The Differential Entropy of Mixtures: New Bounds and Applications

James Melbourne, *Member, IEEE*, Saurav Talukdar, Shreyas Bhaban, Mokshay Madiman, *Senior Member, IEEE*, and Murti V. Salapaka, *Fellow, IEEE*

***Abstract*— Mixture distributions are extensively used as a modeling tool in diverse areas from machine learning to communications engineering to physics, and obtaining bounds on the entropy of mixture distributions is of fundamental importance in many of these applications. This article provides sharp bounds on the entropy concavity deficit, which is the difference between the differential entropy of the mixture and the weighted sum of differential entropies of constituent components. Toward establishing lower and upper bounds on the concavity deficit, results that are of importance in their own right are obtained. In order to obtain nontrivial upper bounds, properties of the skew-divergence are developed and notions of "skew" f-divergences are introduced; a reverse Pinsker inequality and a bound on Jensen-Shannon divergence are obtained along the way. Complementary lower bounds are derived with special attention paid to the case that corresponds to independent summation of a continuous and a discrete random variable. Several applications of the bounds are delineated, including to mutual information of additive noise channels, thermodynamics of computation, and functional inequalities.**

***Index Terms*— Mixture distributions, differential entropy, concavity, f-divergence.**

I. Introduction

MIXTURE models are extensively employed in diverse disciplines including genetics, biology, medicine, economics, speech recognition, as the distribution of a signal at the receiver of a communication channel when the transmitter sends a random element of a codebook, or in models of clustering or classification in machine learning (see, e.g., [29], [71]). A mixture model is described by a density of the form $f = \sum_i p_i f_i(x)$, where each f_i is a probability density function and each p_i is a nonnegative weight with $\sum_i p_i = 1$. Such mixture densities have a natural probabilistic meaning as outcomes of a two stage random process with the first stage being a random draw, i, from the probability mass function p, followed by choosing a vector x according to a probability density function $f_i(\cdot)$ on a Euclidean space; equivalently, it is the density of $X + Z$, where X is a discrete random variable taking values x_i with probabilities p_i, and Z is a dependent variable such that $\mathbb{P}(Z \in A | X = x_i) = \int_A f_i(z + x_i) dz$. The differential entropy of this mixture is of significant interest.

Our original motivation for this paper came from the fundamental study of thermodynamics of computation, in which memory models are well approximated by mixture models, while the erasure of a bit of information is akin to the state described by a single unimodal density. Of fundamental importance here is the entropy of the mixture model which is used to estimate the thermodynamic change in entropy in an erasure process and for other computations [38], [75]. It is not possible, analytically, to determine the differential entropy of the mixture model $\sum p_i f_i$, even in the simplest case where each f_i is a Gaussian density, and hence one is interested in refined bounds on the same. While this was our original motivation, the results of this paper are of broader interest and we strive to give general statements, so as not to limit the applicability.

For a random vector Z taking values in $\mathbb{R}^d$ with probability density f, the *differential entropy* is defined as $h(Z) = h(f) = -\int_{\mathbb{R}^d} f(z) \log f(z) dz$, where the integral is taken with respect to Lebesgue measure. We will frequently omit the qualifier "differential" when this is obvious from context and simply call it the entropy. It is to be noted that, unlike $h(\sum p_i f_i)$, the quantity $\sum_i p_i h(f_i)$ is more readily determinable and thus the *concavity deficit* $h(\sum p_i f_i) - \sum p_i h(f_i)$ is of interest. This quantity can also be interpreted as a generalization of the Jensen-Shannon divergence [15], and its quantum analog (with density functions replaced by density matrices and Shannon entropy replaced by von Neumann entropy) is the Holevo information, which plays a key role in Holevo's theorem bounding the amount of accessible (classical) information in a quantum state [33].

It is a classical fact, going back at least to the origins of information theory,[1] that the entropy h is a concave function,

Manuscript received April 23, 2020; revised February 12, 2021; accepted June 7, 2021. Date of publication January 6, 2022; date of current version March 17, 2022. This work was supported by the National Science Foundation under Grant 1462862 (CMMI), Grant 1544721 (CNS), and Grant 1248100 (DMS). The work of Mokshay Madiman was supported by the National Science Foundation under Grant DMS-1409504. The work of Murti V. Salapaka was supported by the National Science Foundation under Grant CMMI-1462862, Grant CNS 1544721, and Grant ECCS 1809194 (Energy Efficiency in Computing Logical Operations: Fundamental Limits With and Without Feedback). *(Corresponding author: James Melbourne.)*

James Melbourne is with the Centro de Investigación en Matemáticas, A.C. (CIMAT), Guanajuato 36000, México (e-mail: james.melbourne@cimat.mx).

Saurav Talukdar is with Google, Mountain View, CA 94043 USA (e-mail: sauravtalukdar@umn.edu).

Shreyas Bhaban is with the Abbott Diagnostics Division, Santa Clara, CA 95054 USA (e-mail: bhaba001@umn.edu).

Mokshay Madiman is with the Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 USA (e-mail: madiman@udel.edu).

Murti V. Salapaka is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: murtis@umn.edu).

Communicated by T. Courtade, Associate Editor At Large.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3140661.

Digital Object Identifier 10.1109/TIT.2022.3140661

[1] Mathematically, the concavity of entropy is equivalent to the concavity of $\Psi : x \mapsto -x \log x$, seen by integrating $\Psi(\sum_i p_i f_i) \geq \sum_i p_i \Psi(f_i)$. Thus, classical may refer to the 17th century depending on ones interpretation.

which means that the concavity deficit is always non-negative:

$$h(f) - \sum_i p_i h(f_i) \geq 0, \tag{1}$$

with equality when $f_i = f_j$ for all i, j. Let X be a random variable that takes values in a countable set where $X = x_i$ with probability p_i. Intimately related to the entropy h of a mixture distribution $f = \sum p_i f_i$ and the concavity deficit are the quantities $H(p) := -\sum p_i \log p_i$, and the conditional entropies $h(Z|X)$ and $H(X|Z)$. Indeed, it is easy to show an upper bound on the concavity deficit (see, e.g., [11], [19], [85]) in the form

$$h(f) - \sum_i p_i h(f_i) \leq H(p), \tag{2}$$

with equality when the supports of f_i and f_j are disjoint when $i \neq j$. The main thrust of this article will be to provide quantitative refinements of the upper and lower bounds on the concavity deficit, improving upon the basic bounds (1) and (2).

As will be explained in more detail, both improvements can be understood as "stability" results. In general, a stability result for an inequality $A(x) \leq B(x)$, with a known equality case x' such that $A(x') = B(x')$, is a result which shows that when $B(x) - A(x)$ is small, then the distance (in some sense) between x and x' is small as well. Examples of stability results and quantitative sharpenings of information theoretic inequalities in recent research include the following: the entropy power inequality [18], [79], Talagrand's inequality [57], the Gaussian logarithmic Sobolev inequality [10], [24]–[26], and Han's inequality [5].

The main upper bound we establish is inspired by bounds in the quantum setting developed by Audenaert [3] and utilizes the total variation distance. Given two probability densities f_1 and f_2 with respect to a common measure μ, the total variation distance between them is defined as $\|f_1 - f_2\|_{TV} = \frac{1}{2}\int |f_1 - f_2| d\mu$.

We will state the following theorem in terms of the usual differential entropy, on Euclidean space with respect to the Lebesgue measure. Within the article, the statements and proofs will be given for a general Polish[2] measure space (E, γ), from which the result below can be recovered as a special case.

Theorem 1: Suppose $f = \sum_i p_i f_i$, where f_i are probability density functions on $\mathbb{R}^d$, $p_i \in [0, 1)$, $\sum_i p_i = 1$. Define the mixture complement of f_j by $\tilde{f}_j(z) = \sum_{i \neq j} \frac{p_i}{1-p_j} f_i$. Then

$$h(f) - \sum_i p_i h(f_i) \leq \mathcal{T}_f H(p)$$

where

$$\mathcal{T}_f \coloneqq \sup_i \|f_i - \tilde{f}_i\|_{TV}.$$

Note that $\tilde{f}_j$ is characterized by the equality $f = (1 - p_j)\tilde{f}_j + p_j f_j$, which motivates the term "mixture complement". With this decomposition in mind, one can take the convention that $\tilde{f}_j \coloneqq f_j$ when $p_j = 1$ to extend the theorem to $p_j \in [0, 1]$. Theorem 1 shows that as distributions cluster in total variation distance, the concavity deficit vanishes. The above result thus considerably reduces the conservativeness of the upper bound on the concavity deficit given by (2). Indeed consider the following example with ϕ a symmetric unimodal density, $p \in [0, 1]$, $f_1(x) = \phi(x - a)$ and $f_2(x) = \phi(x + a)$ for $a \geq 0$. Then by (2) $h(pf_1 + (1 - p)f_2) \leq H(p) + h(\phi)$. However noting that $\tilde{f}_1 = f_2$ and $\tilde{f}_2 = f_1$ gives,

$$\begin{aligned}\mathcal{T}_f &= \|f_1 - f_2\|_{TV} \\ &= \int_0^\infty (\phi(x - a) - \phi(x + a))\, dx \\ &= 2\int_0^a \phi(x) dx,\end{aligned}$$

so that Theorem 1 reduces to

$$h(pf_1 + (1 - p)f_2) \leq 2\left(\int_0^a \phi(x) dx\right) H(p) + h(\phi). \tag{3}$$

Note that the right side of (3) is of the form $h(\phi) + 2\phi(0)H(p)a + o(a)$. Taking $\phi(x) = \mathbb{1}_{[-1/2, 1/2]}(x)$, we see that equality in Theorem 1 holds for uniform distributions on intervals, since direct computations yields $h(f) = 2\left(\int_0^a \phi(x) dx\right) H(p)$, and $h(f_i) = h(\phi) = 0$.

Let us also point out a mutual information interpretation of the above in this special case. Observe that $2\int_0^a \phi(x) dx = 1 - \mathbb{P}(|\mathcal{Z}| > a)$ where $\mathcal{Z}$ has a symmetric unimodal density ϕ. Thus, an alternative representation of these bounds in this special case is as mutual information bounds

$$I(X; X + \mathcal{Z}) \leq H(X) - \mathbb{P}(|\mathcal{Z}| > a) H(X) \tag{4}$$

where X denotes an independent Bernoulli taking the values $\pm a$ with probability p and $1 - p$.

Another interpretation of Theorem 1, is as a generalization of the classical bounds on the Jensen-Shannon divergence by the total variation distance [43], [78], which is recovered by taking $p_1 = p_2 = \frac{1}{2}$, see Corollary 4.

As a stability result, Theorem 1 shows that a family of distributions $\{f_i\}$ near equality in (2) only when the family possesses an f_i far from $\tilde{f}_i$ in total variation distance. When $n = 2$ this forces $\|f_1 - f_2\|_{TV}$ to its maximum of 1.

The methods and technical development toward establishing Theorem 1 are of independent interest. We develop a notion of skew f-divergence for general f-divergences generalizing the skew divergence (or skew relative entropy) introduced by Lee [40], and in Theorem 4 we show that the class of f-divergences is stable under the skew operation. After proving elementary properties of the skew relative entropy in Proposition 4 and the skew chi-squared divergence in Proposition 5, we adapt arguments due to Audenaert [3] from the quantum setting to prove the two f-divergences are intertwined through a differential equality. Further, it is demonstrated that classical upper bound of the relative entropy by the chi-square divergence can be generalized to the skew setting (see Theorem 5). This is used to obtain a bound of the skew divergence by the total variation in Theorem 6. As a corollary we obtain a reverse Pinsker inequality due to Verdú [82]. With these tools in hand, Theorem 1 is proven and we demonstrate that the bound of the Jensen-Shannon divergence by total variation distance [43], [78] is an immediate special case.

[2]A topological space is Polish when it is homeomorphic to a complete metric space that possess a countable and dense subset.

In the converse direction (providing lower bounds on the concavity deficit), our main result applies to the case where all the component densities come from perturbations of a random vector W in $\mathbb{R}^d$ that has a log-concave and spherically symmetric distribution. We say a random vector W has a *log-concave* distribution when it possesses a density φ satisfying $\varphi((1-t)z+ty) \geq \varphi^{1-t}(z)\varphi^t(y)$. We say W has a *spherically symmetric* distribution when there exists $\psi : \mathbb{R} \to [0,\infty)$ such that the density $\varphi(z) = \psi(|z|)$ for every $z \in \mathbb{R}^d$, where $|z| := \sqrt{z_1^2 + z_2^2 + \ldots z_d^2}$. We employ the notation $B_\lambda = \{x \in \mathbb{R}^d : |x| \leq \lambda\}$ for the centered closed ball of radius λ in $\mathbb{R}^d$, $\mathscr{T}(t) := \mathscr{T}_W(t) := \mathbb{P}(|W| > t)$ for the tail probability of W, and $\|A\| := \sup_{|w|=1} |Aw|$ for the operator norm of a matrix $A : \mathbb{R}^d \to \mathbb{R}^d$.

Theorem 2: Suppose that X is a discrete variable taking values $\{x_i\} \subseteq \mathbb{R}^d$, that satisfy $|x_i - x_j| \geq 2\lambda$ for $i \neq j$ and W is independent, log-concave, and spherically symmetric density, then there exists a constant $C := C(\lambda, W)$

$$I(X+W;X) \geq H(X) - C\, \mathscr{T}^{\frac{1}{2}}(\lambda)$$

with $\lim_{\lambda\to\infty} C(\lambda, W) = \sqrt{d}$.

In fact an explicit bound for $C(\lambda, W)$ can be given. One can take

$$C(\lambda, W) = \sqrt{d} + \left(\int_{B_\lambda^c} \varphi(w) \log^2 \left[3 + \frac{2|w|}{\lambda} \right] dw \right)^{\frac{1}{2}} + A\, \mathscr{T}^{\frac{1}{2}}(\lambda),$$

with $A = 1 + h(W) + \log\left[\left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right) \omega_d^{-1} \right) \right]$, where ω_d denotes the volume of the d-dimensional unit ball, B_λ^c denotes the complement of $B_\lambda \in \mathbb{R}^d$.

Note, $|W|$ is a log-concave variable, and hence $\mathscr{T}(\lambda) := \mathbb{P}^{\frac{1}{2}}(|W| > \lambda)$ is sub-exponential in λ. Theorem 2 quantifies a natural heuristic, for a sufficiently strong signal X, the noise W has an insignificant effect. Up to a sub-exponentially small term, $h(f) - \sum_i p_i h(f_i) = I(X+W;X) \approx H(X)$. As a stability result this shows that within the regime of families comprised of translations of a symmetric log-concave distribution, approaching equality in (1) forces $\mathscr{T}(\lambda)$ large.[3] By the sub-exponential decay of $\mathscr{T}(\lambda)$ for log-concave distributions, λ cannot be too large, and hence at least for one pair of x_i, x_j we must have $|x_i - x_j|$ small, so that f_i and f_j must be close to identical.

The following result generalizes Theorem 2. Building on the same intuition, but allowing for dependent noise.

