

A Family of Bayesian Cramér-Rao Bounds, and Consequences for Log-Concave Priors

Efe Aras, Kuan-Yun Lee, Ashwin Pananjady and Thomas A. Courtade

Department of Electrical Engineering and Computer Sciences

University of California, Berkeley

Email: {efearas96, timkylee, ashwinpm, courtade}@berkeley.edu

Abstract—Under minimal regularity assumptions, we establish a family of information-theoretic Bayesian Cramér-Rao bounds, indexed by probability measures that satisfy a logarithmic Sobolev inequality. This family includes as a special case the known Bayesian Cramér-Rao bound (or van Trees inequality), and its less widely known entropic improvement due to Efroimovich. For the setting of a log-concave prior, we obtain a Bayesian Cramér-Rao bound which holds for any (possibly biased) estimator and, unlike the van Trees inequality, does not depend on the Fisher information of the prior.

I. INTRODUCTION

Throughout, we let $\mathcal{P}(\mathbb{R}^n)$ denote the set of probability measures on $\mathbb{R}^n$, equipped with the Borel σ -algebra. For $\mu \in \mathcal{P}(\mathbb{R}^n)$, we abuse notation slightly and define

$$\text{Var}(\mu) := \inf_{c \in \mathbb{R}^n} \int |x - c|^2 d\mu(x),$$

where $|\cdot|$ denotes Euclidean length on $\mathbb{R}^n$. Thus, $\text{Var}(\mu)$ is the usual variance in dimension $n = 1$; it is the trace of the covariance matrix for arbitrary dimension n . A probability measure $\mu \in \mathcal{P}(\mathbb{R}^n)$ is said to be log-concave if $d\mu(x) = e^{-V(x)} dx$ for convex V . All logarithms are taken with respect to the natural base.

Our results are best stated within the general framework of parametric statistics. To this end, we let $(\mathcal{X}, \mathcal{F}, P_\theta; \theta \in \mathbb{R}^n)$ be a dominated family of probability measures on a measurable space $(\mathcal{X}, \mathcal{F})$; with dominating σ -finite measure λ . To each P_θ , we associate a density $f(\cdot; \theta)$ (w.r.t. λ) according to

$$dP_\theta(x) = f(x; \theta) d\lambda(x).$$

For sufficiently smooth densities, the Fisher information of the parametric family (P_θ) evaluated at θ is defined as

$$\mathcal{I}(\theta) := \int_{\mathcal{X}} \frac{|\nabla_\theta f(x; \theta)|^2}{f(x; \theta)} d\lambda(x),$$

where ∇_θ denotes gradient with respect to θ . Note that $\mathcal{I}$ is distinct from the information theorist's Fisher information $\mathcal{J}$, defined as

$$\mathcal{J}(\mu) := \int_{\mathbb{R}^n} \frac{|\nabla \varrho(\theta)|^2}{\varrho(\theta)} d\theta$$

for a probability measure $\mu \in \mathcal{P}(\mathbb{R}^n)$ having density ϱ with respect to Lebesgue measure. In the special case where θ is a location parameter, the two quantities coincide.

For a real-valued parameter $\theta \in \mathbb{R}$ and an observation $X \sim P_\theta$ the basic question of parametric statistics is how well can

one estimate θ from X . Here, the Cramér-Rao bound is of central importance in proving lower bounds on L^2 estimation error, stating that

$$\text{Var}(\hat{\theta}(X)) = \mathbb{E}(\theta - \hat{\theta}(X))^2 \geq \frac{1}{\mathcal{I}(\theta)} \quad (1)$$

for any unbiased estimator $\hat{\theta}$. The assumption of unbiasedness is quite restrictive, especially since unbiased estimators may not always exist, or may be less attractive than biased estimators for any one of a variety of reasons (computability, performance, etc.). Under the assumption that the parameter θ is distributed according to some prior $\pi \in \mathcal{P}(\mathbb{R})$, the so-called Bayesian Cramér-Rao bound [1], [2] (also known as the van Trees inequality) states, under mild regularity assumptions, that

$$\mathbb{E}(\theta - \hat{\theta}(X))^2 \geq \frac{1}{\mathbb{E}_\pi \mathcal{I}(\theta) + \mathcal{J}(\pi)}, \quad (2)$$

