

Combining this with Lemma 35, we have

$$\|\Delta\|_L^2 \leq \frac{\sigma^4}{\mathcal{L}^2} \|\nabla \ell(w^*)\|_{L^\dagger}^2.$$

Moving on, we will try to bound $\|\nabla \ell(w^*)\|_{L^\dagger}^2$. First, we can write

$$\nabla \ell(w^*) = -\frac{1}{n\sigma} \sum_{i=1}^n \left(\frac{x_i}{\sigma} - \Phi'\left(G(u_i)\right) \right) G'(u_i) m_i.$$

The key here is to notice that $\sigma\Phi'(\cdot)$ is the mean of the exponential family of (5.2) with cumulant function $\Phi(\cdot)$. Hence, we can define a random vector $V \in \mathbb{R}^n$ such that V_i are independent sampled as a mean-shifted version of X/σ, i.e., $V_i = (X' - \mathbb{E}[X'])/\sigma$ for a i.i.d. copy of X' based on the General Pairwise Model with conditions same as in Theorem 32. Define G as the diagonal matrix with diagonal entries $G'(u_1), \ldots, G'(u_n)$. We can write $\nabla \ell(w^*) = -\frac{1}{n\sigma} M^T G V$, and upon defining $Q := \frac{\sigma^2}{\mathcal{L}^2 n^2} M L^\dagger M^T$, we have $\|\Delta\|_L^2 \leq V^T G^T Q G V$. We can use the fact that V_i has variance $\Phi''(u_i)$ and readily calculate

$$\mathbb{E} V^T G^T Q G V = \sum_{i=1}^n Q_{ii} \Phi''\left(u_i\right) G'(u_i)^2 \leq \frac{\mathcal{U}\sigma^2 d}{\mathcal{L}^2 n}.$$

Here, we used the fact that $\mathrm{Tr}(Q) = \frac{\sigma^2(d-1)}{\mathcal{L}^2 n}$ and Assumption 3. The remaining is an application of the Hanson-Wright inequality (see, e.g., [Shah et al., 2016, Lemma 13]), with the identities $\left|\left|\left| G^T Q G \right|\right|\right|_{\mathrm{fro}}^2 \leq (d-1)\frac{\sigma^4}{\mathcal{L}^4 n^2} R^4$ and $\left|\left|\left| G^T Q G \right|\right|\right|_{\mathrm{op}} \leq \frac{\sigma^2}{\mathcal{L}^2 n} R^2$. Here, the notations $|||\cdot|||_{\mathrm{fro}}$ and $|||\cdot|||_{\mathrm{op}}$ correspond to the Frobenius norm and operator norm respectively. Let K be the subgaussian constant of Y_i/σ, and note that we have $K\sigma = s$. Then, we have

$$\begin{aligned} \Pr\left(V^T G^T Q G V - \frac{\mathcal{U}\sigma^2 d}{\mathcal{L}^2 n} > t \right) &\leq \exp\left(-c \min\left\{ \frac{t^2 \mathcal{L}^4 n^2}{K^4 (d-1) \sigma^4 R^4}, \frac{t \mathcal{L}^2 n}{K^2 \sigma^2 R^2} \right\} \right) \\ &= \exp\left(-c \min\left\{ \frac{t^2 \mathcal{L}^4 n^2}{s^4 (d-1) R^4}, \frac{t \mathcal{L}^2 n}{s^2 R^2} \right\} \right) \end{aligned}$$

which implies there exists a universal constant $c' > 0$ such that

$$\Pr\left(\|\Delta\|_L^2 > t\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n}\right) \leq \exp\left(-c' \min\left(t^2\frac{\mathcal{U}^2\sigma^4 d}{s^4R^4}, t\frac{\mathcal{U}\sigma^2 d}{s^2R^2}\right)\right)$$

for all $t \geq 1$. It remains to integrate the above to bound $\mathbb{E}\|\Delta\|_L^2$, which is a matter of algebra. Recalling our assumption that $\frac{s^2R^2}{\mathcal{U}\sigma^2} \leq C$ for some constant $C > 1$, we can write

$$\begin{aligned}
\mathbb{E}\|\Delta\|_L^2 &\leq C\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n} + \int_{C\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n}}^{\infty} \Pr\left(\|\Delta\|_L^2 > t\right) dt \\
&= C\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n} + C\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n}\int_1^{\infty} \Pr\left(\|\Delta\|_L^2 > tC\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n}\right) dt \\
&\lesssim C\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n} + C\frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n}\int_1^{\infty} \exp\left(-c'C\frac{\mathcal{U}\sigma^2 d}{s^2R^2}t\right) dt \\
&\lesssim \frac{\mathcal{U}\sigma^2}{\mathcal{L}^2}\frac{d}{n}.
\end{aligned} \tag{6.20}$$