Theorem 3: Suppose that (Z, X) is an $\mathbb{R}^d \times \mathbb{N}$ valued random variable and $\tau \geq 1$ are such that for each $i \in \mathbb{N}$, $Z|X = i$ has distribution given by $T_i(W)$ where W has density φ, spherically symmetric and log-concave and T_i is a $\sqrt{\tau}$ bi-Lipschitz function. For $i, j \in \mathbb{N}$, take $T_{ij} := T_i^{-1} \circ T_j$ and further assume there exists $\lambda > 0$ such that for any $k \neq i$ $T_{ij}(B_\lambda) \cap T_{kj}(B_\lambda) = \emptyset$. Then

$$I(Z;X) = h(Z) - h(Z|X) \geq H(X) - \tilde{C}(W), \qquad (5)$$

[3]Greater than $\left(\frac{H(X)-\varepsilon}{C}\right)^2$, for $h(f) - \sum_i h(f_i) \leq \varepsilon$.

where $\tilde{C}$ is the following function,

$$\tilde{C}(W) = \mathscr{T}(\lambda)(1 + h(W)) + \mathscr{T}^{\frac{1}{2}}(\lambda)(\sqrt{d} + K(\varphi)) \qquad (6)$$

with

$$K(\varphi) := \log\left[\tau^d \left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right) \omega_d^{-1} \right) \right] \mathbb{P}^{\frac{1}{2}}(|W| > \lambda) + d \left(\int_{B_\lambda^c} \varphi(w) \log^2 \left[1 + 2\tau + \frac{2\tau^2|w|}{\lambda} \right] dw \right)^{\frac{1}{2}}.$$

To obtain Theorem 2 one can take $T_i w = w + x_i$, so that $T_{ij} w = w - x_i + x_j$ are 1 bi-Lipschitz, and observe that the hypothesis $|x_i - x_k| > 2\lambda$ for $k \neq i$ implies that $T_{ij}(B_\lambda) \cap T_{kj}(B_\lambda) = \emptyset$ for $k \neq i$. For a noise dependent example, consider W a standard Gaussian and $T_i(w) = \sigma_i w + \mu_i$ for $\mu_i \in \mathbb{R}^d$ and $\sigma_i \in (0,\infty)$, corresponds to an additive signal dependent Gaussian noise model. Here, the T_i are all $\sqrt{\tau}$ bi-Lipschitz with $\tau = \max_i \sigma_i^2$, and one may take $\lambda = \inf_{j, i\neq k} \frac{|t_{ik}(\mu_j - \mu_i) + (1-t_{ik})(\mu_k - \mu_j)|}{\sigma_j}$ where $t_{ik} = \frac{\sigma_k}{\sigma_i + \sigma_k}$, is determined by the variances.[4]

We note the quantity $H(X|Z)$ connotes the uncertainty in the discrete variable X conditioned on the continuous variable Z; such a quantity needs to be defined/determined from the knowledge of probabilities, p_i, that the discrete variable $X = x_i$ and the description of the conditional probability density function $p(z|x_i) = f_i(z)$. These notions are made precise in Section II. Here, it is also established that (2) can be equivalently formulated as $H(X|Z) \leq H(X)$ for a particular coupling of a discrete variable X taking values with probabilities $\{p_i\}$, and a variable Z with density f_i when conditioned on $X = x_i$. From this perspective the super-concavity bound of Theorem 8 gives $H(X|Z) \leq \tilde{C}(W)$. One should also note that when $\{p_i\}_{i=1}^n$ is a finite sequence, $H(X|Z)$ can be bounded by Fano's inequality, which we now recall. For a Markov triple of random variables $X \to Z \to \hat{X}$, and $e = \{X \neq \hat{X}\}$, Fano's inequality states that

$$H(X|Z) \leq H(e) + \mathbb{P}(e) \log(\#\mathcal{X} - 1), \qquad (7)$$

where $\#\mathcal{X}$ denotes the cardinality of the set $\mathcal{X}$, and we have employed the notation for a measurable set A, $H(A) = -\mathbb{P}(A)\log\mathbb{P}(A) - (1 - \mathbb{P}(A))\log(1 - \mathbb{P}(A))$. This can be restated in terms of mixture distributions in the following way,

$$h(f) \geq \sum_{i=1}^n p_i h(f_i) + H(p) - (H(e) + \mathbb{P}(e)\log(\#\mathcal{X} - 1)).$$

To compare the strength of the bounds derived in Theorem 8 to Fano's we compare $H(e) + \mathbb{P}(e)\log(\#\mathcal{X} - 1)$ and $\tilde{C}(W)$; as is established in Section IV, even in simple cases, $\tilde{C}(W)$ can be arbitrarily small even while $\min_{\hat{X}} H(e) + \mathbb{P}(e)\log(\#\mathcal{X} - 1)$ is arbitrarily large.

The study of entropy of mixtures has a long history and is scattered in a variety of papers that often have other primary emphases. Consequently it is difficult to exhaustively review

[4]To see this, consider $|x|, |y| < \lambda$, then $|T_{ij}x - T_{kj}y| = \left| \frac{\sigma_j x - \mu_i + \mu_j}{\sigma_i} - \frac{\sigma_j y - \mu_k + \mu_j}{\sigma_k} \right| \geq \left| \frac{\mu_j - \mu_i}{\sigma_i} + \frac{\mu_j - \mu_k}{\sigma_k} \right| - \left| \frac{\sigma_j x}{\sigma_i} - \frac{\sigma_j y}{\sigma_k} \right|$. Since $\left| \frac{\sigma_j x}{\sigma_i} - \frac{\sigma_j y}{\sigma_k} \right| < \lambda\sigma_j(\sigma_i^{-1} + \sigma_k^{-1})$, rearranging gives $|T_{ij}x - T_{kj}y| > 0$.

all related work. Nonetheless, the references that we were able to find that attempt to obtain refined bounds under various circumstances are [1], [9], [34], [37], [58], [63]. In all of these papers, however, either the bounds deal with specialized situations, or with a general setup but employing different (and typically far more) information than we require. We emphasize that our bounds deal with general multidimensional situations (including in particular multivariate Gaussian mixtures, which are historically and practically of high interest) and in that sense go beyond the previous literature.

The article is organized as follows. In Section II we will give notation and preliminaries, where we delineate definitions and relationship for entropies, conditional entropies emphasizing a mix of continuous and discrete variables. Section III is devoted to the proof of Theorem 1 (a preliminary version has appeared in the conference paper [54]). In Section IV we prove Theorem 8. These results give a considerable generalization of earlier work of authors in [55], [76]. The result will hinge on a Lemma 2 bounding the sum $\sum_i \varphi(x_i)$ for x_i well spaced and φ log-concave and spherically symmetric, and a concentration result from convex geometry [28]. We close discussing bounds of $\mathbb{P}(|W| \geq t)$ in the case that W is log-concave and strongly log-concave, see Corollary 8. Section V we demonstrate applications of the theorems to a diverse group of problems; hypothesis testing, capacity estimation, nanoscale energetics, and functional inequalities.

II. Notation and Preliminaries

In this part of the article, we will elucidate definitions and results for conditional entropy and mutual information, when a mix of discrete valued and continuous random variables are involved. We will assume that

1) X takes values in a countable set $\mathcal{X}$, indexed by $\mathbb{N}$ and that $\mathbb{P}(X = x_i) = p_i > 0$.
2) Z is a random variable which takes values in a Polish space E. The conditional distribution is described by $\mathbb{P}(Z \in A | X = x_i) = \int_A f_i(z) d\gamma(x)$ where $f_i(z)$ is a density function with respect to a Radon reference measure γ.

We will denote by m the counting measure on $\mathcal{X}$, so that for $A \subseteq \mathcal{X}$, the measure of A is its cardinality,

$$m(A) = \#(A).$$

We denote integration with respect to the counting measure, corresponding to summation, in the following way; for $g : \mathcal{X} \to \mathbb{R}$ such that $\sum_{x \in \mathcal{X}} |g(x)| < \infty$,

$$\int_{\mathcal{X}} g(x) dm \coloneqq \sum_{x \in \mathcal{X}} g(x) = \sum_{i=1}^{\infty} g(x_i)$$

We will denote by $dm\, d\gamma$ the product measure on $\mathcal{X} \times E$ where, for $A \subseteq \mathcal{X}$ and measurable $B \subseteq E$,

$$\int_{\mathcal{X} \times E} \mathbb{1}_{A \times B}(x, z) dm(x)\, d\gamma(z) \coloneqq m(A)\gamma(B),$$

when γ denotes the d-dimensional Lebesgue measure. We will use $|B|_d$ or $|B|$ when there is no risk of confusion to denote the Lebesgue volume of a measureable set B. For measures μ and ν on a shared measure space, we say that μ has a density φ with respect to ν when,

$$\mu(A) = \int_A \varphi \, d\nu \tag{8}$$

holds for every measurable A. Such a φ will also be written as $\frac{d\mu}{d\nu}$.

For random variables U and V whose induced probability measures admit densities with respect to a reference measure γ, in the sense that $\mu(A) \coloneqq \mathbb{P}(U \in A) = \int_A u d\gamma$ and $\nu(A) \coloneqq \mathbb{P}(V \in A) = \int_A v d\gamma$ where u and v are density functions with respect to γ, the relative entropy (or KL divergence) is defined as

$$D(U||V) \coloneqq D(\mu||\nu) \coloneqq D(u||v) \coloneqq \int u \log \frac{u}{v} d\gamma, \tag{9}$$

where we take here and throughout $\log$ to be the natural logarithm. When U is an E valued random variable, with density u with respect to a reference measure γ, denote the entropy

$$h_\gamma(U) \coloneqq h_\gamma(u) = -\int u(z) \log u(z) d\gamma(z), \tag{10}$$

whenever the above integral is well-defined. When $E = \mathbb{R}^d$ and $d\gamma = dx$ is the Lebesgue measure we denote the usual differential entropy,

$$h(U) \coloneqq h(u) \coloneqq -\int u(x) \log u(x) dx, \tag{11}$$

When U is discrete, taking values $x_i \in \mathcal{X}$ with probability p_i, define

$$H(U) \coloneqq H(p) \coloneqq -\sum_{i=1} p_i \log p_i. \tag{12}$$

In the case of a variable taking only two values, we identify $p = (t, 1-t)$ with $t \in [0, 1]$, we define $H(t)$ to the entropy of a Bernoulli random variable with parameter t, $H(t) \coloneqq -(1-t)\log(1-t) - t \log t$, for $t \notin [0, 1]$ we define $H(t) = \infty$. When A is an event, $H(A) \coloneqq H(\mathbb{P}(A))$.

The following propositions elucidate notions of conditional entropy and joint entropy of a mix of discrete and continuous random variables. They are likely known to some readers, but we include their proofs for completeness.

Proposition 1: Suppose X is a discrete random variable with values in a countable set $\mathcal{X}$ and Z is a Borel measurable random variable taking values in E. Suppose, for all $x_i \in \mathcal{X}$, $P(Z \in A | X = x_i) = \int_A f_i(z)\, d\gamma(z)$ for density $f_i(z)$ with respect to a common reference measure γ. Then the following hold.

- The joint distribution of (X, Z) on $\mathcal{X} \times E$, has a density
$$F(x_i, z) = p_i f_i(z) \tag{13}$$
with respect to $dm\, d\gamma$.
- Z has a density
$$f(z) = \sum_{i=1}^{\infty} p_i f_i(z) \tag{14}$$
with respect to γ on E.

- The conditional density of X with respect to $Z = z$ defined as

$$p(x_i|z) = \begin{cases} \frac{p_i f_i(z)}{f(z)} & \text{for } f(z) > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

satisfies $\mathbb{P}(X = x_i) = \int_E p(x_i|z) f(z) dz = p_i$.

Proof: Note that since a set $A \subseteq \mathcal{X} \times E$ can be decomposed into a countable union of disjoint sets $\{x_i\} \times A_i$ where $A_i = \{z \in E : (x_i, z) \in A\}$, to prove $F(x_i, z)$ is the joint density function of (X, Z). Thus, by additivity of measure,[5] it suffices to prove $\mathbb{P}_{XZ}(\{x_i\} \times A_i) = \int_{\{x_i\} \times A_i} F(x_i, z) dm\, d\gamma$. By Bayes' Theorem, $\mathbb{P}_{XZ}(\{x_i\} \times A_i) = \mathbb{P}(X = x_i)\mathbb{P}(Z \in A_i | X = x_i)$, which by definition is $p_i \int_{A_i} f_i(z) d\gamma, = \int_{\{x_i\}} \left[p_i \int_A f_i(z)\, d\gamma\right] dm = \int_{\{x_i\} \times A_i} F(x, z) dm\, d\gamma$. This gives the first claim.

For the second, $\mathbb{P}(Z \in A) = \sum_{i=1}^{\infty} \mathbb{P}(X = x_i, Z \in A) = \sum_{i=1}^{\infty} p_i \int_A f_i(z) d\gamma = \int_A f(z) d\gamma$.

The last assertion is immediate, $\int_E p(z|x_i) f(z) d\gamma = \int_E \frac{p_i f_i(z)}{f(z)} f(z) d\gamma = p_i$. □

Proposition 1 allows the following definitions of conditional entropies.

$$\begin{aligned} h_\gamma(Z|X) &\coloneqq \mathbb{E}_x[h_\gamma(Z|X = x)] \\ &\coloneqq -\sum_{i=1}^{\infty} p_i \int_{z \in E} f_i(z) \log f_i(z) d\gamma \\ &= \sum_{i=1}^{\infty} p_i h_\gamma(f_i), \end{aligned} \tag{16}$$

and

$$\begin{aligned} H(X|Z) &\coloneqq \mathbb{E}_Z[H(X|Z = z)] \\ &\coloneqq -\int_E \left(\sum_{i=1}^{\infty} p(x_i|z) \log p(x_i|z)\right) f(z) d\gamma. \end{aligned}$$

Let us note how the entropy of a mixture can be related to its relative entropy with respect to a dominating distribution g. The entropy concavity deficit of a convex combination of densities f_i, is the convexity deficit of the relative entropy with respect to a reference measure in the following sense.

Proposition 2: For a density g such that $\sum_i p_i D(f_i||g) < \infty$,

$$h_\gamma(f) - \sum_i p_i h_\gamma(f_i) = \sum_i p_i D(f_i||g) - D(f||g). \tag{17}$$

Proof:

$$\begin{aligned} \sum_i p_i D(f_i||g) - D(f||g) &= \sum_i p_i \int \left(f_i \log \frac{f_i}{g} - f_i \log \frac{f}{g}\right) d\gamma \\ &= \sum_i p_i \int \left(f_i \log f_i - f_i \log f\right) d\gamma \\ &= h_\gamma(f) - \sum_i p_i h_\gamma(f_i). \end{aligned}$$

□

[5]Indeed, $\mathbb{P}_{XZ}(A)\mathbb{P}_{XZ}(\cup_i \{x_i\} \times A_i) = \sum_{i=1}^{\infty} \mathbb{P}_{XZ}(\{x_i\} \times A_i)$ while $\int_A F(x, z) dm\, d\gamma = \int_{\cup_i \{x_i\} \times A_i} F(x, z) dm\, d\gamma = \sum_{i=1}^{\infty} \int_{A_i} F(x_i, z)\, d\gamma$. and the result would follow.

Note that the left hand side of Proposition 2 is invariant with respect to g. Thus for g_1, g_2 such that $\sum_i p_i D(f_i||g_j) < \infty$,

$$\sum_i p_i D(f_i||g_1) - D(f||g_1) = \sum_i p_i D(f_i||g_2) - D(f||g_2). \tag{18}$$

Taking $g_1 = g$ and $g_2 = f = \sum_i p_i f_i$ yields the compensation identity,

$$\sum_i p_i D(f_i||g) = \sum_i p_i D(f_i||f) + D(f||g), \tag{19}$$

which is often used to obtain its immediate corollary

$$\min_g \sum_i p_i D(f_i||g) = \sum_i p_i D(f_i||f). \tag{20}$$

We define the mutual information between probability measures $\mathbb{P}_U$ and $\mathbb{P}_V$ with joint distribution $\mathbb{P}_{UV}$ and their product distribution $\mathbb{P}_U \mathbb{P}_V$, as the relative entropy of their joint distribution with respect to the product distribution,

$$I(\mathbb{P}_U; \mathbb{P}_V) = D(\mathbb{P}_{UV}||\mathbb{P}_U \mathbb{P}_V).$$

For the random variables U and V inducing probability measures $\mathbb{P}_U$ and $\mathbb{P}_V$, we will write $I(U; V) = I(\mathbb{P}_U; \mathbb{P}_V)$.