where the expectation is over $\theta \sim \pi$ and, conditioned on θ , $X \sim P_\theta$. As noted by Tsybakov [3, Section 2.7.3], this inequality is quite powerful since it does not impose any restriction on unbiasedness, is relatively simple to apply, and often leads to sharp results (including sharp constants). Tsybakov states that the primary disadvantage of (2) is that it applies only to L^2 loss. Although it does not appear to be widely known, this is actually not true. Indeed, Efroimovich proved in [4] that

$$\frac{1}{2\pi e} e^{2h(\theta|X)} \geq \frac{1}{\mathbb{E}_\pi \mathcal{I}(\theta) + \mathcal{J}(\pi)}, \quad (3)$$

which is stronger than (2) by the max-entropy property of Gaussians. Efroimovich's inequality can be rearranged to give an upper bound on the mutual information

$$\begin{aligned} I(\pi; P_\theta) &\equiv I(\theta; X) \\ &:= \iint f(x; \theta) \log \frac{f(x; \theta)}{\int_{\mathbb{R}^n} f(x; \theta') d\pi(\theta')} d\lambda(x) d\pi(\theta). \end{aligned}$$

Such a general upper bound on $I(\pi; P_\theta)$ can be useful in settings beyond those where (2) applies. For example, it can be used to give one direction of the key estimate in Clarke and Barron's work showing that Jeffrey's prior is least favorable [5]. It can also be applied to characterize Bayes risk measured under losses other than L^2 when coupled with a lower bound on mutual information (see, e.g., [6]). We remark

that several systematic techniques exist for lower bounding the mutual information $I(\pi; P_\theta)$ in terms of Bayes risk (e.g., Fano's method, or the Shannon lower bound for the rate distortion function), so finding a good upper bound is often the challenge. A typical heuristic is to bound $I(\pi; P_\theta)$ from above by the capacity of the channel $\theta \mapsto X \sim P_\theta$, but this method has the disadvantages that (i) it discards information about the prior π ; and (ii) capacity expressions are only explicitly known for very special parametric families (P_θ) (e.g., Gaussian channels). Efroimovich's inequality overcomes both of these obstacles, but has the undesirable property of being degenerate when $\mathcal{J}(\pi) = +\infty$. This can be a serious disadvantage in applications since many natural priors have infinite Fisher information, for example uniform measures on convex bodies¹.

Contributions

We make two main contributions, which we describe in rough terms here. Precise statements are given in Section II. First, we establish a family of Bayesian Cramér-Rao-type bounds indexed by probability measures that satisfy a logarithmic Sobolev inequality on $\mathbb{R}^n$. This generalizes Efroimovich's inequality (3), which corresponds to the special case where the reference measure is taken to be Gaussian. Second, we specialize the first result to obtain an explicit Bayesian Cramér-Rao-type bound under the assumption of a log-concave prior π . In dimension one, the result implies

$$e^{2h(\theta|X)} \geq \frac{4}{e^2 \mathbb{E}_\pi \mathcal{I}(\theta)} \geq \frac{1}{2 \mathbb{E}_\pi \mathcal{I}(\theta)}, \quad (4)$$

provided $\text{Var}(\pi) \geq 1/\mathbb{E}_\pi \mathcal{I}(\theta)$; a correction is needed if this condition is not met² (see Theorem 2 for a precise statement). In particular,

$$\mathbb{E}(\theta - \hat{\theta}(X))^2 \geq \text{Var}(\theta - \hat{\theta}(X)) \geq C/\mathbb{E}_\pi \mathcal{I}(\theta)$$

holds under our assumptions for a universal constant C , regardless of whether $\hat{\theta}$ is biased. This should be compared to the classical Cramér-Rao bound: morally speaking, (1) continues to hold (up to a modest constant factor) for any estimator $\hat{\theta}$, provided we are working with a log-concave prior and $\text{Var}(\pi) \geq 1/\mathbb{E}_\pi \mathcal{I}(\theta)$. Note that the crucial (and somewhat surprising) advantage relative to (3) is that the Fisher information $\mathcal{J}(\pi)$ does not appear.