Finally, we can use the fact that the nullspace of L is spanned by 1_d and $\langle 1_d, \Delta\rangle = 0$ to conclude $\|\Delta\|_L^2 \geq \lambda_2(L)|\Delta|^2 = \frac{1}{n}\lambda_2(X^TX)|\Delta|^2$. Combining this with (6.20) completes the proof.

Chapter 7

Concluding remarks and outlook

While we have covered a variety of new results within the dissertation, there remains many interesting questions to be answered.

7.1 Outlook

We mark several avenues for possible future work.

1. Improved bounds with further optimization of reference measure

Given specific pairs (π, P_θ), it may be possible to optimize over *all* index measures μ (satisfying LSI) to get the tightest possible bound on $I(\pi; P_\theta)$.

2. Unification of Theorems 4 and 7

As discussed in Chapter 2, Theorems 4 and 7 stem from a similar choice of reference measure but with different analytical methodologies. These differences lead to pros and cons of either results. Theorem 4 preserves a log-determinant behavior that is consistent with well-known channel capacity expressions such as that of the Gaussian observation model, but does not capture the case where $\mathcal{I}_X \to 0$ well. On the other hand, Theorem 7 successfully captures the behavior in the $\mathcal{I}_X \to 0$ regime and is uniformly tighter than Theorem 4 when Σ_π and $\mathcal{I}_X$ are multiples of the identity matrix. However, Theorem 7 is of a log-trace form and does not capture well cases where Σ_π or $\mathcal{I}_X$ has extreme eigenvalues. It is therefore of interest whether there is a unified inequality that preserves the pros of both theorems.

3. Combination with advancements in the hyperplane conjecture

As discussed in Chapters 3 and 4, our new inequalities for log-concave priors have good synergy with the hyperplane conjecture, due to its entropic formulation by Bobkov and Madiman [2010]. Concurrent developments in the hyperplane conjecture have been rapid,

and while we cannot foresee the future, new developments on the hyperplane conjecture should lead to interesting applications with our results.

4. Extensions to other applications

We first note that any application of the van Trees inequality has the potential of being extended by replacing the van Trees inequality by inequalities developed within this dissertation. This leads to possibilities for improvements of previous and future developments. For example, recent work by Barnes et al. [2020] apply the van Trees inequality to lower bounds on minimax risk of learning distributions under communication constraints. It is interesting to see whether our bounds provide utility in similar scenarios. Moreover, Gill and Levit [1995] were able to apply the van Trees inequality to give lower bounds on minimax risk of nonparametric models through identifying and establishing lower bounds on minimax risk for difficult parametric sub-models. Successful applications to nonparametric problems range from linear estimators [van Rooij and Ruymgaart, 1996, Belitser and Levit, 1996] to estimation of quantum states [Lahiry and Nussbaum, 2022]. For all of these applications, our bounds can naturally be applied in a similar fashion and has the potential of providing improvements.

It is of sufficient interest to extend beyond this framework; for example, Barnes and Ozgur [2021] discuss *lower bounds* on mutual information expressed in terms of Fisher information under certain subgaussian constraints. While their setup is different from our parametric framework, there is potential that our *upper bounds* on mutual information may be a good complement.

In the context of machine learning, our lower bounds on risk can in some cases be interpreted as lower bounds on the training error in the speak of the machine learning community. It is of broad practical interest to ask whether our bounds on mutual information can be converted into bounds for generalization error, which has gained massive interest in recent times; see, e.g., [Pensia et al., 2018, Lopez and Jog, 2018, Rigollet, 2007, Murphy, 2005, Meir and Zhang, 2003] for some recent developments.

Finally, there are a wide variety of statistical models of practical interest that are closely related to the GLM, to which our techniques for bounding risk can be applied without too much difficulty. These models include, for example, matrix completion and recovery [Recht, 2011, Candes and Plan, 2011, 2010, Keshavan et al., 2010], covariance matrix estimation [Friedman et al., 2008, Cai et al., 2010] and phase retrieval [Cai et al., 2016, Candes et al., 2015, Candes et al., 2015, Fienup, 1982].