Proposition 3: For X discrete with $\mathbb{P}(X = x_i) = p_i$ and Z satisfying $\mathbb{P}(Z \in B | X = x_i) = \int_B f_i(z) dz$,

$$I(X; Z) = H(X) - H(X|Z) \tag{21}$$

$$= \sum_i p_i D(f_i||f). \tag{22}$$

$$= h_\gamma(Z) - h_\gamma(Z|X) \tag{23}$$

Proof: By Proposition 1, $\mathbb{P}_{XZ}$ has density $F(x_i, z) = p_i f_i(z)$ with respect to $dm(x_i)\, d\gamma(z)$ the product of the counting measure m and γ. The product measure $\mathbb{P}_X \mathbb{P}_Z$, has density $G(x_i, z) = p_i f(z)$ with respect to $dm\, d\gamma$ and it follows that

$$\frac{d\mathbb{P}_{XZ}}{d\mathbb{P}_X \mathbb{P}_Z}(x_i, z) = \frac{\frac{d\mathbb{P}_{XZ}}{dm\, d\gamma}(x_i, z)}{\frac{d\mathbb{P}_X \mathbb{P}_Z}{dm\, dz}(x_i, z)} = \frac{F(x_i, z)}{G(x_i, z)} = \frac{f_i(z)}{f(z)}. \tag{24}$$

By equation (24),

$$\begin{aligned} D(\mathbb{P}_{XZ}||\mathbb{P}_X \mathbb{P}_Z) &= \int_{\mathcal{X} \times E} F(x_i, z) \log \frac{f_i(z)}{f(z)} dm\, d\gamma \\ &= \int_E \sum_i p_i f_i(z) \log \frac{f_i(z)}{f(z)} d\gamma(z). \end{aligned}$$

Recalling $p(z|x)$ from Proposition 1, using the algebra of logarithms and Fubini-Tonelli,

$$\begin{aligned} &\int_E \sum_i p_i f_i(z) \log \frac{f_i(z)}{f(z)} d\gamma(z) \\ &\quad = -\sum_i p_i \log p_i \int_E f_i(z) d\gamma(z) \\ &\qquad + \int_E f(z) \sum_i p(x_i|z) \log p(x|z) d\gamma(z) \\ &\quad = H(X) - H(X|Z), \end{aligned}$$

giving (21). By Fubini-Tonelli,

$$\int_E \sum_i p_i f_i(z) \log \frac{f_i(z)}{f(z)} d\gamma(z) = \sum_i p_i \int_E f_i(z) \log \frac{f_i(z)}{f(z)} d\gamma(z),$$

which gives expression (22). By Proposition 2,

$$\sum_i p_i D(f_i||f) = h_\gamma(f) - \sum_i p_i h_\gamma(f_i) = h(Z) - h(Z|X),$$

(23) follows. □

Using Proposition 3, we can give a simple information theoretic proof of a result proved analytically in [11], [85].

Corollary 1: When $\mathcal{X} \subseteq E$, and γ is a Haar measure,[6] then X and Z satisfy,

$$h_\gamma(X+Z) \leq H(X) + h_\gamma(Z|X) \tag{25}$$

which reduces to

$$h_\gamma(X+Z) \leq H(X) + h_\gamma(Z) \tag{26}$$

in the case that X and Z are independent.

Proof: Applying Proposition 3 to X and $\tilde{Z} = X + Z$ we have

$$h_\gamma(X+Z) = H(X) + h_\gamma(X+Z|X) - H(X|X+Z) \tag{27}$$

$$= H(X) + h_\gamma(Z|X) - H(X|X+Z), \tag{28}$$

where the second equality follows from the translation invariance of Haar measures.[7] Since $H(X|X+Z) \geq 0$, (25) follows, while (26) follows from $h_\gamma(Z|X) = h_\gamma(Z)$ under the assumption of independence. □

Incidentally, the main use of Corollary 1 in [85] is to give a rearrangement-based proof of the entropy power inequality (see [48] for much more in this vein).

We will first introduce the notion of f-divergences.

Definition 1: For a convex function $f : (0,\infty) \to \mathbb{R}$, satisfying $f(1) = 0$, and probability measures μ and ν, with densities $u = \frac{d\mu}{d\gamma}$ and $v = \frac{d\nu}{d\gamma}$ with respect to a common reference measure γ, the f divergence from μ to ν is

$$\begin{aligned} D_f(\mu||\nu) &:= D_f(u||v) \\ &:= \int f\left(\frac{u}{v}\right) v d\gamma \\ &:= \int_{\{uv>0\}} f\left(\frac{u}{v}\right) v d\gamma \\ &\quad + f(0)\nu\{u=0\} + f^*(0)\mu\{v=0\} \end{aligned} \tag{29}$$

where $f(0) := \lim_{t \downarrow 0} f(t)$, and $f^*(0) := \lim_{t\to\infty} \frac{f(t)}{t}$, with the convention that $0 \cdot \infty = 0$.

Note that a common reference measure for measures μ and ν always exists, take $\frac{1}{2}(\mu+\nu)$ for instance. To see that the definition of $D_f(\mu||\nu)$ is independent of the choice of reference measure, consider $\tilde{\gamma}$ a measure that γ is absolutely continuous with respect to. By the chain rule for Radon-Nikodym derivatives, $\tilde{u} := \frac{d\mu}{d\tilde{\gamma}} = u\frac{d\gamma}{d\tilde{\gamma}}$ and $\tilde{v} := \frac{d\nu}{d\tilde{\gamma}} = v\frac{d\gamma}{d\tilde{\gamma}}$, so that $\int f(u/v)v d\gamma = \int f(\tilde{u}/\tilde{v})\tilde{v} d\tilde{\gamma}$. To compare two proposed reference measures, γ_0 and γ_1, not necessarily absolutely continuous with respect to one another, one needs only to observe that the γ_i are both absolutely continuous with respect to $\gamma_0 + \gamma_1$.

When the value of a functional of a pair of probability distributions μ, ν is given by (29), we will call the functional an f-divergence. An f-divergence satisfies the following: (i) Non-negativity, $D_f(\mu||\nu) \geq 0$, by Jensen's inequality and (ii) The map $(\mu,\nu) \mapsto D_f(\mu||\nu)$ is convex, which follows from the convexity of $(x,y) \mapsto yf(x/y)$ for f convex. We direct the reader to [42], [67], [69] for further background on f-divergences and their properties. When $f(x) = x\log x$, the divergence induced is the relative entropy.

III. Upper Bounds

In this section we will provide upper bounds on $h_\gamma(f) - \sum_i p_i h_\gamma(f_i)$ through a generalization of Theorem 1. This section is organized as follows. We first introduce the concept of skewing an f-divergence. In short, comparing the f-divergence from a convex combination $(1-t)\mu + t\nu$ to ν, in place of the f-divergence from μ to ν. Skewing provides a more regular version of the original divergence measure. For example the Radon-Nikodym derivative of μ with respect to $(1-t)\mu + t\nu$ always exists even if Radon-Nikodym derivative of μ with respect to ν may not, whereby skew divergence is well-defined unlike divergence while still preserving important features of the original divergence. We first state elementary properties of the skew relative information (corresponding to skewing the relative entropy), with proofs given in an appendix, and then introduce a skew χ^2-divergence which interpolates between the well known Neyman χ^2-divergence and the Pearson χ^2-divergence.

We will pause to demonstrate that the class of f-divergences is stable under skewing and recover as a special case, a recent result of Nielsen [62], that the generalized Jensen-Shannon divergence is an f-divergence. Then we establish several inequalities between the skew relative information and the skew χ^2-divergence. We will show in Theorem 5 that the skew relative information can be controlled by the skew χ^2 divergence extending the classical bound of relative entropy by Pearson χ^2-divergence. Further, using an argument due to Audenart in the quantum setting [3], we show that the rate of decrease of the skew relative information with respect to the skewing parameter can be described exactly as a multiple of the skew χ^2-divergence.

Theorem 6 also appropriates a quantum argument [3] to show that though neither the Neyman or Pearson divergences can be controlled by total variation, their skewed counterparts can be. We harness this bound along side the differential relationship between the two skew divergences to bound the skew relative entropy by the total variation as well. As a brief aside we demonstrate that the this bound is equivalent to a reverse Pinsker type inequality from [82], before using Theorem 6 to give our proof of Theorem 1. Finally to conclude

[6] An inner and outer regular measure γ on a commutative group is a Haar measure when it is finite on compact sets and satisfies $\gamma(x+A) = \gamma(A)$, for all points x and measurable sets A.

[7] Note that by the translation invariance of γ, the conditional densities of $\tilde{Z}$ with respect to γ are of the form $z \mapsto f_i(z - x_i)$. Thus, $h_\gamma(\tilde{Z}|X) = -\sum_i p_i \int f_i(z-x_i)\log f_i(z-x_i) d\gamma(z) = -\sum_i p_i f_i \log f_i d\gamma$.

the section, we demonstrate that one may obtain the classical result of Lin [43], bounding the Jensen-Shannon divergence by total variation as a special case of Theorem 1.

A. Skew Relative Information

We will consider the following generalization of the relative entropy due to Lee.

Definition 2 [40]: For probability measures μ and ν on a common set $\mathcal{Y}$ and $t \in [0,1]$ define their skew relative information

$$S_t(\mu||\nu) \coloneqq D_f(\mu||\nu),$$

where $f(x) = x\log(x/(tx+(1-t))$. In the case that $d\mu = ud\gamma$, and $d\nu = vd\gamma$ we will also write

$$S_t(u||v) = S_t(\mu||\nu).$$

We state some important properties of skew relative information with the proofs provided in the Appendix.

Proposition 4: For probability measures μ and ν on a common set and $t \in [0,1]$ the skew relative information satisfies the following properties.

1) $S_t(\mu||\nu) = D(\mu||t\mu+(1-t)\nu)$. In particular, $S_0(\mu||\nu) = D(\mu||\nu)$.
2) $S_t(\mu||\nu) = 0$ iff $t = 1$ or $\mu = \nu$.
3) For $0 < t < 1$ the Radon-Nikodym derivative of μ with respect to $t\mu + (1-t)\nu$ does exist, and $S_t(\mu||\nu) \leq -\log t$.
4) $S_t(\mu||\nu)$ is convex, non-negative, and decreasing in t.
5) S_t is an f-divergence with $f(x) = x\log(x/(tx+(1-t))$.

Motivated by the fact that the act of skewing the relative entropy preserves its status as an f-divergence we introduce the act of skewing of an f-divergence.

Definition 3: Given a convex function $f : [0,\infty) \to \mathbb{R}$ with $f(1) = 0$ and its associated divergence $D_f(\cdot||\cdot)$, define the r,t-skew of D_f by

$$S_{f,r,t}(\mu||\nu) \coloneqq D_f(r\mu + (1-r)\nu||t\mu + (1-t)\nu). \tag{30}$$

It can be shown that for $t \in (0,1)$, $S_{f,r,t}(\mu||\nu) < \infty$.

Theorem 4: The class of f-divergences is stable under skewing. That is, if f is convex, satisfying $f(1) = 0$, then

$$\hat{f}(x) \coloneqq (tx+(1-t))f\left(\frac{rx+(1-r)}{tx+(1-t)}\right) \tag{31}$$

is convex with $\hat{f}(1) = 0$ as well, so that the r,t skew of D_f defined in (30) is an f-divergence as well.

Proof: If μ and ν have respective densities u and v with respect to a reference measure γ, then $r\mu + (1-r)\nu$ and $t\mu + 1 - t\nu$ have densities $ru + (1-r)v$ and $tu + (1-t)v$

$$S_{f,r,t}(\mu||\nu) = \int f\left(\frac{ru+(1-r)v}{tu+(1-t)v}\right)(tu+(1-t)v)d\gamma \tag{32}$$

$$= \int f\left(\frac{r\frac{u}{v}+(1-r)}{t\frac{u}{v}+(1-t)}\right)(t\frac{u}{v}+(1-t))vd\gamma \tag{33}$$

$$= \int \hat{f}\left(\frac{u}{v}\right)vd\gamma. \tag{34}$$

Since $\hat{f}(1) = f(1) = 0$, we need only prove $\hat{f}$ convex. For this, recall that the conic transform g of a convex function f defined by $g(x,y) = yf(x/y)$ for $y > 0$ is convex. Indeed, for $\lambda \in (0,1)$,

$$f\left(\frac{(1-\lambda)x_1+\lambda x_2}{(1-\lambda)y_1+\lambda y_2}\right)$$

$$= f\left(\frac{(1-\lambda)y_1}{(1-\lambda)y_1+\lambda y_2}\frac{x_1}{y_1} + \frac{\lambda y_2}{(1-\lambda)y_1+\lambda y_2}\frac{x_2}{y_2}\right)$$

$$\leq \frac{(1-\lambda)y_1}{(1-\lambda)y_1+\lambda y_2}f\left(\frac{x_1}{y_1}\right) + \frac{\lambda y_2}{(1-\lambda)y_1+\lambda y_2}f\left(\frac{x_2}{y_2}\right).$$

Thus,

$$((1-\lambda)y_1+\lambda y_2)f\left(\frac{(1-\lambda)x_1+\lambda x_2}{(1-\lambda)y_1+\lambda y_2}\right)$$

$$\leq (1-\lambda)y_1f\left(\frac{x_1}{y_1}\right) + \lambda y_2 f\left(\frac{x_2}{y_2}\right).$$

Our result follows since $\hat{f}$ is the composition of the affine function $A(x) = (rx+(1-r), tx\ +(1-t))$ with the conic transform of f,

$$\hat{f}(x) = g(A(x)). \tag{35}$$

□

Let us note that in the special case that D_f corresponds to relative entropy, Theorem 4 demonstrates that the "Generalized Jensen-Shannon divergence" developed recently by Nielsen see [62, Definition 1] is in fact an f-divergence, as it is defined as the weighted sum of r_i, t-skew divergences associated to the relative entropy.

Corollary 2: For a vector $\alpha \in [0,1]^k$ and $w_i > 0$ such that $\sum_i w_i = 1$, the (α, w)-Jensen-Shannon divergence between two densities p, q defined by:

$$JS^{\alpha,w}(p:q) \coloneqq \sum_{i=1}^{k} w_i D((1-\alpha_i)p+\alpha_i q||(1-\bar{\alpha})p+\bar{\alpha}q)$$

with $\bar{\alpha} = \sum_i w_i\alpha_i$, is an f-divergence.

Proof: By Theorem 4 the mapping $(p,q) \mapsto D((1-\alpha_i)p+\alpha_i q||(1-\bar{\alpha})p+\bar{\alpha}q)$ is an f-divergence, and the result follows since the class of f-divergences is stable under non-negative linear combinations. □

We note that an application of Theorem 7 is used in [53] to upper bound the $JS^{\alpha,w}(p:q)$ by a constant[8] multiple of the total variation. This extends the classical bound of the Jensen-Shannon divergence, see Corollary 4 below. A complimentary lower bound for the $JS^{\alpha,w}$ by the square of the total variation is derived in [53] as well.

We will only further pursue the case that $r = 1$, and write $S_{f,t}(\mu||\nu) \coloneqq S_{f,1,t}(\mu||\nu)$.

We now skew Pearson's χ^2-divergence, which we recall below.

[8]Explicitly, $JS^{\alpha,w}(p:q) \leq H(w)\mathcal{A}\|p-q\|_{TV}$, where $\mathcal{A} \coloneqq \max_i |\alpha_i - \sum_{j\neq i} w_j\alpha_j/(1-w_i)|$. Note that this reduces to the classical bound for the Jensen-Shannon divergence when $k = 2$, $w = (1/2, 1/2)$, and $\alpha = (0,1)$.

Definition 4 [66]: For measures μ and ν define the χ^2-divergence

$$\chi^2(\mu;\nu) = D_f(\mu||\nu),$$

where $f(x) = (1-x)^2$.

Note that the χ^2-divergence is clearly an f divergence, and for $d\mu = ud\gamma$, $d\nu = vd\gamma$, $\chi^2(\mu;\nu) = \int \frac{(u-v)^2}{v} d\gamma$. Its dual divergence, the χ^2-divergence of Neyman [60] differs only by a notational convention $\chi^2_N(\nu;\mu) = \chi^2(\mu;\nu)$, see [42] for more modern treatment and [20] for background on the distances significance in statistics. Now let us present a skew χ^2-divergence, which interpolates the Pearson and Neyman χ^2-divergences.