Organization

The sequel is organized as follows: main results, along with assumptions and brief discussion are provided in Section II. The proofs of all results can be found in Section III.

¹Mollification may be a useful heuristic to compensate for infinite $\mathcal{J}(\pi)$ in low dimensions, but this problem becomes more fundamental in high dimensions where mollification picks up dimensional dependence, and alters the boundary of a set where the measure concentrates.

²It is easy to see why a condition like this is needed: if there were no such assumption, then we could let π approximate a point mass, effectively showing that the Cramér-Rao bound holds – up to an absolute constant – for any estimator. This clearly can not be true (consider the trivial constant biased estimator $\hat{\theta}$ not equal to θ).

II. MAIN RESULTS

A. Assumptions

As is typical of Cramér-Rao-type bounds, our main results require us to assume some mild regularity. In particular, for a given measure $\mu \in \mathcal{P}(\mathbb{R}^n)$, we will refer to the following standard condition on the densities associated to (P_θ) :

$$\int_{\mathcal{X}} \nabla_\theta f(x; \theta) d\lambda(x) = 0, \quad \mu - a.e. \theta, \quad (5)$$

where ∇_θ denotes the gradient with respect to θ . We remark that this holds whenever the orders of differentiation with respect to θ and integration with respect to x can be exchanged (Leibniz rule).

B. Statement of Results

Our first main result establishes a family of Cramér-Rao-type bounds on the mutual information $I(\pi; P_\theta)$ in terms of logarithmic Sobolev inequalities on $\mathbb{R}^n$. To this end, we recall the standard definitions of relative entropy and relative Fisher information (the parlance in which logarithmic Sobolev inequalities are framed). Consider $\mu, \nu \in \mathcal{P}(\mathbb{R}^n)$, with $\nu \ll \mu$ and $d\nu = h d\mu$. The entropy of ν , relative to μ , is defined as

$$D_\mu(\nu) \equiv D_\mu(h) := \int_{\mathbb{R}^n} h \log h d\mu.$$

If the density h is weakly differentiable, the Fisher information of ν , relative to μ , is defined according to

$$I_\mu(\nu) \equiv I_\mu(h) := \int_{\mathbb{R}^n} \frac{|\nabla h|^2}{h} d\mu.$$

If h is not weakly differentiable, we adopt the convention that $I_\mu(h) = +\infty$ so that our expressions make sense even in the general case.

A probability measure μ is said to satisfy a logarithmic Sobolev inequality with constant $C > 0$ (or, $\text{LSI}(C)$ for short) if, for all probability measures $\nu \ll \mu$,

$$D_\mu(\nu) \leq \frac{C}{2} I_\mu(\nu).$$

The standard Gaussian measure $d\gamma(x) := (2\pi)^{-n/2} e^{-|x|^2/2} dx$ on $\mathbb{R}^n$ is a prototypical example of a measure that satisfies an LSI, and does so with constant $C = 1$. More generally, if $d\mu(x) = e^{-V(x)} dx$ with $\text{Hess}(V) \geq K \cdot I_n$ for $K > 0$ and I_n the $n \times n$ identity matrix, then μ satisfies $\text{LSI}(1/K)$; this result is known as the Bakry-Émery theorem [7], and we shall need it later in the proof of Theorem 2.

With these definitions in hand, our first result is the following:

Theorem 1. *Let $\mu \in \mathcal{P}(\mathbb{R}^n)$ satisfy $\text{LSI}(C)$ and assume the regularity condition (5) holds. For any probability measure $\pi \ll \mu$ on $\mathbb{R}^n$,*

$$I(\pi; P_\theta) + D_\mu(\pi) \leq \frac{C}{2} \left(I_\mu(\pi) + \int_{\mathbb{R}^n} \mathcal{I}(\theta) d\pi(\theta) \right). \quad (6)$$

Inequality (6) improves the LSI for μ . Indeed, taking P_θ independent of θ renders $I(\pi; P_\theta) = \mathcal{I}(\theta) = 0$, so that the LSI for μ is recovered. However, the proof of (6) follows from a relatively simple application of the LSI for μ and some basic analysis, so the two inequalities should be viewed as being formally equivalent in this sense.