7.2 Concluding remarks

We first make a brief recap of the dissertation: We introduced a family of inequalities indexed by reference measures satisfying an LSI (Theorem 2). Under a suitably chosen Gaussian reference measure, we recover Efroimovich's inequality (Corollary 3) and consequently the

celebrated van Trees inequality (2.9). We established two new inequalities under log-concave priors (Theorems 4 and 7). Procedures of applying our new inequalities to lower bound risk via tools in rate distortion theory were shown in Chapter 4. We established a variety of lower bounds for the GLM and GPM is Chapters 5 and 6 respectively.

Some of the most interesting contributions that thread through the entire dissertation are the previously unknown connections between the LSI, information inequalities and statistical applications. This dissertation serves to give a glimpse of some of the exciting results under the framework of Theorem 2. Of course, the results presented here are not the end of the story and likely the tip of the iceberg. We believe there is strong potential for more interesting results to emerge through future work.

Bibliography

[1] Felix Abramovich and Vadim Grinshtein. Model Selection and Minimax Estimation in Generalized Linear Models. *IEEE Transactions on Information Theory*, 62(6):3721–3730, 2016.

[2] Aviv Adler, Jennifer Tang, and Yury Polyanskiy. Quantization of Random Distributions under KL Divergence. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 2762–2767. IEEE, 2021.

[3] James Aitchison. Goodness of Prediction Fit. *Biometrika*, 62(3):547–554, 1975.

[4] Hirotogu Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[5] Charalambos D. Aliprantis and Kim C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

[6] Noga Alon, Oded Schwartz, and Asaf Shapira. An Elementary Construction of Constant-Degree Expanders. *Combinatorics, Probability & Computing*, 17(3):319, 2008.

[7] Efe Aras, Kuan-Yun Lee, Ashwin Pananjady, and Thomas A. Courtade. A Family of Bayesian Cramér-Rao Bounds, and Consequences for Log-Concave Priors. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 2699–2703. IEEE, 2019.

[8] Suguru Arimoto. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

[9] Dominique Bakry and Michel Émery. Diffusions Hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

[10] Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 103. Springer, 2014.

[11] Keith Ball. Logarithmically Concave Functions and Sections of Convex Sets in $\mathbb{R}^n$. *Studia Math*, 88(1):69–84, 1988.

[12] Keith Ball and Van Hoang Nguyen. Entropy Jumps for Isotropic Log-Concave Random Vectors and Spectral Gap. *arXiv preprint arXiv:1206.5098*, 2012.

[13] Stefan Banach. Sur les Opérations dans les Ensembles Abstraits et Leur Application aux Équations Intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922. URL `http://eudml.org/doc/213289`.

[14] Leighton Pate Barnes and Ayfer Ozgur. Fisher Information and Mutual Information Constraints. *arXiv preprint arXiv:2102.05802*, 2021.

[15] Leighton Pate Barnes, Yanjun Han, and Ayfer Ozgur. Lower Bounds for Learning Distributions under Communication Constraints via Fisher Information. *Journal of Machine Learning Research*, 21(236):1–30, 2020.

[16] Eduard N. Belitser and Boris Y. Levit. Asymptotically Minimax Nonparametric Regression in L_2. *Statistics: A Journal of Theoretical and Applied Statistics*, 28(2): 105–122, 1996.

[17] Toby Berger. Rate-Distortion Theory. *Wiley Encyclopedia of Telecommunications*, 2003.

[18] Jacob Binia, Moshe Zakai, and Jacob Ziv. On the ϵ-Entropy and the Rate-Distortion Function of Certain Non-Gaussian Processes. *IEEE Transactions on Information Theory*, 20(4):517–524, 1974.

[19] Lucien Birgé and Pascal Massart. Gaussian Model Selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[20] Lucien Birgé and Pascal Massart. Minimal Penalties for Gaussian Model Selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

[21] Richard Blahut. Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

[22] Sergey Bobkov and Mokshay Madiman. Entropy and the Hyperplane Conjecture in Convex Geometry. In *2010 IEEE International Symposium on Information Theory*, pages 1438–1442. IEEE, 2010.