Definition 5: For $t \in [0,1]$ and measures μ, and ν, define the skew χ^2_t via:

$$\chi^2_t(\mu;\nu) \coloneqq D_f(\mu||\nu),$$

where $f(x) = (x-1)^2/(1+t(x-1))$.

The skew χ^2-divergence, may be more appropriately known as the Györfi-Vajda divergence, who originally introduced it in [32]. In addition to interpolating the Pearson and Neyman χ^2-divergences, the case $t = \frac{1}{2}$ corresponds to the Vincze-Le Cam divergence [39], [83]. We also direct the reader to recent studies of the object in [53], [64].

Proposition 5: The skew χ^2_t divergence satisfies the following,

1) When $d\mu = ud\gamma$ and $d\nu = vd\gamma$ with respect to some reference measure γ, then
$$\chi^2_t(\mu;\nu) = \int \frac{(u-v)^2}{tu+(1-t)v} d\gamma,$$
with the understanding that the integrand is 0 when u and v are both 0.
2) $(1-t)^2\chi^2_t(\mu;\nu) = \chi^2(\mu; t\mu + (1-t)\nu)$.
3) $\chi^2_t(\mu;\nu) = \chi^2_{1-t}(\nu;\mu)$.
4) χ^2_t is an f-divergence with $f(x) = (x-1)^2/(1+t(x-1))$.
5) The skew χ^2_t interpolates the divergences of Pearson and Neyman, $\chi^2_0(\mu;\nu) = \chi^2(\mu;\nu)$ and $\chi^2_1(\mu;\nu) = \chi^2_N(\mu;\nu)$.

Proof: For (1), the formula follows from direct computation. Breaking the integral into three pieces, first

$$\begin{aligned}\int_{\{uv>0\}} \frac{(u-v)^2}{tu+(1-t)v} d\gamma &= \int_{\{uv>0\}} \frac{\left(1-\frac{u}{v}\right)^2}{t\frac{u}{v}+(1-t)} v d\gamma \\ &= \int_{\{uv>0\}} f\left(\frac{u}{v}\right) v d\gamma.\end{aligned}$$

Then the two limiting cases,

$$\begin{aligned}\int_{\{u>0,v=0\}} \frac{(u-v)^2}{tu+(1-t)v} d\gamma &= \int_{\{v=0\}} \frac{u}{t} d\gamma \\ &= f^*(0)\mu\{v=0\},\end{aligned}$$

and

$$\begin{aligned}\int_{\{u=0,v>0\}} \frac{(u-v)^2}{tu+(1-t)v} d\gamma &= \int_{\{u=0\}} \frac{v}{1-t} d\gamma \\ &= f(0)\nu\{u=0\},\end{aligned}$$

gives our first result.

To prove (2), we use (1). Note that $d\mu = ud\gamma$ and $d\nu = vd\gamma$ implies that $d(t\mu + (1-t)\nu) = (tu + (1-t)v)d\gamma$ so that

$$\begin{aligned}\chi^2(\mu; t\mu+(1-t)\nu) &= \int \frac{(u-(tu+(1-t)v))^2}{tu+(1-t)v} d\gamma \\ &= (1-t)^2 \int \frac{(u-v)^2}{tu+(1-t)v} d\gamma \\ &= (1-t)^2 \chi^2_t(\mu;\nu).\end{aligned}$$

It is immediate from (1) that (3) holds. That χ^2_t is an f-divergence follows from (2) and Theorem 4, so that (4) follows. To prove (5), note that $\chi^2_0(\mu;\nu) = \chi^2(\mu;\nu)$ is immediate from the definition. Applying this and symmetry from (3) we have $\chi^2_1(\mu;\nu) = \chi^2_0(\nu;\mu) = \chi^2(\nu;\mu) = \chi^2_N(\mu;\nu)$. □

The skew divergence and skew χ^2 inherit bounds from $t = 0$ case, and enjoy an interrelation unique to the skew setting as described below.

Theorem 5: For probability measures μ and ν and $t \in (0,1)$

$$S_t(\mu||\nu) \le (1-t)^2 \chi^2_t(\mu;\nu) \tag{36}$$

and

$$\frac{d}{dt} S_t(\mu||\nu) = (t-1)\chi^2_t(\mu;\nu). \tag{37}$$

Proof: Recall that when $t = 0$, the concavity of logarithm bounds $\log x$ by its tangent line $x-1$ so that,

$$\begin{aligned}\int \log\left(\frac{u}{v}\right) u d\gamma &\le \int \left(\frac{u}{v}-1\right) u d\gamma \\ &= \int \left(\frac{u}{v}-1\right)^2 v d\gamma,\end{aligned}$$

giving the classical bound,

$$D(\mu||\nu) \le \chi^2(\mu;\nu). \tag{38}$$

Applying (38) to the identities Proposition 4, (1) and Proposition 5, (2) gives

$$\begin{aligned}S_t(\mu||\nu) &= D(\mu||t\mu+(1-t)\nu) \\ &\le \chi^2(\mu; t\mu+(1-t)\nu) \\ &= (1-t)^2\chi^2_t(\mu;\nu).\end{aligned}$$

Applying the identity $(1-t)(y-1) = y - (ty + (1-t))$ we have

$$\begin{aligned}(1-t)\chi^2_t(\mu;\nu) &= \int \frac{(\frac{u}{v}-1)(\frac{u}{v} - (t\frac{u}{v}+(1-t)))}{t\frac{u}{v}+(1-t)} v d\gamma \\ &= \int \frac{\frac{u}{v}-1}{t\frac{u}{v}+(1-t)} u d\gamma - \int (\frac{u}{v}-1) v d\gamma \\ &= \int \frac{u-v}{tu+(1-t)v} u d\gamma. \end{aligned} \tag{39}$$

Observing the expression

$$S_t(\mu||\nu) = \int u \log\left(\frac{u}{tu+(1-t)v}\right) d\gamma,$$

we compute directly,

$$\frac{d}{dt} S_t(\mu||\nu) = -\int \frac{u-v}{tu+(1-t)v} u d\gamma.$$

□

Recall the total variation norm for a signed measure γ to be $\sup_A |\gamma(A)|$, and adopting the notation $x_+ = \max\{x, 0\}$ then we that

$$\|\mu - \nu\|_{TV} = D_f(\mu||\nu) = \int (u - v)_+ d\gamma,$$

with $f(x) = (x-1)_+$.

Theorem 6: For μ and ν, and $t \in (0, 1)$,

$$\chi^2_t(\mu;\nu) \leq \frac{\|\mu - \nu\|_{TV}}{t(1-t)} \tag{40}$$

$$S_t(\mu||\nu) \leq -\log t \|\mu - \nu\|_{TV}. \tag{41}$$

Proof: From identity (39) we have,

$$\begin{aligned}\chi^2_t(\mu;\nu) &= \frac{1}{1-t}\int \frac{u(u-v)}{tu + (1-t)v} d\gamma \\ &\leq \frac{1}{t(1-t)}\int \left(\frac{u}{v} - 1\right)_+ v d\gamma \\ &= \frac{\|\mu - \nu\|_{TV}}{t(1-t)}.\end{aligned}$$

Define the function

$$\varphi(\lambda) \coloneqq S_{e^{-\lambda}}(\mu||\nu),$$

for $\lambda \in [0, \infty)$ and note that $\varphi(0) = D(\mu||\mu) = 0$. Thus by (37), we can write

$$\begin{aligned}\varphi(\lambda) &= \int_0^\lambda \frac{d}{ds} S_{e^{-s}}(\mu||\nu) ds \\ &= \int_0^\lambda e^{-s}(1 - e^{-s})\chi^2_{e^{-s}}(\mu;\nu) ds.\end{aligned}$$

Applying (40) gives

$$S_{e^{-\lambda}}(\mu||\nu) = \varphi(\lambda) \leq \int_0^\lambda \|\mu - \nu\|_{TV} ds = \lambda \|\mu - \nu\|_{TV}.$$

The substitution $t = e^{-\lambda}$ gives (41). □

Observe that (41) of Theorem 6 recovers a reverse Pinsker inequality due to Verdú [82], see also [68] for related upper bounds on relative entropy and Rényi divergences.

Corollary 3 ([82] Theorem 7): For probability measures μ and γ such that $\frac{d\mu}{d\gamma} \leq \frac{1}{\beta}$ with $\beta \in (0, 1)$

$$\|\mu - \gamma\|_{TV} \geq \frac{1-\beta}{\log \frac{1}{\beta}} D(\mu||\gamma).$$

Proof: The hypothesis implies that $\nu = \frac{\gamma - \beta\mu}{1-\beta}$ is a probability measure satisfying $\gamma = \beta\mu + (1-\beta)\nu$. Applying (41)

$$D(\mu||\gamma) = S_\beta(\mu||\nu) \leq -\log\beta \|\mu - \nu\|_{TV} = \frac{-\log\beta}{1-\beta}\|\mu - \gamma\|_{TV}.$$

□

It is easily seen that the two results, (41) and Theorem 7 of [82] are actually equivalent. In contrast the proof of (41) hinges on foundational properties of the divergence metrics, while Verdú leverages the monotonicity of $x \log x/(x-1)$ for $x > 1$.

Theorem 7: Suppose $f = \sum_i p_i f_i$, where f_i are probability density functions with respect to a reference measure γ on a Polish space E, $p_i \in [0, 1)$, $\sum_i p_i = 1$. For the mixture complement of f_j, $\tilde{f}_j(z) = \sum_{i \neq j} \frac{p_i}{1-p_j} f_i$,

$$h_\gamma(f) - \sum_i p_i h_\gamma(f_i) \leq \mathcal{T}_f H(p)$$

where

$$\mathcal{T}_f \coloneqq \sup_i \|f_i - \tilde{f}_i\|_{TV}.$$

Theorem 1 follows by taking γ to be the Lebesgue measure and $E = \mathbb{R}^d$.

Proof: From Proposition 2

$$\begin{aligned}h_\gamma(\sum_i p_i f_i) &= \sum_i p_i h_\gamma(f_i) + \sum_i p_i D(f_i||f) \\ &= \sum_i p_i h_\gamma(f_i) + \sum_i p_i S_{p_i}(f_i||\tilde{f}_i).\end{aligned}$$

By Theorem 6, $S_{p_i}(f_i||\tilde{f}_i) \leq \log\frac{1}{p_i}\|f_i - \tilde{f}_i\|_{TV}$. Applying Hölder's inequality completes the proof,

$$\begin{aligned}\sum_i p_i S_{p_i}(f_i||\tilde{f}_i) &\leq \sum_i p_i \log\frac{1}{p_i}\|f_i - \tilde{f}_i\|_{TV} \\ &\leq \mathcal{T}_f \sum_i p_i \log\frac{1}{p_i},\end{aligned}$$

where we recall $\mathcal{T}_f \coloneqq \sup_i \|f_i - \tilde{f}_i\|_{TV}$. □

Since the total variation of any two measures is bounded above by 1 this is indeed a sharpening of (2). Expressed in random variables it is

$$h_\gamma(Z) \leq \mathcal{T}_f H(X) + h_\gamma(Z|X).$$

When γ is a Haar measure and we apply the above to $\tilde{Z} = X + Z$ this gives

$$h_\gamma(X + Z) \leq \mathcal{T}_f H(X) + h_\gamma(Z|X). \tag{42}$$

The right hand side of (42) reduces to

$$h_\gamma(X + Z) \leq \mathcal{T}_f H(X) + h_\gamma(Z)$$

in the case that X and Z are independent.

Note that the quantity $h_\gamma(\sum_i p_i f_i) - \sum_i p_i h_\gamma(f_i) = \sum_i p_i D(f_i||f)$ can be considered a generalized Jensen-Shannon divergence, as the case that $n = 2$ and $p_1 = p_2 = \frac{1}{2}$ this is exactly the Jensen-Shannon divergence.

Definition 6: For probability measures μ and ν define the Jensen-Shannon divergence,

$$JSD(\mu||\nu) = \frac{1}{2}\left(D(\mu||2^{-1}(\mu + \nu)) + D(\nu||2^{-1}(\mu + \nu)\right).$$

Theorem 1 recovers the classical bound of the Jensen-Shannon divergence by the total variation, due to Lin, see also [77], [78] for other proofs.

Corollary 4 [43]: For μ and ν probability measures,

$$JSD(\mu||\nu) \leq \|\mu - \nu\|_{TV} \log 2.$$

Proof: Apply Theorem 1 to the Jensen-Shannon divergence, and observe that $\mathcal{T}_f = \|\mu - \nu\|_{TV}$ in the case of two summands. □

IV. Lower Bounds

Observe that when a mixture $f = \sum_i p_i f_i$ is composed of densities f_i (with respect to γ), with disjoint support, in the sense that $f_i f_j = 0$ for $i \neq j$, then $h_\gamma(\sum_i p_i f_i) - \sum_i p_i h_\gamma(f_i) = H(p)$ and the trivial upper bound is attained. The results of this section focus on the case that γ is the Lebesgue measure and can be understood as stability results. The results quantify for log-concave densities, a natural heuristic, densities with small overlap, are near equality in (1). Let us state our assumptions and notations for this section.

1) X is a random variable taking values in countable space $\mathcal{X}$, such that for $i \in \mathcal{X}$, $\mathbb{P}(X = i) = p_i$.
2) Z is an $\mathbb{R}^d$ valued random variable, with conditional densities, f_i satisfying,
$$\mathbb{P}(Z \in A | X = i) = \int_A f_i(z)dz = \mathbb{P}(T_i(W) \in A),$$
for T_i a $\sqrt{\tau}$ bi-Lipschitz[9] and W is a spherically symmetric log-concave random vector with density φ.
3) There exists $\lambda, M > 0$ such that for any i, j,
$$\#\{k : T_{kj}(B_\lambda) \cap T_{ij}(B_\lambda) \neq \emptyset\} \leq M,$$
with $\#$ denoting cardinality and $T_{ij} \coloneqq T_i^{-1} \circ T_j$.

Our assumption that W is log-concave and spherically symmetric is equivalent to W possessing a density φ that is spherically symmetric in the sense that $\varphi(x) = \varphi(y)$ for $|x| = |y|$ and log-concave in the sense that $\varphi((1-t)x+ty) \geq \varphi^{1-t}(x)\varphi^t(y)$ holds for $t \in [0,1]$ and $x, y \in \mathbb{R}^d$. By the spherical symmetry of φ, there exists $\psi : [0,\infty) \to [0,\infty)$ such that $\varphi(x) = \psi(|x|)$. By Radamacher's theorem, Lipschitz continuous functions are almost everywhere differentiable, and since by definition, bi-Lipschitz functions have Lipschitz inverses, the following expression for the density of Z, based on 1, is well-defined,

$$f(z) = \sum_i p_i f_i(z) = \sum_i p_i \varphi(T_i^{-1}(z)) \det((T_i^{-1})'(z)).$$

Note that T_i being $\sqrt{\tau}$ bi-Lipschitz implies T_i^{-1} is $\sqrt{\tau}$ bi-Lipschitz as well, thus T_{ij} is τ-bi-Lipschitz, thus after potentially adjusting T'_{ij} on set of measure zero, we have $\frac{1}{\tau} \leq \|T'_{ij}(z)\| \leq \tau$. Under these assumptions we will prove the following generalization of 3.