Clearly, the statement of Theorem 1 allows us the freedom to choose the measure μ so as to obtain the tightest possible bound on $I(\pi; P_\theta)$. However, a notable example is obtained when μ is taken to be the standard Gaussian measure on $\mathbb{R}^n$. In this case, upon simplification we obtain

$$1 + \log(2\pi e) \leq \frac{2}{n} h(\theta|X) + \mathcal{J}(\pi) + \int_{\mathbb{R}^n} \mathcal{I}(\theta) d\pi(\theta). \quad (7)$$

Of note, (7) is not invariant to rescalings of the parameter θ . So, just as one passes from Lieb's inequality to the entropy power inequality, we may optimize over all such scalings to obtain the following multidimensional version of (3):

$$\frac{1}{2\pi e} \exp\left(\frac{2}{n} h(\theta|X)\right) \geq \frac{n}{\mathcal{J}(\pi) + \int_{\mathbb{R}^n} \mathcal{I}(\theta) d\pi(\theta)}.$$

Remark 1. *Efroimovich's work [4] contains a slightly stronger multidimensional form, stated in terms of determinants of Fisher information matrices. As defined, our Fisher information quantities $\mathcal{I}$ and $\mathcal{J}$ correspond to traces of the same matrices, leading to a weaker inequality by the arithmetic-geometric mean inequality. Nevertheless, the two inequalities should really be regarded as essentially equivalent, as they are both direct consequences of the one-dimensional inequality (where the two results coincide). See [4, Proof of Theorem 5] for details. It is unclear whether a similar claim holds for non-Gaussian μ in (6).*

We remark that (3) was discovered by Efroimovich in 1979, but does not appear to be widely known (we could not find a statement of the result outside the Russian literature). At the time of Efroimovich's initial discovery of (3), the study of logarithmic Sobolev inequalities was just getting started, being largely initiated by Gross's work on the Gaussian case in 1975 [8]. In particular, the derivation of (3) (and, less generally, the van Trees inequality) from the Gaussian logarithmic Sobolev inequality does not appear to have been observed previously. So, from a conceptual standpoint, one contribution of Theorem 1 is that it shows how Efroimovich's result (and the weaker van Trees inequality) emerges as one particular instance in the broader context of LSIs which, to our knowledge, have not found direct use in parametric statistics beyond their implications for measure concentration (see, e.g., [9]).

A nontrivial consequence of Theorem 1 is a general Cramér-Rao-type bound on $I(\pi; P_\theta)$, assuming only that π is log-concave. Specifically, our second main result is the following:

Theorem 2. *Assume the parametric family (P_θ) satisfies (5) for μ equal to Lebesgue measure. Let $d\pi(x) = e^{-V(x)} dx$ satisfy $\text{Hess}(V) \geq K \cdot I_n$ for some scalar $K \geq 0$, where I_n*

is the $n \times n$ identity matrix. Define $P := \frac{1}{n} \text{Var}(\pi)$, $J := \frac{1}{n} \int_{\mathbb{R}^n} \mathcal{I}(\theta) d\pi(\theta)$. It holds that

$$I(\pi; P_\theta) \leq n \cdot \phi\left(\sqrt{(KP)^2 + JP} - KP\right), \quad (8)$$

where

$$\phi(x) := \begin{cases} x & \text{if } 0 \leq x < 1 \\ 1 + \log x & \text{if } x \geq 1. \end{cases}$$

Remark 2. *The one-dimensional inequality (4) follows directly from Theorem 2 for $K = 0$, combined with the entropy lower bound for log-concave random variables $h(\theta) \geq \frac{1}{2} \log(4\text{Var}(\theta))$ due to Marsiglietti and Kostina [10]. Similar statements hold for general dimension n , albeit with a correction factor that depends on dimension (no correction is needed if the hyperplane conjecture is true; see [11]).*