[23] Sergey Bobkov and Mokshay Madiman. The Entropy Per Coordinate of A Random Vector is Highly Constrained Under Convexity Conditions. *IEEE Transactions on Information Theory*, 57(8):4940–4954, 2011.

[24] Sergey Bobkov and Mokshay Madiman. An Equipartition Property for High-Dimensional Log-Concave Distributions. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 482–488. IEEE, 2012.

[25] Sergey G. Bobkov and Fedor L. Nazarov. On Convex Bodies and Log-Concave Probability Measures with Unconditional Basis. In *Geometric Aspects of Functional Analysis*, pages 53–69. Springer, 2003.

[26] Christer Borell. Convex Measures on Locally Convex Spaces. *Arkiv för matematik*, 12 (1-2):239–252, 1974.

[27] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

[28] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[29] Herm Jan Brascamp and Elliott H. Lieb. On Extensions of the Brunn-Minkowski and Prékopa-Leindler Theorems, Including Inequalities for Log Concave Functions, and with an Application to the Diffusion Equation. In *Inequalities*, pages 441–464. Springer, 2002.

[30] Silouanos Brazitikos, Apostolos Giannopoulos, Petros Valettas, and Beatrice-Helen Vritsiou. *Geometry of Isotropic Convex Bodies*, volume 196. American Mathematical Soc., 2014.

[31] Florentina Bunea, Alexandre B. Tsybakov, and Marten H. Wegkamp. Aggregation for Gaussian Regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

[32] T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal Rates of Convergence for Covariance Matrix Estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[33] T. Tony Cai, Xiaodong Li, and Zongming Ma. Optimal Rates of Convergence for Noisy Sparse Phase Retrieval via Thresholded Wirtinger Flow. *The Annals of Statistics*, 44 (5):2221–2251, 2016.

[34] Emmanuel J. Candes and Yaniv Plan. Matrix Completion with Noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[35] Emmanuel J. Candes and Yaniv Plan. Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[36] Emmanuel J. Candes, Yonina C. Eldar, Thomas Strohmer, and Vladislav Voroninski. Phase Retrieval via Matrix Completion. *SIAM review*, 57(2):225–251, 2015.

[37] Emmanuel J. Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase Retrieval via Wirtinger Flow: Theory and Algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[38] Eva Cantoni and Elvezio Ronchetti. Robust Inference for Generalized Linear Models. *Journal of the American Statistical Association*, 96(455):1022–1030, 2001.

[39] Eric A Carlen. Superadditivity of Fisher's Information and Logarithmic Sobolev Inequalities. *Journal of Functional Analysis*, 101(1):194–211, 1991.

[40] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[41] Hong-Bin Chen, Sinho Chewi, and Jonathan Niles-Weed. Dimension-Free Log-Sobolev Inequalities for Mixture Distributions. *Journal of Functional Analysis*, 281(11):109236, 2021.

[42] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes Risk Lower Bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.

[43] Yuansi Chen. An Almost Constant Lower Bound of the Isoperimetric Coefficient in the KLS Conjecture. *arXiv preprint arXiv:2011.13661*, 2020.

[44] Mung Chiang and Stephen Boyd. Geometric Programming Duals of Channel Capacity and Rate Distortion. *IEEE Transactions on Information Theory*, 50(2):245–258, 2004.

[45] B. S. Clarke and A. R. Barron. Jeffreys' Prior is Asymptotically Least Favorable Under Entropy Risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[46] Thomas A. Courtade and Max Fathi. Stability of the Bakry-Émery Theorem on $\mathbb{R}^n$. *Journal of Functional Analysis*, 279(2):108523, 2020.

[47] Thomas A. Courtade and Tsachy Weissman. Multiterminal Source Coding under Logarithmic Loss. *IEEE Transactions on Information Theory*, 60(1):740–761, 2013.

[48] Thomas A. Courtade and Richard D. Wesel. Multiterminal Source Coding with an Entropy-Based Distortion Measure. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2040–2044. IEEE, 2011.

[49] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[50] Harald Cramér. Mathematical Methods of Statistics. Technical report, 1946.

[51] Harald Cramér. *Mathematical Methods of Statistics*, volume 43. Princeton University Press, 1999.

[52] G. Darmois. Sur les Limites de la Dispersion de Certaines Estimations. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 13(1/4):9–15, 1945. ISSN 03731138. URL `http://www.jstor.org/stable/1400974`.