Theorem 8: For X and Z satisfying the assumptions of Section (IV), and $\varepsilon > 0$,

$$h(Z) - h(Z|X) \geq H(X) - \tilde{C}(W),$$

where $\tilde{C}$ is the following function of φ,

$$\tilde{C}(W) = \tilde{C}(\varphi) = (M-1) + \mathscr{T}(\lambda)(M + h(\varphi)) + \mathscr{T}^{\frac{1}{2}}(\lambda)(\sqrt{d} + K(\varphi))$$

with

$$K(\varphi) \coloneqq \log\left[\tau^d M\left(\|\varphi\|_\infty + \left(\frac{3}{\varepsilon}\right)\omega_d^{-1}\right)\right]\mathscr{T}^{\frac{1}{2}}(\lambda) + d\left(\int_{B_\lambda^c} \varphi(w)\log^2\left[1 + 2\frac{\varepsilon\tau + \tau^2|w|}{\lambda}\right]dw\right)^{\frac{1}{2}} \tag{43}$$

where ω_d denoting the volume of the d-dimensional unit ball, B_λ^c denotes the complement of $B_\lambda \in \mathbb{R}^d$, and we recall the notation $\mathscr{T}(\lambda) = \mathbb{P}(|W| > \lambda)$.

Note that when $M = 1$, Theorem 8 reduces to Theorem 3. Let us observe that the $K(\varphi)$ bound can be simplified by choosing $\varepsilon = \lambda$, and that $\log^2(x)$ is a concave function for $x \geq e$, and $1 + 2\left(\tau + \frac{\tau^2|w|}{\lambda}\right) \geq 3 \geq e$, so that by Jensen's inequality,

$$\left(\int_{B_\lambda^c} \varphi(w)\log^2\left[1 + 2\left(\tau + \frac{\tau^2|w|}{\lambda}\right)\right]dw\right)^{\frac{1}{2}} \leq \left(\int_{\mathbb{R}^d} \varphi(w)\log^2\left[1 + 2\left(\tau + \frac{\tau^2|w|}{\lambda}\right)\right]dw\right)^{\frac{1}{2}} \leq \log\left(1 + 2\left(\tau + \frac{\tau^2\int\varphi(w)|w|dw}{\lambda}\right)\right).$$

Thus we can further bound

$$K(\varphi) \leq \log\left[\tau^d M\left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right)\omega_d^{-1}\right)\right]\mathbb{P}^{\frac{1}{2}}(|W| > \lambda) + d\log\left(1 + 2\left(\tau + \frac{\tau^2\int\varphi(w)|w|dw}{\lambda}\right)\right).$$

We will have use for the following lemma.

Lemma 1: Let $M \geq 1$ be an integer. Suppose $\mathcal{A} = (A_i)$ is a family of subsets indexed by I such for all i we have $\#\{j \in I : A_i \cap A_j \neq \emptyset\} \leq M$. Then $\mathcal{A}$ can be partitioned into parts $\mathcal{A}_1, \ldots, \mathcal{A}_M$ such that the members of $\mathcal{A}_i$ are pairwise non-intersecting.

Proof: We proceed by induction. If $M = 1$, the result is immediate, so assume the lemma holds for $k < M$. Choose $\mathcal{A}'$ to be a maximal subset of $\mathcal{A}$ such that $A_i, A_j \in \mathcal{A}'$. This implies that $A_i \cap A_j = \emptyset$. For $A_k \notin \mathcal{A}'$ we have

$$\#\{A_j \in \mathcal{A} - \mathcal{A}' : A_k \cap A_j \neq \emptyset\} \leq M - 1.$$

Indeed for every $A_j \in \mathcal{A}$, A_j intersects at most M others, and since $A_k \in \mathcal{A}'$, A_k intersects at least one element of $\mathcal{A}'$ and must intersect at most $M-1$ elements of $\mathcal{A}-\mathcal{A}'$. By induction $\mathcal{A}-\mathcal{A}'$ can be partitioned into no more than $M-1$ collections of disjoint subsets, and the result follows. □

We now derive some implications of our assumptions on T_{ji}; a partitioning result on $T_{ji}(B_\lambda)$ based on the axiom of choice and for the reader's convenience we prove some elementary consequences of the boundedness of the derivatives of T_{ij}.

Proposition 6: For T_i and T_j, $\sqrt{\tau}$ bi-Lipschitz, and $T_{ij} = T_i^{-1} \circ T_j$,

$$T_{ij}(0) + B_{\lambda/\tau} \subseteq T_{ij}(B_\lambda) \subseteq T_{ij}(0) + B_{\lambda\tau}. \tag{44}$$

Proof: As the composition of $\sqrt{\tau}$-bi-Lipschitz functions T_i^{-1} and T_j, T_{ij} is τ-bi-Lipschitz. Thus (44) is a set theoretic

[9]Recall that for $\varepsilon > 0$ a function f is ε bi-Lipschitz when f and its inverse function f^{-1}, satisfy $|f(x) - f(y)| \leq \varepsilon|x - y|$, $|f^{-1}(x) - f^{-1}(y)| \leq \varepsilon|x - y|$ for all x, y. This can be written in a single line as $\frac{|x-y|}{\varepsilon} \leq |f(x) - f(y)| \leq \varepsilon|x - y|$.

statement of the fact that T_{ij} is τ-bi-Lipschitz at 0, that $\frac{|x|}{\tau} \leq |T_{ij}(x) - T_{ij}(0)| \leq \tau|x|$. □

We will need the following concentration result for the information content of a log-concave vector [8], [28], [61], [84].

Theorem 9: For a log-concave density function φ on $\mathbb{R}^d$,

$$\int \left(\log \frac{1}{\varphi(x)} - h(\varphi)\right)^2 \varphi(x)dx \leq d,$$

where $h(\varphi)$ is the entropy of the density φ.

See [27] for a generalization to convex measures, which can be heavy-tailed.

The following upper bounds the sum of a sequence whose values are obtained by evaluating a spherically symmetric density at well spaced points.

Lemma 2: If ϕ is a density on $\mathbb{R}^d$, not necessarily log-concave, given by $\phi(x) = \psi(|x|)$ for $\psi : [0, \infty) \to [0, \infty)$ decreasing, $\lambda > 0$, and a discrete set $\mathcal{X} \subseteq \mathbb{R}^d$ admitting a partition $\mathcal{X}_1, \ldots, \mathcal{X}_M$ such that distinct $x, y \in \mathcal{X}_k$ satisfy $|x-y| \geq 2\lambda$, then there exists an absolute constant $c \leq 3$ such that

$$\sum_{x\in\mathcal{X}} \phi(x) \leq M\left(\|\phi\|_\infty + \left(\frac{c}{\lambda}\right)^d \omega_d^{-1}\right), \tag{45}$$

where $\omega_d = |\{x : |x| \leq 1\}|_d$, where we recall $|\cdot|_d$ as the d-dimensional Lebesgue volume. In particular, if for all $x_0 \in \mathcal{X}$

$$\#\{x \in \mathcal{X} : |x-x_0| < 2\lambda\} \leq M, \tag{46}$$

then (45) holds.

Note, that when $M = 1$ and ϕ is the uniform distribution on a d-dimensional ball of radius R, implicitly determined by $\|\phi\|_\infty = 1$, this reduces to a sphere packing bound,

$$\#\{\text{disjoint } \lambda\text{-balls contained in } B_{R+\lambda}\} \leq 1 + \left(\frac{Rc}{\lambda}\right)^d.$$

From which it follows, due to classical bounds of Minkowski, that $c \geq \frac{1}{2}$.

For the proof below we use the notations $B_\lambda(x) = \{w \in \mathbb{R}^d | \ |w - x| \leq \lambda\}$ and we identify $B_\lambda \equiv B_\lambda(0)$.

Proof: Let us first see that it is enough to prove the result when $M = 1$, as with this case in hand,

$$\sum_{x\in\mathcal{X}} \phi(x) = \sum_{k=1}^M \sum_{x\in\mathcal{X}_k} \phi(x) \leq M\left(\|\phi\|_\infty + \left(\frac{c}{\lambda}\right)^d \omega_d^{-1}\right).$$

We proceed in the case that $M = 1$ and observe that ψ non-increasing enables the following Riemann sum bound,

$$1 = \int \phi(x)dx \tag{47}$$

$$\geq \sum_{k=1}^\infty \psi(k\lambda)\omega_d\lambda^d\left(k^d - (k-1)^d\right), \tag{48}$$

where $\omega_d\lambda^d((k+1)^d - k^d)$ is the volume of the annulus $B_{(k+1)\lambda} - B_{k\lambda}$. Define

$$\Lambda_k \coloneqq \left\{x \in \mathcal{X} : |x| \in [k\lambda, (k+1)\lambda)\ \right\}.$$

Then,

$$\sum_{x\in\mathcal{X}} \phi(x) = \sum_{k=0}^\infty \sum_{x\in\Lambda_k} \phi(x) \tag{49}$$

$$\leq \sum_{k=0}^\infty \#\Lambda_k \psi(k\lambda), \tag{50}$$

as ψ is non-increasing. Let us now bound $\#\Lambda_k$. Using the assumption that any two elements x and y in $\mathcal{X}$ satisfy $|x - y| \geq 2\lambda$,

$$\left|\bigcup_{x\in\Lambda_k}\{x + B_\lambda\}\right| = \#\Lambda_k|B_\lambda| \tag{51}$$

$$= \#\Lambda_k\omega_d\lambda^d, \tag{52}$$

so that it suffices to bound $|\bigcup_{x\in\Lambda_k}\{x + B_\lambda\}|$. Observe that we also have $\bigcup_{x\in\Lambda_k}\{x + B_\lambda\}$ contained in an annulus,

$$\bigcup_{x\in\Lambda_k} B_\lambda(x) \subseteq \{x : |x| \in [(k-1)\lambda, (k+2)\lambda)\},$$

which combined with (52) gives

$$\#\Lambda_k\omega_d\lambda^d \leq |\{x : |x| \in [(k-1)\lambda, (k+2)\lambda)\}|_d$$
$$= \omega_d\lambda^d\left((k+2)^d - (k-1)^d\right),$$

so that

$$\#\Lambda_k \leq (k+2)^d - (k-1)^d. \tag{53}$$

Note the following bound, for $k \geq 1$

$$(k+2)^d - (k-1)^d \leq 3^d\left(k^d - (k-1)^d\right). \tag{54}$$

To see this, observe it is enough to prove $(x+3)^d - x^d \leq 3^d((x+1)^d - x^d)$ for $x > 0$, which after dividing by x^d, and substituting $w = \frac{1}{x}$, this is equivalent to $(1+3w)^d - 1 \leq 3^d((1+w)^d - 1)$ for $w > 0$. Binomial expansion shows that our desired inequality is

$$\sum_{l=1}^d \binom{d}{l} 3^l w^l \leq \sum_{l=1}^d \binom{d}{l} 3^d w^l,$$

which is obviously true. Thus for $k \geq 1$, (53) and (54) give,

$$\#\Lambda_k \leq 3^d\left((k+1)^d - k^d\right).$$

Applying this inequality to (49) gives

$$\sum_{x\in\mathcal{X}} \phi(x) \leq \sum_{k=0}^\infty \sum_{x\in\Lambda_k} \psi(\lambda k) \tag{55}$$

$$\leq \|\phi\|_\infty + \sum_{k=1}^\infty \psi(\lambda k)\#\Lambda_k \tag{56}$$

$$\leq \|\phi\|_\infty + \sum_{k=1}^\infty \psi(\lambda k)3^d\left(k^d - (k-1)^d\right) \tag{57}$$

$$\leq \|\phi\|_\infty + \omega_d^{-1}\left(\frac{3}{\lambda}\right)^d, \tag{58}$$

where (57) follows from the fact that $\#\Lambda_0 \leq 1$ (any $x \in \Lambda_0$ has $0 \in \{x + B_\lambda\}$), and the last inequality follows from the Riemann sum bound (48).

To see that any $\mathcal{X}$ satisfying (46) necessarily satisfy (45), we need only to apply Lemma 1 to $\mathcal{A} = \{A_x\}$ where $A_x = B_{2\lambda}(x)$ to see that such a collection satisfies the hypothesis of the current lemma. □

Lemma 3: For $\{T_i\}$ a collection of $\sqrt{\tau}$-bi-Lipshitz maps, fix j and suppose that $\#\{l : |T_{lj}(0) - T_{ij}(0)| < 2\lambda\} \leq M$ holds for all i, then for $\varepsilon > 0$,

$$\#\{l : |T_{ij}(w) - T_{lj}(w)| < \varepsilon\} \leq M \left(\frac{\lambda + \varepsilon + 2\tau|w|}{\lambda}\right)^d, \tag{59}$$

holds for every i. In particular,

$$\#\{l : |T_{ij}(0) - T_{lj}(0)| < \varepsilon\} \leq M \left(\frac{\varepsilon + \lambda}{\lambda}\right)^d, \tag{60}$$

holds for every i.

Proof: We will first prove and then leverage (60). Note that there is nothing to prove when $\varepsilon \leq 2\lambda$ as (60) is weaker than the assumption. When $2\lambda < \varepsilon$ choose (by Zorn's lemma for instance) Λ to be a maximal subset of $\mathbb{N}$ such that for $k \in \Lambda$, $|T_{kj}(0) - T_{ij}(0)| < \varepsilon$ and $|T_{kj}(0) - T_{k'j}(0)| \geq 2\lambda$ for $k, k' \in \Lambda$, with $k \neq k'$. By construction, for a fixed j, $T_{kj}(0) + B_\lambda$ are disjoint over $k \in \Lambda$ contained in $T_{ij}(0) + B_{\lambda+\varepsilon}$. Thus

$$\begin{aligned}\lambda^d \omega_d \#\Lambda &= \left|\bigcup_{k\in\Lambda} \{T_{kj}(0) + B_\lambda\}\right| \\ &\leq |\{T_{ij}(0) + B_{\lambda+\varepsilon}\}| \\ &= (\lambda + \varepsilon)^d \omega_d,\end{aligned}$$

and we have the following bound on the cardinality of Λ,

$$\#\Lambda \leq \left(\frac{\lambda + \varepsilon}{\lambda}\right)^d. \tag{61}$$

Applying (61), the assumed cardinality bounds, and the maximality of Λ, imply that $\bigcup_{k\in\Lambda} \{T_{kj}(0) + B_{2\lambda}\}$ contains every $T_{lj}(0)$ such that $|T_{lj}(0) - T_{ij}(0)| < \varepsilon$, we have

$$\begin{aligned}&\#\{l : |T_{ij}(0) - T_{lj}(0)| < \varepsilon\} \\ &\qquad \leq \sum_{k\in\Lambda} \#\{m : |T_{mj}(0) - T_{kj}(0)| < 2\lambda\} \\ &\qquad \leq M \left(\frac{\lambda + \varepsilon}{\lambda}\right)^d.\end{aligned}$$

Towards (59), by the mean value theorem, there exists $t \in [0, 1]$ such that

$$T_{ij}(w) - T_{lj}(w) = T_{ij}(0) - T_{lj}(0) + (T'_{ij}(tw) - T'_{lj}(tw))w.$$

Note that if $|T_{ij}(w) - T_{lj}(w)| < \varepsilon$, then

$$\begin{aligned}&|T_{ij}(0) - T_{lj}(0)| \\ &\qquad = |T_{ij}(w) - T_{lj}(w) - (T'_{ij}(tw) - T'_{lj}(tw))w| \\ &\qquad \leq |T_{ij}(w) - T_{lj}(w)| + |(T'_{ij}(tw) - T'_{lj}(tw))w| \\ &\qquad \leq \varepsilon + 2\tau|w|.\end{aligned}$$

Thus

$$\begin{aligned}&\#\{l : |T_{ij}(w) - T_{lj}(w)| < \varepsilon\} \\ &\qquad \leq \#\{l : |T_{ij}(0) - T_{lj}(0)| < \varepsilon + 2\tau|w|\}.\end{aligned}$$

Applying (60),

$$\#\{l : |T_{ij}(w) - T_{lj}(w)| < \varepsilon\} \leq M \left(\frac{\lambda + \varepsilon + 2\tau|w|}{\lambda}\right)^d.$$