The upper bound (8) should be viewed as a function of two nonnegative quantities: the products KP and JP . By the Brascamp-Lieb inequality [12], we always have $KP \leq 1$; this quantity only depends on the prior π and distills what quantitative information is known about its degree of log-concavity. In particular, if π is only known to be log-concave, then $K = 0$ gives $I(\pi; P_\theta) \leq n \cdot \phi(\sqrt{JP})$. In the other extreme case, if $KP = 1$ (e.g., if π is scaled standard Gaussian), we have the slightly improved bound $I(\pi; P_\theta) \leq n \cdot \phi(\sqrt{1 + JP} - 1)$. These bounds both essentially behave as $\frac{n}{2} \log(JP)$ for JP modestly large, so knowledge of KP (i.e., additional information about the measure π) only significantly affects the behavior of the upper bound (8) for JP small. To be precise, for JP near zero, the upper bound behaves as nJ/K when $K > 0$, and $n\sqrt{JP}$ if $K = 0$. Applications in asymptotic statistics consider a sequence of observations $X_1, \dots, X_m$, conditionally independent given θ . In this case, J grows linearly with m , so that the logarithmic behavior of the bound dominates, regardless of what is known about K .

Let us now make a brief observation on the sharpness of Theorem 2. To this end, consider the classical Gaussian sequence model $X = \theta + Z$, where $Z \sim N(0, \sigma^2 I_n)$ is independent of $\theta \sim \pi$. In this case, the typical quantity of relevance is the signal-to-noise ratio $\text{snr} := \frac{\text{Var}(\pi)}{n\sigma^2} = n^{-2} \text{Var}(\pi) \int \mathcal{I}(\theta) d\pi(\theta)$, in terms of which we have the sharp upper bound

$$I(\pi; P_\theta) \leq \frac{n}{2} \log(1 + \text{snr}) = \frac{n}{2} \log(1 + JP). \quad (9)$$

Thus, in view of the previous discussion, we clearly see that Theorem 2 provides a sharp estimate in the regime where JP is moderately large. We do not yet know whether the bound $I(\pi; P_\theta) \leq n \cdot \phi(\sqrt{JP})$ is sharp for small JP and $K = 0$, but we believe that it should be.

Finally, we remark that all results have correct dependence on dimension for product measures.

C. Remarks on Applications

Applications of Cramér-Rao-type bounds are numerous, and our results will generally apply in Bayesian settings. In

particular, we believe corollaries such as (4) may be especially useful for proving lower bounds on Bayes risk when the prior π is log-concave. Of course, since mutual information is invariant to one-to-one transformations of the arguments, Theorem 2 holds even when π is not log-concave, but one must reparameterize the problem data in terms of a log-concave proxy for π for purposes of computing the quantities J, P (a nontrivial task in general). That being said, we remark that our results are quite general in form, and therefore should not be interpreted as being restricted to applications in parametric statistics.

III. PROOFS

This section outlines the proofs of main results. Omitted details can be found in [13].

A. Proof of Theorem 1

We may assume that the RHS of equation (6) is finite; else the claim is trivially true. Let $d\pi = h d\mu$, and note that $h(\theta)f(x; \theta)$ is the joint density of (π, P_θ) with respect to $\mu \times \lambda$. Define $f(x) = \int_{\mathbb{R}^n} f(x; \theta) d\pi(\theta)$, and $h_x(\theta) = h(\theta)f(x; \theta)/f(x)$, which is well-defined $(\pi \times \lambda)$ -a.e. Now, since μ satisfies LSI(C), we have for λ -a.e. x

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

where we write ∇ in place of ∇_θ for brevity. Integrating both sides with respect to the density $f d\lambda$, we have

$$\begin{aligned} & \int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^n} h_x(\theta) \log h_x(\theta) d\mu(\theta) \right) d\lambda(x) \\ & \leq \frac{C}{2} \int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^n} \frac{|\nabla h_x(\theta)|^2}{h_x(\theta)} d\mu(\theta) \right) d\lambda(x). \end{aligned}$$

Now, observe that

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

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

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

The exchange of integrals to obtain the last line is justified by Tonelli's theorem. Therefore, by Fubini's theorem,

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

where the last equality follows by the regularity assumption. Summarizing, we have

$$\begin{aligned} & \int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^n} \frac{|\nabla h_x(\theta)|^2}{h_x(\theta)} d\mu(\theta) \right) d\lambda(x) \\ & = I_\mu(\pi) + \int_{\mathbb{R}^n} \mathcal{I}(\theta) d\pi(\theta). \end{aligned}$$