[53] Philip J. Davis. Leonhard Euler's Integral: A Historical Profile of the Gamma Function: In Memoriam: Milton Abramowitz. *The American Mathematical Monthly*, 66(10):849–869, 1959.

[54] Lee D. Davisson. Universal Noiseless Coding. *IEEE Transactions on Information Theory*, 19(6):783–795, 1973.

[55] Annette J. Dobson and Adrian G. Barnett. *An Introduction to Generalized Linear Models*. CRC press, 2018.

[56] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax Optimal Procedures for Locally Private Estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[57] Jacques Dutka. The Early History of the Factorial Function. *Archive for History of Exact Sciences*, pages 225–249, 1991.

[58] S. Yu Efroimovich. Information Contained in a Sequence of Observations. *Robl. Peredachi Inf.*, 15:3:24–39, 1979.

[59] James R. Fienup. Phase Retrieval Algorithms: A Comparison. *Applied optics*, 21(15): 2758–2769, 1982.

[60] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

[61] Maurice Fréchet. Sur l'Extension de Certaines Evaluations Statistiques au Cas de Petits Echantillons. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 11(3/4):182–205, 1943. ISSN 03731138. URL `http://www.jstor.org/stable/1401114`.

[62] A. Gerrish and P. Schultheiss. Information Rates of Non-Gaussian Processes. *IEEE Transactions on Information Theory*, 10(4):265–271, 1964.

[63] Richard D. Gill and Boris Y. Levit. Applications of the van Trees Inequality: a Bayesian Cramér-Rao Bound. *Bernoulli*, 1(1-2):59–79, 1995.

[64] Herbert Gish and John Pierce. Asymptotically Efficient Quantizing. *IEEE Transactions on Information Theory*, 14(5):676–683, 1968.

[65] Robert M. Gray, Tamás Linder, and Jia Li. A Lagrangian Formulation of Zador's Entropy-Constrained Quantization Theorem. *IEEE Transactions on Information Theory*, 48(3):695–707, 2002.

[66] Leonard Gross. Logarithmic Sobolev Inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.

[67] Leonard Gross and Oscar Rothaus. Herbst Inequalities for Supercontractive Semigroups. *Journal of Mathematics of Kyoto University*, 38(2):295–318, 1998.

[68] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-Optimal Inference from Partial Rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.

[69] Jaroslav Hájek. Local Asymptotic Minimax and Admissibility in Estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 175–194, 1972.

[70] Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin J. Wainwright. Approximate Ranking from Pairwise Comparisons. *arXiv preprint arXiv:1801.01253*, 2018.

[71] Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active Ranking from Pairwise Comparisons and When Parametric Assumptions Do Not Help. *The Annals of Statistics*, 47(6):3099–3126, 2019.

[72] Sandra Heldsinger and Stephen Humphry. Using the Method of Pairwise Comparison to Obtain Reliable Teacher Assessments. *The Australian Educational Researcher*, 37 (2):1–19, 2010.

[73] Julien Hendrickx, Alex Olshevsky, and Venkatesh Saligrama. Minimax Rate for Learning From Pairwise Comparisons in the BTL Model. *International Conference on Machine Learning*, 2020.

[74] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, pages 569–576, 2007.

[75] Richard Holley and Daniel Stroock. Simulated Annealing via Sobolev Inequalities. *Communications in Mathematical Physics*, 115(4):553–569, 1988.

[76] Jiantao Jiao, Thomas A. Courtade, Kartik Venkat, and Tsachy Weissman. Justification of Logarithmic Loss Via the Benefit of Side Information. *IEEE Transactions on Information Theory*, 61(10):5357–5365, 2015.

[77] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix Completion from a Few Entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[78] Bo'az Klartag and Joseph Lehec. Bourgain's Slicing Problem and KLS Isoperimetry up to Polylog. *arXiv preprint arXiv:2203.15551*, 2022.

[79] Tobias Koch. The Shannon Lower Bound Is Asymptotically Tight. *IEEE Transactions on Information Theory*, 62(11):6155–6161, 2016.

[80] Victoria Kostina. When is Shannon's Lower Bound Tight at Finite Blocklength? In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 982–989. IEEE, 2016.

[81] Victoria Kostina. Data Compression with Low Distortion and Finite Blocklength. *IEEE Transactions on Information Theory*, 63(7):4268–4285, 2017.