□

Corollary 5: For a density $\phi(w) = \psi(|w|)$, with ψ decreasing, $\varepsilon > 0$, T_i $\sqrt{\tau}$-bi-Lipschitz, $T_{ij} = T_i^{-1} \circ T_j$, such that there exists $M \geq 1$ such that for a fixed j,

$$|\{k : T_{ij}(B_\lambda) \cap T_{kj}(B_\lambda) \neq \emptyset\}| \leq M,$$

holds for every i, we have,

$$\begin{aligned}&\sum_i \phi(T_{ij}(w)) \det(T'_{ij}(w)) \\ &\qquad \leq \tau^d M \left(1 + 2\frac{\varepsilon\tau + \tau^2|w|}{\lambda}\right)^d \left(\|\phi\|_\infty + \left(\frac{3}{\varepsilon}\right)^d \omega_d^{-1}\right).\end{aligned}$$

Proof: For any k suppose $|T_{ij}(0) - T_{kj}(0)| < 2\lambda/\tau$, then $\{T_{ij}(0) + B_{\lambda/\tau}\} \cap \{T_{kj}(0) + B_{\lambda/\tau}\} \neq \emptyset$, which by Proposition 6 implies $T_{ij}(B_\lambda) \cap T_{kj}(B_\lambda) \neq \emptyset$. Thus, $\#\{k : |T_{ij}(0) - T_{kj}(0)| < 2\lambda/\tau\} \leq M$. By Lemma 3,

$$\begin{aligned}\#\{l : |T_{ij}(w) - T_{lj}(w)| < 2\varepsilon\} &\leq M \left(\frac{\frac{\lambda}{\tau} + 2\varepsilon + 2\tau|w|}{\frac{\lambda}{\tau}}\right)^d \\ &= M \left(1 + 2\frac{\varepsilon\tau + \tau^2|w|}{\lambda}\right)^d.\end{aligned} \tag{62}$$

This shows that (46) holds for $x_i = T_{ij}(w)$, $\lambda = \varepsilon$ and M in (46) identified with $M \left(1 + 2\frac{\varepsilon\tau + \tau^2|w|}{\lambda}\right)^d$. It follows from Lemma 2 that

$$\begin{aligned}&\sum_i \phi(T_{ij}(w)) \\ &\qquad \leq M \left(1 + 2\frac{\varepsilon\tau + \tau^2|w|}{\lambda}\right)^d \left(\|\phi\|_\infty + \left(\frac{3}{\varepsilon}\right)^d \omega_d^{-1}\right).\end{aligned}$$

This, combined with $\|T_{ij}(x)\| \leq \tau$, which implies the determinant bounds $\det(T'_{ij}(w)) \leq \tau^d$, yields,

$$\begin{aligned}&\sum_i \phi(T_{ij}(w)) \det(T'_{ij}(w)) \\ &\qquad \leq \tau^d \sum_i \phi(T_{ij}(w)) \\ &\qquad \leq \tau^d M \left(1 + 2\frac{\varepsilon\tau + \tau^2|w|}{\lambda}\right)^d \left(\|\phi\|_\infty + \left(\frac{3}{\varepsilon}\right)^d \omega_d^{-1}\right).\end{aligned}$$

□

Corollary 6: Consider T_i, $\sqrt{\tau}$-bi-Lipschitz and suppose there exists $M, \lambda > 0$ such that $\#\{k : T_{kj}(B_\lambda) \cap T_{ij}(B_\lambda) \neq \emptyset\} \leq M$ holds for any i, j. Then for any $\varepsilon > 0$, and a spherically symmetric log-concave density $\varphi(x) = \psi(|x|)$,

$$\begin{aligned}&\int_{B_\lambda^c} \varphi(w) \log \left(\sum_i p_i \sum_j \varphi(T_{ij}(w)) \det(T'_{ij}(w))\right) \\ &\qquad \leq K(\varphi) \mathbb{P}^{\frac{1}{2}}(|W| > \lambda),\end{aligned}$$

where

$$K(\varphi) := \log\left[\tau^d M\left(\|\varphi\|_\infty + \left(\frac{3}{\varepsilon}\right)^d \omega_d^{-1}\right)\right] \mathbb{P}^{\frac{1}{2}}(|W| > \lambda) + d\left(\int_{B_\lambda} \varphi(w) \log^2\left[1+2\left(\frac{\tau\varepsilon}{\lambda} + \frac{\tau^2|w|}{\lambda}\right)\right] dw\right)^{\frac{1}{2}}. \tag{63}$$

Note that $K(\varphi)$ depends only on the statistics of φ, its maximum in the first term and the second term is controlled by the mean logarithm of $|W|$, roughly the average number of digits in $|W|$, and not on the configuration of the mixture. The proof does not leverage log-concavity, a stronger assumption than ψ non-increasing,[10] except to ensure that the relevant statistics are finite.

Proof: Applying Corollary 5 with $\varepsilon = \lambda$, gives

$$\sum_i p_i \sum_j \varphi(T_{ij}(w)) \det(T'_{ij}(w)) \le \tau^d M\left(1 + 2\left(\tau + \frac{\tau^2|w|}{\lambda}\right)\right)^d \left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right)^d \omega_d^{-1}\right)$$

Integrating this inequality against $\varphi(w)\mathbb{1}_{B_\lambda^c}$ and applying Cauchy-Schwarz gives the result. □

Proof of Theorem 8: Using $f_i(x) = \varphi(T_i^{-1}(x)) \det((T_i^{-1})'(x))$, applying the substitution $x = T_i(w)$, and recalling the definition of $T_{ij} = T_i^{-1} \circ T_j$ we can write

$$\int f_i(x) \log\left(1 + \frac{\sum_{j\neq i} p_j f_j(x)}{p_i f_i(x)}\right) dx = \int \varphi(w) \log\left(1 + \frac{\sum_{j\neq i} p_j \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{p_i \varphi(w)}\right) dw.$$

By Jensen's inequality,

$$\sum_i p_i \log\left[1 + \frac{\sum_{j\neq i} p_j \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{p_i \varphi(w)}\right] \le \log\left[1 + \frac{\sum_i \sum_{j\neq i} p_j \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right] = \log\left[1 + \frac{\sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right].$$

Thus,

$$\sum_i p_i \int f_i(x) \log\left[1 + \frac{\sum_{j\neq i} p_j f_j(x)}{p_i f_i(x)}\right] dx \le \int \varphi(w) \log\left(1 + \frac{\sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right) dw.$$

[10]Under spherical symmetry and log-concavity $\|\varphi\|_\infty = \varphi(0)$. Indeed, $\varphi(0) = \varphi(\frac{-x+x}{2}) \ge \sqrt{\varphi(-x)\varphi(x)} = \varphi(x)$. Using log-concavity again for $t \in (0,1)$, $\psi(t|x|) = \varphi((1-t)0 + tx) \ge \varphi^{1-t}(0)\varphi^t(x) \ge \varphi(x) = \psi(|x|)$. Thus it follows that ψ is non-increasing.

We will split the integral into two pieces. Using $\log(1+x) \le x$ on B_λ

$$\begin{aligned}
&\int_{B_\lambda} \varphi(w) \log\left(1 + \frac{\sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right) dw \\
&\le \int_{B_\lambda} \sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w)) dw \\
&= \sum_j p_j \sum_{i\neq j} \int_{B_\lambda} \varphi(T_{ji}(w)) \det(T'_{ji}(w)) dw \\
&= \sum_j p_j \left(\sum_{i\neq j} \int_{T_{ji}(B_\lambda)} \varphi(x) dx\right) \\
&= \sum_j p_j \left(\sum_{\{i\neq j : T_{ji}(B_\lambda) \cap B_\lambda \neq \emptyset\}} \int_{T_{ji}(B_\lambda)} \varphi(x) dx\right) \qquad (64) \\
&\quad + \sum_j p_j \left(\sum_{\{i : T_{ji}(B_\lambda) \cap B_\lambda = \emptyset\}} \int_{T_{ji}(B_\lambda)} \varphi(x) dx\right) \\
&\le \sum_j p_j \left(\sum_{\{i\neq j : T_{ji}(B_\lambda) \cap B_\lambda \neq \emptyset\}} \int_{T_{ji}(0) + B_{\lambda\tau}} \varphi(x) dx\right) \qquad (65) \\
&\quad + \sum_j p_j \left(M \int_{B_\lambda^c} \varphi(x) dx\right) \\
&\le \sum_j p_j \left((M-1) \int_{B_{\lambda\tau}} \varphi(x) dx\right) + M \int_{B_\lambda^c} \varphi(x) dx \\
&= (M-1)\mathbb{P}(|W| \le \lambda\tau) + M\mathbb{P}(|W| > \lambda). \qquad (66)
\end{aligned}$$

Inequality (65) follows from the fact that $T_{ji}(B_\lambda) \cap B_\lambda = \emptyset$ implies that $T_{ji}(B_\lambda) \subseteq B_\lambda^c$, giving the bound $\int_{T_{ji}(B_\lambda)} \varphi(x)dx \le \int_{B_\lambda^c} \varphi(x)dx$, while condition (3) of Section IV and by Proposition IV.1 demonstrates that this family can be split into at most M, subfamilies whose members are pairwise disjoint, that $\#\{i : T_{ji}(B_\lambda) \cap B_\lambda = \emptyset\} \le M$. Inequality (66) follows from another application of Proposition 6 and the fact that the map $s(x) = \int_{x+B_{\lambda\tau}} \varphi(z)dz$ is maximized at 0. To see this, observe that s can be realized as the convolution of two spherically symmetric unimodal functions, explicitly $s(x) = \varphi * \mathbb{1}_{B_{\lambda\tau}}(x)$. Since the class of such functions is stable under convolution, see for instance [41, Proposition 8], s is unimodal and spherically symmetric which obviously implies $s(0) \ge s(x)$ for all x.

Using the fact that $T_{ii}(w) = w$,

$$\begin{aligned}
&\int_{B_\lambda^c} \varphi(w) \log\left(1 + \frac{\sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right) dw \\
&= \int_{B_\lambda^c} \varphi(w) \log\left(\sum_j p_j \sum_i \varphi(T_{ji}(w)) \det(T'_{ji}(w))\right) dw \\
&\quad - \int_{B_\lambda^c} \varphi(w) \log \varphi(w) dw \\
&\le K(\varphi)\mathbb{P}^{\frac{1}{2}}(|W| > \lambda) - \int_{B_\lambda^c} \varphi(w) \log \varphi(w) dw,
\end{aligned}$$

where the bound $K(\varphi)$ is defined from Corollary 6. By Cauchy-Schwarz, followed by Theorem 9,

$$\begin{aligned}
&-\int_{B_\lambda^c} \varphi(w)\log\varphi(w)dw \\
&=h(W)\mathbb{P}(|W|\geq\lambda)+\int_{B_\lambda^c}\varphi(w)\left(\log\frac{1}{\varphi(w)}-h(W)\right)dw \\
&\leq h(W)\mathbb{P}(|W|\geq\lambda) \\
&\quad+\sqrt{\int(\log\frac{1}{\varphi(w)}-h(\varphi))^2\varphi(w)dw\int_{B_\lambda^c}\varphi(w)dw} \\
&\quad\leq h(W)\mathbb{P}(|W|\geq\lambda)+\sqrt{d}\,\mathbb{P}^{\frac{1}{2}}(|W|\geq\lambda).
\end{aligned}$$

□

A. Commentary on Theorem 8

Let us comment on the nature of $\mathbb{P}(|W|\geq\lambda)$ for W log-concave. It is well known that in broad generality (see [12]–[14], [30], [44]), log-concave random variables satisfy "sub-exponential" large deviation inequalities. The following is enough to suit our needs.

Lemma 4: [44, Theorem 2.8] When W is a log-concave random vector, $r\geq 1$, and $t>0$, then,

$$\mathbb{P}(|W|>rt)\leq\mathbb{P}(|W|>t)^{\frac{r+1}{2}}.$$

Corollary 7: For a spherically symmetric log-concave random vector W such that $\mathbb{E}W_1^2=\sigma^2$, where W_1 is the random variable given by the first coordinate of W, and $t\geq\sqrt{2\sigma^2 d}$

$$\mathbb{P}(|W|>t)\leq Ce^{-ct},$$

where $C=2^{-1/2}$, $c=\frac{\log 2}{2\sqrt{2d\sigma^2}}$.

Proof: By Chebyshev's inequality $\mathbb{P}(|W|>\sqrt{2\sigma^2 d})\leq\frac{1}{2}$. Hence for $r>1$, by Lemma 4

$$\begin{aligned}
\mathbb{P}(|W|>r\sqrt{2\sigma^2 d})&\leq\mathbb{P}(|W|>\sqrt{2\sigma^2 d})^{\frac{r+1}{2}} \\
&\leq 2^{-\frac{r+1}{2}}.
\end{aligned}$$

Taking $t=r\sqrt{2\sigma^2 d}$ gives the result. □

Lemma 5: Suppose that a spherically symmetric W is unimodal with respect to the Gaussian, in the sense that its density φ has the form $\varphi(x)=\rho(|x|)e^{-|x|^2/2}$ for ρ non-increasing, then,

$$\mathbb{P}(|W|>t)\leq\mathbb{P}(|\mathcal{Z}|>t),$$

where $\mathcal{Z}$ is a standard normal vector.

Note that W unimodal with respect to the Gaussian, includes the so called strongly log-concave vectors (see [70]), those with densities can be represented as $e^{-V(x)-|x|^2/2}$ with V convex.

Proof: Since ρ is non-increasing, define $t_0=\inf_t\rho(t)\leq 1$, and $\Psi(t)=\mathbb{P}(|W|\leq t)-\mathbb{P}(|\mathcal{Z}|\leq t)$. It follows directly, that Ψ is non-decreasing on $[0,t_0]$ and non-increasing on $[t_0,\infty)$. Since $\Psi(0)=\lim_{t\to\infty}\Psi(t)=0$, we have $\Psi(t)\geq 0$, which is equivalent to the claim. □

Corollary 8: Suppose X and z are variables satisfying the conditions of Section IV for $\tau, M=1$, λ, and W possessing spherically symmetric log-concave density φ. Then

$$H(X|Z)\leq\tilde{C}\,2^{-\frac{1+\lambda/\sqrt{2\sigma^2 d}}{4}}.$$

Furthermore, if W is strongly log-concave then

$$H(X|Z)\leq\tilde{C}\,\mathbb{P}^{\frac{1}{2}}(|\mathcal{Z}|>t),$$

where $\mathcal{Z}$ is a standard Gaussian vector and $\tilde{C}=\left(1+\sqrt{d}+h(W)+K(\varphi)\right)$, with $K(\varphi)$ defined as in (43).

Proof: After recalling from Proposition 3 that $H(X|Z)=H(X)-h(Z)+h(Z|X)$, the proof is an immediate application of Corollary 7 and Lemma 5 to Theorem 8. □

V. Applications

Mixture distributions are ubiquitous, and their entropy is a fundamental quantity. In this section we will demonstrate some applications of our results. First we give special attention to how the ideas of Section IV can be sharpened in the case that W is a Gaussian.

Proposition 7: When X and Z satisfy the assumptions of Section IV, for τ, M, λ, T_{ij} and $W\sim\varphi(w)=e^{-|x|^2/2\sigma^2}/(2\pi\sigma^2)^{d/2}$, then

$$H(X|Z)\leq(M-1)\mathbb{P}(|W|\leq\tau\lambda)+J_d(\varphi)\mathbb{P}(|W|>\lambda)$$

with

$$\begin{aligned}
J_d(\varphi)=\log\Bigg[e^{(\lambda/\sigma)^2+M}(\tau e)^d M\left(1+2\tau+2\tau^2+\frac{2\tau^2 d\sigma}{\lambda}\right)^d \\
\times\left(1+\left(\frac{3\sqrt{2\pi}\sigma}{\lambda}\right)^d\omega_d^{-1}\right)\Bigg]
\end{aligned}\tag{67}$$

for $d\geq 2$ and

$$\begin{aligned}
J_1(\varphi)=\log\Bigg[e^{(\lambda/\sigma)^2+M+2}\tau M\left(1+2\tau+2\tau^2+\frac{2\tau^2\sigma^2}{\lambda^2}\right) \\
\times\left(1+\sqrt{\frac{9\pi}{2}}\frac{\sigma}{\lambda}\right)\Bigg],
\end{aligned}$$

when $d=1$.