To finish, we observe that

$$\begin{aligned} & \int_{\mathcal{X}} f(x) \left(\int_{\mathbb{R}^n} h_x(\theta) \log h_x(\theta) d\mu(\theta) \right) d\lambda(x) \\ & = \int_{\mathcal{X}} \int_{\mathbb{R}^n} f(x) h_x(\theta) \log h_x(\theta) d\mu(\theta) d\lambda(x) \\ & = \int_{\mathcal{X}} \int_{\mathbb{R}^n} f(x;\theta) h(\theta) \log \frac{h_x(\theta)}{h(\theta)} d\mu(\theta) d\lambda(x) \\ & \quad + \int_{\mathcal{X}} \int_{\mathbb{R}^n} f(x;\theta) h(\theta) \log h(\theta) d\mu(\theta) d\lambda(x) \\ & = \int_{\mathcal{X}} \int_{\mathbb{R}^n} f(x;\theta) h(\theta) \log \frac{f(x;\theta)}{f(x)} d\mu(\theta) d\lambda(x) \\ & \quad + \int_{\mathbb{R}^n} h(\theta) \log h(\theta) d\mu(\theta) \\ & = I(\pi; P_\theta) + D_\mu(\pi), \end{aligned}$$

which proves the claim.

B. Proof of Theorem 2

We require the following proposition, the proof of which is the most difficult part of the argument. The ideas of the proof are independent from Theorem 2, so it is deferred to [13] due to space constraints. The key idea is to show the existence of the specified m_δ using the Banach fixed-point theorem. The proposition itself may be of independent interest.

Proposition 1. Let $\rho = e^{-V}$ be a probability density on $\mathbb{R}^n$, with V convex.

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

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

(ii) For m_δ as in part (i), and each $\delta \geq 0$

$$\begin{aligned} & -\log \left(\int_{\mathbb{R}^n} e^{-\delta|x-m_\delta|^2/2} \rho(x) dx \right) \\ & \leq \begin{cases} \frac{\delta}{2} \text{Var}(\rho) & \text{if } 0 \leq \delta < \frac{n}{\text{Var}(\rho)}, \\ \frac{n}{2} (1 + \log(\frac{\delta}{n} \text{Var}(\rho))) & \text{if } \delta \geq \frac{n}{\text{Var}(\rho)}. \end{cases} \end{aligned}$$

To begin the proof, consider the log-concave density $d\pi(x) = e^{-V(x)} dx$, where $\text{Hess}(V) \geq K \cdot \text{I}_n$. For $\delta > 0$, let μ_δ be the probability measure with density

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

where $C_\delta = \int_{\mathbb{R}^n} e^{-V(x) - \delta|x-m_\delta|^2/2} dx$ is a normalizing constant and $m_\delta \in \mathbb{R}^n$ is such that $\int_{\mathbb{R}^n} x d\mu_\delta = m_\delta$, which exists as a consequence of Proposition 1(i). Note that π has density $C_\delta e^{\delta|x-m_\delta|^2/2}$ with respect to μ_δ . Therefore, we may readily compute

$$\begin{aligned} D_{\mu_\delta}(\pi) &= \frac{\delta}{2} \int_{\mathbb{R}^n} |x - m_\delta|^2 e^{-V(x)} dx + \log C_\delta \\ &= \frac{1}{2\delta} I_{\mu_\delta}(\pi) + \log C_\delta. \end{aligned}$$

By the Bakry-Émery theorem, μ_δ satisfies $\text{LSI}(1/(K + \delta))$, so it follows from Theorem 1 that

$$\begin{aligned} & I(\pi; P_\theta) \\ & \leq -D_{\mu_\delta}(\pi) + \frac{1}{2(K + \delta)} \cdot I_{\mu_\delta}(\pi) + \frac{1}{2(K + \delta)} \int \mathcal{I}(\theta) d\pi(\theta) \\ & = -\frac{K}{2\delta(K + \delta)} \cdot I_{\mu_\delta}(\pi) + \frac{1}{2(K + \delta)} \int \mathcal{I}(\theta) d\pi(\theta) - \log C_\delta \\ & = -\frac{K\delta}{2(K + \delta)} \cdot \int |x - m_\delta|^2 e^{-V(x)} dx \\ & \quad + \frac{1}{2(K + \delta)} \int \mathcal{I}(\theta) d\pi(\theta) - \log C_\delta. \end{aligned}$$