[82] Paul F.M. Krabbe. Thurstone Scaling as a Measurement Method to Quantify Subjective Health Outcomes. *Medical Care*, 46(4):357–365, 2008.

[83] Samriddha Lahiry and Michael Nussbaum. Minimax Nonparametric Estimation of Pure Quantum States. *The Annals of Statistics*, 50(1):430–459, 2022.

[84] Rafał Latała and Jakub Onufry Wojtaszczyk. On the Infimum Convolution Inequality. *arXiv preprint arXiv:0801.4036*, 2008.

[85] Michel Ledoux. *The Concentration of Measure Phenomenon*. Number 89. American Mathematical Soc., 2001.

[86] Kuan-Yun Lee and Thomas Courtade. Minimax Bounds for Generalized Linear Models. *Advances in Neural Information Processing Systems*, 33, 2020.

[87] Kuan-Yun Lee and Thomas A. Courtade. Linear Models are Most Favorable among Generalized Linear Models. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 1213–1218. IEEE, 2020.

[88] Kuan-Yun Lee and Thomas A. Courtade. Minimax Bounds for Generalized Pairwise Comparisons. In *2021 International Conference on Machine Learning (ICML) Workshop on Information-Theoretic Methods for Rigorous, Responsible, and Reliable Machine Learning*, 2021.

[89] Tamas Linder and Ram Zamir. On the Asymptotic Tightness of the Shannon Lower Bound. *IEEE Transactions on Information Theory*, 40(6):2026–2031, 1994.

[90] Peter John Loewen, Daniel Rubenson, and Arthur Spirling. Testing The Power Of Arguments In Referendums: A Bradley–Terry Approach. *Electoral Studies*, 31(1): 212–221, 2012.

[91] Po-Ling Loh and Martin J. Wainwright. Regularized M-Estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

[92] Adrian Tovar Lopez and Varun Jog. Generalization Error Bounds using Wasserstein Distances. In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2018.

[93] R. Duncan Luce. *Individual Choice Behavior*. John Wiley, 1959.

[94] Cheng Mao, Ashwin Pananjady, and Martin J. Wainwright. Towards Optimal Estimation of Bivariate Isotonic Matrices with Unknown Permutations. *arXiv preprint arXiv:1806.09544*, 2018.

[95] Peter McCullagh. *Generalized Linear Models*. Routledge, 2019.

[96] Colin McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998.

[97] Ron Meir and Tong Zhang. Generalization Error Bounds for Bayesian mixture Algorithms. *Journal of Machine Learning Research*, 4(Oct):839–860, 2003.

[98] Vitali D. Milman. Inégalité de Brunn-Minkowski Inverse et Applications à la Théorie Locale des Espaces Normés. *C. R. Acad. Sci. Paris Sér. I Math., 302(1):25–28*, 1986.

[99] Hans-Georg Müller and Ulrich Stadtmüller. Generalized Functional Linear Models. *the Annals of Statistics*, 33(2):774–805, 2005.

[100] Susan A. Murphy. A Generalization Error for Q-Learning. 2005.

[101] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative Ranking from Pair-Wise Comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012.

[102] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank Centrality: Ranking From Pairwise Comparisons. *Operations Research*, 65(1):266–287, 2017.

[103] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistical Science*, 27(4):538–557, 2012.

[104] John Ashworth Nelder and Robert WM Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

[105] Robert M. Nosofsky. Luce's Choice Model and Thurstone's Categorical Judgment Model Compared: Kornbrot's Data Revisited. *Perception & Psychophysics*, 37(1): 89–91, 1985.

[106] Ashwin Pananjady, Cheng Mao, Vidya Muthukumar, Martin J. Wainwright, and Thomas A. Courtade. Worst-Case Versus Average-Case Design for Estimation From Partial Pairwise Comparisons. *Annals of Statistics*, 48(2):1072–1097, 2020.

[107] Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization Error Bounds for Noisy, Iterative Algorithms. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 546–550. IEEE, 2018.

[108] Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94. Cambridge University Press, 1999.

[109] Robin L. Plackett. The Analysis of Permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[110] Maxim Raginsky and Igal Sason. Concentration of Measure Inequalities in Information Theory, Communications, and Coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

[111] C. Radhakrishna Rao. Information and Accuracy Attainable in the Estimation of Statistical Parameters. *Bull. Calcutta Math. Soc.*, 37:81–91, 1945.