The proof of the case $d\geq 2$ is given below, a similar argument in the $d=1$ case is given in the appendix as Proposition 11.

Proof: As in the proof of Theorem 8,

$$\begin{aligned}
&H(X|Z)= \\
&\sum_i p_i\int\varphi(w)\log\left(1+\frac{\sum_{j\neq i}p_j\varphi(T_{ji}(w))\det(T'_{ji}(w))}{p_i\varphi(w)}\right)dw \\
&\leq\int_{B_\lambda^c}\varphi(w)\log\left(1+\frac{\sum_j p_j\sum_{i\neq j}\varphi(T_{ji}(w))\det(T'_{ji}(w))}{\varphi(w)}\right)dw \\
&+\int_{B_\lambda}\varphi(w)\log\left(1+\frac{\sum_j p_j\sum_{i\neq j}\varphi(T_{ji}(w))\det(T'_{ji}(w))}{\varphi(w)}\right)dw.
\end{aligned}$$

We use the general bound from the proof of Theorem 8 for

$$\int_{B_\lambda} \varphi(w) \log\left(1 + \frac{\sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right) dw \leq (M-1)\mathbb{P}(|W| \leq \lambda\tau) + M\mathbb{P}(|W| > \lambda). \tag{68}$$

Splitting the integral,

$$\int_{B_\lambda^c} \varphi(w) \log\left(1 + \frac{\sum_j p_j \sum_{i\neq j} \varphi(T_{ji}(w)) \det(T'_{ji}(w))}{\varphi(w)}\right) dw = \int_{B_\lambda^c} \varphi(w) \log\left(\sum_j p_j \sum_i \varphi(T_{ji}(w)) \det(T'_{ji}(w))\right) dw - \int_{B_\lambda^c} \varphi(w) \log \varphi(w) dw.$$

Integrating the pointwise inequality,

$$\log\left[\sum_j p_j \sum_i \varphi(T_{ji}(w)) \det(T'_{ji}(w))\right] \leq \log\left[\tau^d M \left(1 + 2\tau + \frac{2\tau^2|w|}{\lambda}\right)^d \left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right)^d \omega_d^{-1}\right)\right]$$

obtained from Corollary 5 against $\mathbb{1}_{B_\lambda^c}\varphi(w)$, and then applying Jensen's inequality we have

$$\int_{B_\lambda^c} \varphi(w) \log\left(\sum_j p_j \sum_i \varphi(T_{ji}(w)) \det(T'_{ji}(w))\right) dw \tag{69}$$

$$\leq \int_{B_\lambda^c} \varphi(w) \log\left[\tau^d M \left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right)^d \omega_d^{-1}\right)\right] dw \tag{70}$$

$$+ \int_{B_\lambda^c} \varphi(w) \log\left(1 + 2\tau + \frac{2\tau^2|w|}{\lambda}\right)^d dw. \tag{71}$$

Note, (70) is exactly

$$\mathbb{P}(|W| > \lambda) \log\left[\tau^d M \left(\|\varphi\|_\infty + \left(\frac{3}{\lambda}\right)^d \omega_d^{-1}\right)\right], \tag{72}$$

while applying Jensen's inequality to (71), allows an upper bound of

$$\mathbb{P}(|W| > \lambda) \log\left[1 + 2\tau + \frac{2\tau^2 \int \mathbb{1}_{\{|w|>\lambda\}} \varphi(w)|w| dw}{\mathbb{P}(|W| > \lambda)\lambda}\right]^d \leq \mathbb{P}(|W| > \lambda) \log\left[1 + 2\tau + \frac{2\tau^2(\lambda + d\sigma)}{\lambda}\right]^d, \tag{73}$$

where the inequality is an application of Proposition 9, $\int_{B_\lambda^c} \varphi(w)|w| dw \leq (\lambda + d\sigma)\mathbb{P}(|W| > \lambda)$.

Then applying Proposition 8,

$$-\int_{B_\lambda^c} \varphi(w) \log \varphi(w) dw \leq \left(\frac{d}{-}2 \log 2\pi e^2\sigma^2 + \lambda^2/\sigma^2\right) \mathbb{P}(|W| > \lambda) \tag{74}$$

Combining (72), (73), and (74) gives the result. □

For example, when X takes values $\{x_i\} \in \mathbb{R}$ such that $|x_i - x_j| \geq 2\lambda$ and Y is given by $X + W$ where W is independent Gaussian noise with variance σ^2, then Y has density $\sum_i p_i f_i(y)$ with $f_i(y) = \varphi_\sigma(y - x_i)$. Let $T_i(y) = y - x_i$. Thus $T_{ij}(B_\lambda) = B_\lambda - x_j + x_i$, so that $\{T_{kj}(B_\lambda)\}_k$ are disjoint and we can take $M = 1$ and $\tau = 1$. Applying Proposition 7, we have

$$\begin{aligned} H(X|Y) &= H(X|X+W) \\ &\leq J_1(\varphi)\mathbb{P}(|W| > \lambda) \\ &= J_1(\varphi)\mathbb{P}(|\mathcal{Z}| > \lambda/\sigma), \end{aligned}$$

where $\mathcal{Z}$ is a standard Gaussian vector, and

$$J_1(\varphi) = \log\left[e^{\frac{\lambda^2}{2\sigma^2}+2}\left(5 + 2\frac{\sigma^2}{\lambda^2}\right)\left(1 + \sqrt{\frac{9\pi}{2}}\frac{\sigma}{\lambda}\right)\right].$$

We collect this observation in the following Corollary.

Corollary 9: When X is a discrete $\mathbb{R}$ valued random variable taking values $\{x_i\}$ such that $|x_i - x_j| \geq 2\lambda$ for $i \neq j$ and W is an independent Gaussian variable with variance σ^2, then

$$H(X|X+W) \leq \log\left[e^{\frac{\lambda^2}{2\sigma^2}+2}\left(5 + 2\frac{\sigma^2}{\lambda^2}\right)\left(1 + \sqrt{\frac{9\pi}{2}}\frac{\sigma}{\lambda}\right)\right] \times \mathbb{P}\left(|\mathcal{Z}| > \lambda/\sigma\right).$$

A. Fano's Inequality

A multiple hypothesis testing problem is described in the following, an index $i \in \mathcal{X}$ is drawn according to a variable X and then subsequently samples are drawn from the distribution f_i, with a goal of determining the value i. If Z denotes a random variable with $P(Z \in A|X = i) = \int_A f_i(z)dz$, then by the commutativity of mutual information proven in Proposition 3, $H(X|Z) = h(Z|X) - h(Z) + H(X)$. Thus bounds on the mixture distribution are equivalent to bounds on $H(X|Z)$. For $\hat{X} = g(Z)$, Fano's inequality provides the following bound

$$H(X|Z) \leq H(e) + \mathbb{P}(e)\log(\#\mathcal{X} - 1) \tag{75}$$

where $e = \{\hat{X} \neq X\}$ is the occurrence of an error. Fano and Fano-like inequalities are important in multiple hypothesis testing, as they can be leveraged to deliver bounds on the Bayes risk (and hence min/max risk); we direct the reader to [7], [31], [89] for more background. Fano's inequality gives a lower bound on the entropy of a mixture distribution, that can also give a non-trivial improvement on the concavity of entropy through the equality $H(X|Z) = H(X) + h(Z|X) - h(Z)$. Combined with (75),

$$h(\sum_i p_i f_i) - \sum_i p_i h(f_i) \geq H(p) - \left(H(e) + \mathbb{P}(e)\log(|\mathcal{X}| - 1)\right). \tag{76}$$

In concert with Theorem 1 we have the following corollary.

Corollary 10: For X distributed on indices $i \in \mathcal{X}$, and Z such that $Z|\{X = i\}$ is distributed according to f_i, then given an estimator $\tilde{X} = f(Z)$, with $e = \{X \neq \tilde{X}\}$

$$(1 - \mathcal{T}_f)H(X) \leq H(e) + \mathbb{P}(e)\log(|\mathcal{X}| - 1).$$

Proof: By Fano's inequality $H(e) + \mathbb{P}(e)\log(N-1) \geq H(X|Z)$. Recalling that $H(X|Z) = H(X) - (h(\sum_i p_i f_i) - \sum_i p_i h(f_i))$ and by Theorem 1

$$H(X) - (h(\sum_i p_i f_i) - \sum_i p_i h(f_i)) \geq H(X) - \mathcal{T}_f H(X),$$

gives our result. □

Heuristically, this demonstrates that "good estimators" are only possible for hypothesis distributions discernible in total variation distance. For example in the simplest case of binary hypothesis testing where $n = 2$, the inequality is $(1 - \|f_1 - f_2\|_{TV})H(X) \leq H(e)$, demonstrating how the quality of an estimator of X is limited explicitly by the total variation distance of the two densities.

We note that the pursuit of good estimators $\hat{X}$ is a non-trivial problem in most interesting cases, so much so that Fano's inequality is often used to provide a lower bound on the potential performance of a general estimator by the ostensibly simpler quantity $H(X|Z)$, as determining an optimal value for $\mathbb{P}(e)$ is often intractable. A virtue of Theorem 8 is that it provides upper bounds on $H(X|Z)$, in terms of tail bounds of a single log-concave variable $|W|$. Thus, Theorem 8 asserts that for a large class of models, $H(X|Z)$ can be controlled by a single easily computable quantity, which in the case that $M = 1$, decays sub-exponentially in λ to 0. However, the example delineated below, demonstrates that even in simple cases where an optimal estimator of X admits explicit computation, the bounds derived from Theorem 8 may outperform the best possible bounds based on Fano's inequality.

Suppose that X is uniformly distributed on $\{1, 2, \ldots, N\}$ and that W is an independent, symmetric log-concave variable with density φ, and $Z = X + W$, then Z has density $f(z) = \sum_{i=1}^{N} \frac{f_i(z)}{N}$ with $f_i(z) = \varphi(z - i)$. The optimal (Bayes) estimator of X is given by $\Theta(z) = \arg\max_i \{f_i(z) : i \in \{1, 2, \ldots, N\}\}$, which by the assumption of symmetric log-concavity can be expressed explicitly as:

$$\Theta(z) = \mathbb{1}_{(-\infty, \frac{3}{2})} + \sum_{i=2}^{N-1} i\mathbb{1}_{(i-\frac{1}{2}, i+\frac{1}{2})} + N\mathbb{1}_{(N-\frac{1}{2}, \infty)}.$$

Thus, $\mathbb{P}(\Theta \neq X)$ can be written explicitly as well. Indeed,

$$\begin{aligned}
&\mathbb{P}(X = \Theta(Z)) \\
&\quad = \sum_i \mathbb{P}(X = i, i = \Theta(i + W)) \\
&\quad = \frac{1}{N}\mathbb{P}\left(W \leq \frac{1}{2}\right) + \frac{1}{N}\mathbb{P}\left(W \geq -\frac{1}{2}\right) \\
&\qquad\qquad + \sum_{i=2}^{N-1} \frac{\mathbb{P}\left(W \in (-\frac{1}{2}, \frac{1}{2})\right)}{N} \\
&\quad = \mathbb{P}\left(W \in \left(-\frac{1}{2}, \frac{1}{2}\right)\right) + \frac{2}{N}\mathbb{P}\left(W \geq \frac{1}{2}\right).
\end{aligned}$$

Thus writing $P(e) = \mathbb{P}(X \neq \Theta)$, we have

$$\begin{aligned}
P(e) &= 2\mathbb{P}(W \geq \frac{1}{2})\left(1 - \frac{1}{N}\right) \\
&= \mathbb{P}(|W| \geq 1/2)\left(1 - \frac{1}{N}\right),
\end{aligned}$$

Thus the optimal bounds achievable through Fano's inequality are described by,

$$H(X|X + W) \leq H(e) + P(e)\log(N - 1) \tag{77}$$

with $P(e) = \mathbb{P}(|W| \geq 1/2)\left(1 - \frac{1}{N}\right)$. Note that with $N \to \infty$, the bounds attainable through Fano's inequality become meaningless since $\lim_{N\to\infty} H(e) + P(e)\log(N-1) = \infty$ for any W with support not contained in $[-1/2, 1/2]$. For example in the Gaussian case, Corollary 9 gives the following bound, with $\lambda = \frac{1}{2}$:

$$\begin{aligned}
&H(X|X + W) \leq \\
&\qquad \log\left[e^{\frac{1}{8\sigma^2}+2}\left(5 + 8\sigma^2\right)\left(1 + \sqrt{18\pi}\sigma\right)\right]\mathbb{P}\left(|W| \geq 1/2\right),
\end{aligned} \tag{78}$$

independent of N.

Comparing (77) and (78) to understand at what scale the bounds derived here should be used, we write $p = \mathbb{P}(|W| > 1/2)$ and $\tilde{N} = 1 - \frac{1}{N}$, to compute,

$$\begin{aligned}
&H(e) + P(e)\log(N-1) \\
&= -p\tilde{N}\log(p\tilde{N}) - (1 - p\tilde{N})\log(1 - p\tilde{N}) + p\tilde{N}\log(N-1) \\
&\geq -p\tilde{N}\log(p\tilde{N}) + (1 - p\tilde{N})p\tilde{N} + p\tilde{N}\log(N-1) \\
&\geq p\tilde{N}\left(-\log\tilde{N} + (1 - \tilde{N}) + \log(N-1)\right) \\
&= p\left[\frac{N-1}{N}\left(\frac{1}{N} + \log N\right)\right] \\
&\geq (.98)\log(N-1)\mathbb{P}(|W| \geq 1/2).
\end{aligned}$$

Where the inequalities follow from $-x\log x \geq -x$, the equation being decreasing in p, and the last through calculus or numerical verification.

Thus, it follows that the optimal bounds derived from Fano's inequality are outperformed by equation (78) for $N \geq \log\left[e^{\frac{1}{8\sigma^2}+1.02}\left(5 + 8\sigma^2\right)\left(1 + \sqrt{18\pi}\sigma\right)\right] + 1$. A depiction of such N is given below.

B. Channel Capacity

In the case of a channel that admits discrete inputs (and possibly continuous inputs as well) with output density $f_i(z) = p(z|i)$ when conditioned on an input i. Suppose the input X takes value i with probability p_i then the output Z distribution will have a density function $\sum_i p_i f_i$. Thus

$$\begin{aligned}
I(Z; X) &= h(Z) - h(Z|X) \\
&= H(X) - H(X|Z) \\
&= h(\sum_i p_i f_i) - \sum_i p_i h(f_i).
\end{aligned}$$

Thus, any choice of input X gives a lower bound on the capacity of the channel. In the context of additive white

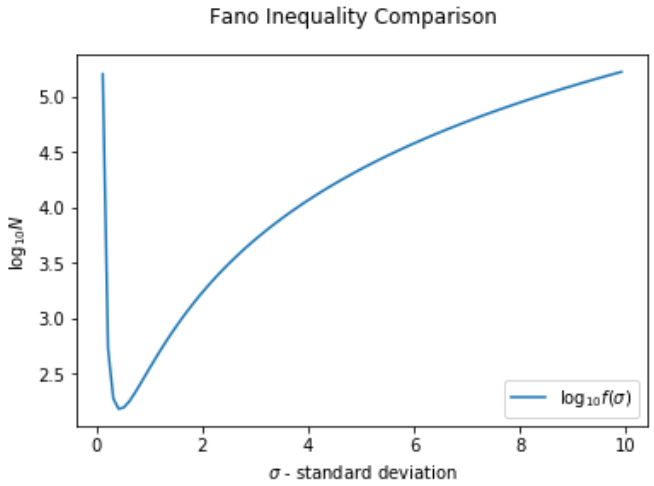

Fig. 1. For $N \geq f(\sigma) := \log\left[e^{\frac{1}{8\sigma^2}+1.02}\left(5+8\sigma^2\right)\left(1+\sqrt{18\pi}\sigma\right)\right]+1$, the bounds derived from Theorem 2 out perform the optimal bounds achievable through Fano's inequality.