By Proposition 1(ii) and the inequality

$$\int |x - m_\delta|^2 e^{-V(x)} dx \geq \text{Var}(\pi)$$

holding by definition of $\text{Var}(\pi)$, we have

$$\begin{aligned} I(\pi; P_\theta) & \leq -\frac{K\delta}{2(K + \delta)} \cdot nP + \frac{1}{2(K + \delta)} \cdot nJ \\ & \quad + \begin{cases} \frac{\delta}{2} \cdot nP & \text{if } 0 \leq \delta < \frac{1}{P} \\ \frac{n}{2} (1 + \log(\delta P)) & \text{if } \delta \geq \frac{1}{P}, \end{cases} \quad (10) \end{aligned}$$

where J, P are as defined in the statement of the theorem. Since the above holds for arbitrary $\delta > 0$, we now particularize by (optimally) choosing

$$\delta = \sqrt{K^2 + J/P} - K$$

if $JP < 1 + 2KP$, and otherwise choosing

$$\delta = \frac{1}{2} \left((K^2 P + J - 2K) + \sqrt{(K^2 P + J)^2 - 4K(K^2 P + J)} \right).$$

It can be verified that if $JP < 1 + 2KP$, then this choice of δ ensures $\delta < 1/P$. On the other hand, if $JP \geq 1 + 2KP$, then this choice of δ ensures $\delta \geq 1/P$. Hence, substitution into equation (10) and simplifying yields:

$$I(\pi; P_\theta) \leq n \cdot \psi(KP, JP)$$

where ψ is defined piecewise according to

$$\psi(a, b) := \sqrt{a^2 + b} - a \quad \text{if } b < 2a + 1,$$

and, otherwise if $b \geq 2a + 1$, then

$$\begin{aligned} \psi(a, b) &:= \frac{1}{2} \left(1 - a + \frac{2(a^2 + b)}{a^2 + b + \sqrt{(a^2 + b)^2 - 4a(a^2 + b)}} \right. \\ & \quad \left. + \log \left(\frac{a^2 + b + \sqrt{(a^2 + b)^2 - 4a(a^2 + b)}}{2} - a \right) \right). \end{aligned}$$

This bound is actually better than what is stated in the theorem, but is clearly a bit cumbersome. Since $KP \leq 1$, we note the simpler (yet, still essentially as good) bound holding for ψ in the range $0 \leq a \leq 1$, completing the proof

$$\psi(a, b) \leq \begin{cases} \sqrt{a^2 + b} - a & \text{if } b < 2a + 1 \\ 1 + \log(\sqrt{a^2 + b} - a) & \text{otherwise.} \end{cases}$$

ACKNOWLEDGEMENT

This research was supported in part by NSF grants CCF-1704967, CCF-0939370 and CCF-1750430. We thank an anonymous referee for their careful reading and suggestions.

REFERENCES

- [1] R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.
- [2] H. L. van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 1968.
- [3] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag New York, 2009.
- [4] S. Y. Efroimovich. Information contained in a sequence of observations (in Russian). *Problems in Information Transmission*, 15(3):24–39, 1979.
- [5] 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.
- [6] Y. Wu. Lecture notes for information-theoretic methods for high-dimensional statistics, July 2017.
- [7] D. Bakry and M. Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.
- [8] L. Gross. Logarithmic Sobolev inequalities. *American Journal of Mathematics*, 97(4):1061–1083, 1975.
- [9] M. Ledoux. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.
- [10] A. Marsiglietti and V. Kostina. A lower bound on the differential entropy of log-concave random vectors with applications. *Entropy*, 20(3):185, 2018.
- [11] S. Bobkov and M. 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.
- [12] H. J. Brascamp and E. 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. *Journal of Functional Analysis*, 22(4):366–389, 1976.
- [13] E. Aras, K.-Y. Lee, A. Pananjady, and T. A. Courtade. A family of Bayesian Cramér-Rao bounds, and consequences for log-concave priors. *arXiv preprint*, 2019.