[112] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-Balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[113] Benjamin Recht. A Simpler Approach to Matrix Completion. *Journal of Machine Learning Research*, 12(12), 2011.

[114] Philippe Rigollet. Generalization Error Bounds in Semi-Supervised Classification Under the Cluster Assumption. *Journal of Machine Learning Research*, 8(7), 2007.

[115] Philippe Rigollet. Kullback–Leibler Aggregation and Misspecified Generalized Linear Models. *The Annals of Statistics*, 40(2):639–665, 2012.

[116] O. S. Rothaus. Diffusion on Compact Riemannian Manifolds and Logarithmic Sobolev Inequalities. *Journal of Functional Analysis*, 42(1):102–109, 1981.

[117] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic Programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

[118] Adrien Saumard and Jon A. Wellner. Log-Concavity and Strong Log-Concavity: a review. *Statistics Surveys*, 8:45, 2014.

[119] Louis L. Scharf and Cédric Demeure. *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*. Prentice Hall, 1991.

[120] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence. *The Journal of Machine Learning Research*, 17(1):2049–2095, 2016.

[121] Nihar B. Shah, Sivaraman Balakrishnan, and Martin J. Wainwright. Feeling the Bern: Adaptive Estimators for Bernoulli Probabilities of Pairwise Comparisons. *IEEE Transactions on Information Theory*, 65(8):4854–4874, 2019.

[122] Claude E. Shannon. Coding Theorems for a Discrete Source with a Fidelity Criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

[123] Claude Elwood Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[124] Jack Sherman and Winifred J Morrison. Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

[125] Aart J. Stam. Some Inequalities Satisfied by the Quantities of Information of Fisher and Shannon. *Information and Control*, 2(2):101–112, 1959.

[126] John A. Swets. The Relative Operating Characteristic in Psychology: A Technique For Isolating Effects of Response Bias Finds Wide Use in The Study of Perception and Cognition. *Science*, 182(4116):990–1000, 1973.

[127] Harry H. Tan and Kung Yao. Evaluation of Rate-Distortion Functions for a Class of Independent Identically Distributed Sources Under an Absolute-Magnitude Criterion. *IEEE Transactions on Information Theory*, 21(1):59–64, 1975.

[128] Louis L. Thurstone. A Law of Comparative Judgment. *Psychological Review*, 34(4): 273, 1927.

[129] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

[130] Sara A. Van de Geer. High-Dimensional Generalized Linear Models and the Lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

[131] Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

[132] Arnoud CM van Rooij and Frits H Ruymgaart. Asymptotic Minimax Rates for Abstract Linear Estimators. *Journal of Statistical Planning and Inference*, 53(3):389–402, 1996.

[133] Harry L. van Trees. *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, 1968.

[134] Nicolas Verzelen. Minimax Risks for Sparse Regressions: Ultra-High Dimensional Phenomenons. *Electronic Journal of Statistics*, 6:38–90, 2012.

[135] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[136] Xianfu Wang. Volumes of Generalized Unit Balls. *Mathematics Magazine*, 78(5): 390–395, 2005.

[137] Terry A. Welch. A Technique for High-Performance Data Compression. *Computer*, 17 (06):8–19, 1984.

[138] Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The Context-Tree Weighting Method: Basic Properties. *IEEE Transactions on Information Theory*, 41 (3):653–664, 1995.

[139] Max A. Woodbury. *Inverting Modified Matrices*. Statistical Research Group, 1950.

[140] Y. Wu. Lecture Notes For Information-Theoretic Methods For High-Dimensional Statistics, July 2017.

[141] Yihong Wu. Lecture Notes for Information-Theoretic Methods for High-Dimensional Statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, 2017.

[142] Yoshio Yamada, Saburo Tazaki, and R. Gray. Asymptotic Performance of Block Quantizers with Difference Distortion Measures. *IEEE Transactions on Information Theory*, 26(1):6–14, 1980.

[143] Paul Laszlo Zador. *Development and Evaluation of Procedures for Quantizing Multivariate Distributions*. Stanford University, 1964.

[144] Jacob Ziv and Abraham Lempel. Compression of Individual Sequences via Variable-Rate Coding. *IEEE Transactions on Information Theory*, 24(5):530–536, 1978.