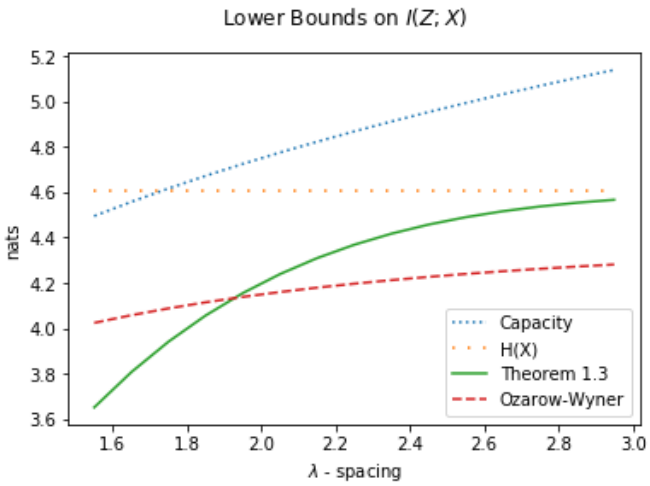

Fig. 2. AWGN channel with average power restraint on input X, corresponding to $N = 100$ uniform input and λ-spacing.

Gaussian noise channel [65] gave rigorous bounds to the findings of [80], that finite input can nearly achieve capacity.

Ozarow and Wyner gave the following bounds, analytically confirming an observation of Ongerboeck, in his celebrated paper [80].

Theorem 10 (Ozarow-Wyner [65]): Suppose X is uniformly distributed on N evenly spaced points, $\{2\lambda, 4\lambda, \ldots, 2N\lambda\}$ and its variance $\sigma^2 = \mathbb{E}X^2 - \mathbb{E}^2X = \lambda^2\left(\frac{N^2-1}{3}\right)$. If $Z = X + W$, where W is Gaussian with variance one and independent of X. Then

1)

$$I(Z;X) \geq (1-(\pi K)^{-1/2}e^{-K})H(X) - H((\pi K)^{-1/2}e^{-K}) \tag{79}$$

where

$$K = \frac{3}{2\alpha^2}(1-e^{-2C})$$
$$\alpha = Ne^{-C}$$
$$C = \frac{1}{2}\log\left(1+\lambda^2\frac{N^2-1}{3}\right).$$

2)

$$I(Z;X) \geq C - \frac{1}{2}\log\frac{\pi e}{6} - \frac{1}{2}\log\frac{1+\alpha^2}{\alpha^2}.$$

In the notation of this paper $K = \frac{\lambda^2}{2}\left(1-\frac{1}{N^2}\right)$. Defining,

$$p_o := \frac{e^{-\frac{\lambda^2}{2}\left(1-\frac{1}{N^2}\right)}}{\sqrt{\frac{\pi\lambda^2}{2}\left(1-\frac{1}{N^2}\right)}},$$

we can re-write (79) as

$$I(Z;X) \geq H(X) - \left(p_oH(X) + H(p_o)\right). \tag{80}$$

Note that $N2^{-C} = \frac{N}{\sqrt{1+\sigma^2}} = \sqrt{\frac{1+3(\sigma/\lambda)^2}{1+\sigma^2}}$, so that $\alpha \approx \sqrt{3}/\lambda$ for σ large. Thus (79) gives a bound with sub-Gaussian-like convergence in λ to $H(X)$ for fixed N, but gives worse than trivial bounds for fixed λ and $N \to \infty$. In contrast, if we take for example W to be Gaussian with variance 1 by Corollary 9,

$$I(Z;X) \geq H(X) - \log\left[3e^{\lambda^2+3}\left(1+\sqrt{\frac{9\pi}{2\lambda^2}}\right)\right]\mathbb{P}\left(|\mathcal{Z}| > \lambda\right). \tag{81}$$

Comparing the bound on the gap between $I(X|Z)$ and $H(X)$ provided by (80) and Corollary 9, we see that Corollary 9 outperforms Theorem 10 for large λ. Indeed, one can easily find an explicit rational function q such that

$$\frac{p_oH(X) + h(p_o)}{\log\left[3e^{\lambda^2+3}\left(1+\sqrt{\frac{9\pi}{2}}\frac{\sigma}{\lambda}\right)\right]\mathbb{P}\left(|\mathcal{Z}| > \lambda\right)} \geq q(\lambda)e^{\frac{\lambda^2}{2N}}.$$

Additionally, (81) gives a universal bound, independent of N.

These results have been of recent interest, see for example [21], [22], where the results improving and generalizing (2) have been studied in a form

$$H(X) - gap^* \leq I(Z;X) \leq H(X),$$

with an emphasis on achieving gap^* bounds that are independent of N, and viable for more general noise models. The significance of the results of Theorem 8 in this context is that the gap^* bounds provided converge exponentially fast to zero in λ, independent of $H(X)$, while for example in [22], the gap^* satisfies

$$gap^* \geq \frac{1}{2}\log\frac{2\pi e}{12}.$$

Additionally, the tools developed can be extended to perturbations of the Y and signal dependent noise through Theorem 8.

A related investigation of recent interest is the relationship between finite input approximations of capacity achieving distributions, particularly the number of "constellations" needed to approach capacity. For example [86], [87] the rate of convergence in n of the capacity of an n input power constrained additive white Gaussian noise channel to the usual additive

white noise Gaussian channel is obtained. In many practical situations, although a Gaussian input is capacity achieving, discrete inputs are used. We direct the reader to [88] for background on the role this practice plays in Multiple Input-Multiple Output channels pivotal in the development of 5G technology.

Additionally in the amplitude constrained discrete time additive white Gaussian noise channel, the capacity achieving distribution is itself discrete [74]. In fact, many important channels achieve capacity for discrete distributions, see for example [2], [16], [35], [73], [81]. Thus in the case that the noise model is independent of input, the capacity achieving output will be a mixture distribution, and the capacity of the channel is given by calculating the entropy of said mixture.

Theorem 8 shows that for sparse input, relative to the strength of the noise, the mutual information of the input and output distributions is sub-exponentially close to the entropy of the input in the case of log-concave noise, and sub-Gaussian from the entropy of the input in the case of strongly log-concave noise, which includes Gaussian noise as a special case. In contrast Theorem 1 gives a reverse inequality, demonstrating that when the mixture distributions are close to one another in the sense that their total variation distance from the mixture "with themselves removed" is small, then the mutual information is quantifiably lessened.

C. Energetics of Non-Equillibrium Thermodynamics

For a process x_t satisfying the overdamped Langevin stochastic differential equation, $dx_t = -\frac{\nabla U(x_t,t)}{\gamma}dt + \sqrt{2D}d\zeta_t$, where $U : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ is a time varying potential, and ζ_t is a Brownian motion with $D = k_B T/\gamma$, where γ is the viscosity constant, k_B is Boltzman's constant, and T is temperature, one can define natural thermodynamic quantities. In particular, trajectory dependent notions of work done $\mathcal{W}$ on the system (see [36]) and heat dissipated $\mathcal{Q}$, respectively,

$$\mathcal{W} \coloneqq \int_0^{t_f} \partial_t U(x_t, t)dt$$

and

$$\mathcal{Q} \coloneqq -\int_0^{t_f} \nabla_x U(x_t, t) \circ dx_t,$$

where the above is a Stratonovich stochastic integral. Recall that Stratonovich integrals satisfy a chain rule $dU(x_t, t) = \nabla_x U(x_t, t) \circ dx_t + \frac{\partial U}{\partial t}(x_t, t)dt$ so that we immediately have a first law of thermodynamics

$$\begin{aligned}\Delta U &\coloneqq U(t_f, x(t_f)) - U(0, x(0)) \\ &= \int_0^{t_f} \partial_t U(x_t, t)dt + \int_0^{t_f} \nabla_x U(x_t, t) \circ dx_t \\ &= \mathcal{W} - \mathcal{Q}.\end{aligned}$$

Further, if ρ_t denotes the distribution of x_t at time t, satisfying the Fokker-Planck equation then it can be shown [4] (see also [17], [52], [72]),

$$\mathbb{E}\mathcal{Q} = k_B T \left(h(\rho_0) - h(\rho_{t_f})\right) + \int_0^{t_f} \mathbb{E}|v(t, x_t)|^2 dt,$$

where v is mean local velocity (see [4] or as the current velocity in [59]). In the quasistatic limit where the non-negative term $\int_0^{t_f} \langle |v(t, x_t)|^2 \rangle dt$ goes to 0, one has a fundamental lower bound on the efficiency of a process's evolution, the average heat dissipated in a transfer from configuration ρ_0 to ρ_{t_f} is bounded below by the change in entropy.

$$\mathbb{E}\mathcal{Q} \;\geq\; k_B T \left(h(\rho_0) - h(\rho_{t_f})\right). \tag{82}$$

A celebrated example of this inequality is Landauer's principle [38], which proposes fundamental thermodynamic limits to the efficiency of a computer utilizing logically irreversable computations (see also [6]). More explicitly (82) suggests that the average heat dissipated in the erasure of a bit, that is, the act of transforming a random bit to a deterministic 0 is at least $k_B T \log 2$. This can be reasoned to in the above, by presuming the entropy of a random bit should satisfy $h(\rho_0) = \log 2$ and that the reset bit should satisfy $h(\rho_{t_f}) = 0$.

In the context of nanoscale investigations, (like protein pulling or the intracellular transport of cargo by molecular motors) it is often the case that phenomena take one of finitely many configurations with an empirically derived probability. However at this scale, thermal fluctuations can make discrete modeling of the phenomena unreasonable, and hence the distributions ρ_0 and ρ_{t_f} in such problems are more accurately modeled as a discrete distribution disrupted by thermal noise, and are thus, mixture distributions. Consequently, bounds on the entropy of mixture distributions translate directly to bounds on the energetics of nanoscale phenomena [56], [76]. For example, in the context of Landauer's bound, the distribution of the position of a physical bit is typically modeled by a Gaussian bistable well, explicitly by the density

$$f_p(z) = p\frac{e^{-(x-a)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} + (1-p)\frac{e^{-(x+a)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}. \tag{83}$$

The variable p connotes the probability that the bit takes the value 1, and $(1-p)$ the probability the bit takes the value 0. This can be modeled by X_p a Bernoulli variable taking values $\pm a$ and $Z_p = X_p + \sigma\mathcal{Z}$ where $\mathcal{Z}$ is a standard normal, so that Z_p has distribution f_p.

Corollary 11: The average heat dissipated $\mathcal{Q}_0$ in an optimal erasure protocol, resetting a random bit to zero in the framework of (83) can be bounded above and below,

$$\begin{aligned}\tilde{C}_L\ \mathbb{P}(|\mathcal{Z}| > a/\sigma) &\leq \mathbb{E}\mathcal{Q}_0 - k_B T \log 2 \\ &\leq \tilde{C}_U\ \mathbb{P}(|\mathcal{Z}| > a/\sigma)\end{aligned} \tag{84}$$

where

$$\tilde{C}_L = -k_B T \left(\log\left[e^{\frac{a^2}{2\sigma^2}+2}\left(5 + 2\frac{\sigma^2}{a^2}\right)\left(1 + \sqrt{\frac{9\pi}{2}}\frac{\sigma}{a}\right)\right]\right)$$

and

$$\tilde{C}_U = k_B T \left(\log\left[e^{\frac{a^2}{2\sigma^2}+2}\left(5 + 2\frac{\sigma^2}{a^2}\right)\left(1 + \sqrt{\frac{9\pi}{2}}\frac{\sigma}{a}\right)\right]\right).$$

More generally, in the case that the erasure is imperfect, so that the probability of failure is non-negligible we have the

following bound,

$$C_L\,\mathbb{P}(|\mathcal{Z}|>a/\sigma) \le \mathbb{E}\mathcal{Q}_0 - k_BT\left(H(p_0)-H(p_1)\right) \quad (85)$$
$$\le C_U\,\mathbb{P}(|\mathcal{Z}|>a/\sigma) \quad (86)$$

where

$$C_L = \tilde{C}_L + k_BTH(X_{p_1})$$
$$C_U = \tilde{C}_U - k_BTH(X_{p_0}).$$

Proof: First let us note that we understand a random bit to be the case that $p_0 = \frac{1}{2}$, while an erasured bit is to be understood as a deterministic X with $p_1 = 0$. Thus, (84) follows immediately from (85) If we let $\rho_0 = f_{p_0}$ and $\rho_{t_f} = f_{p_1}$, and let $Z_{p_i} = X_{p_i} + \sigma\mathcal{Z}$ denote a variable then (82) gives

$$\mathbb{E}\mathcal{Q}_0 = k_BT\left(h(f_{p_0}) - h(f_{p_1})\right) \quad (87)$$
$$= k_BT\left(I(Z_{p_0};X_{p_0}) - I(Z_{p_1};X_{p_1})\right), \quad (88)$$

where the second equality follows from the fact that both Z_{p_i} variables are conditionally Gaussian of the same variance. Using the Corollary 9,

$$I(Z_{p_i};X_{p_i}) - H(X_{p_i}) \ge$$
$$-\log\left[e^{\frac{a^2}{2\sigma^2}+2}\left(5+2\frac{\sigma^2}{a^2}\right)\left(1+\sqrt{\frac{9\pi}{2}}\frac{\sigma}{a}\right)\right]\mathbb{P}\left(|\mathcal{Z}|>a/\sigma\right),$$

while Theorem 1 applied as in (4) as

$$I(Z_{p_i};X_{p_i}) \le \mathcal{T}_fH(X_{p_i})$$
$$= H(X_{p_i}) - \mathbb{P}(|\mathcal{Z}|>a/\sigma)H(X_{p_i}),$$

since $T_f = 1-\mathbb{P}(|\mathcal{Z}|>a/\sigma)$. Combining these results gives

$$I(Z_{p_0};X_{p_0}) - I(Z_{p_1};X_{p_1}) - H(X_{p_0}) + H(X_{p_1}) \le$$
$$\mathbb{P}(|\mathcal{Z}|>a/\sigma)\log\left[e^{\frac{a^2}{2\sigma^2}+2}\left(5+2\frac{\sigma^2}{a^2}\right)\left(1+\sqrt{\frac{9\pi}{2}}\frac{\sigma}{a}\right)\right]$$
$$-\mathbb{P}(|\mathcal{Z}|>a/\sigma)H(X_{p_0})$$

and

$$I(Z_{p_0};X_{p_0}) - I(Z_{p_1};X_{p_1}) - H(X_{p_0}) + H(X_{p_1}) \ge$$
$$-\mathbb{P}(|\mathcal{Z}|>a/\sigma)\log\left[e^{\frac{a^2}{2\sigma^2}+2}\left(5+2\frac{\sigma^2}{a^2}\right)\left(1+\sqrt{\frac{9\pi}{2}}\frac{\sigma}{a}\right)\right]$$
$$+\mathbb{P}(|\mathcal{Z}|>a/\sigma)H(X_{p_1}).$$

Inserting these equations into (88) completes the proof. □

D. Functional Inequalities

Mixture distributions arise naturally in mathematical contexts as well. For example in [11] Bobkov and Marsiglietti found interesting application of $h(X+Z) \le H(X)+h(Z)$ for X discrete and Z independent and continuous in the investigation of entropic Central Limit Theorem for discrete random variables under smoothing.

In the study of stability in the Gaussian log-Sobolev inequalities, Eldan *et al.* [23], it is proven as Proposition 5 that the deficit in the Gaussian log-Sobolev inequality, defined as

$$\delta(\mu) = \frac{I(\mu||\gamma)}{2} - D(\mu||\gamma)$$

for a measure μ and γ the standard d-dimensional Gaussian measure, and I is relative Fisher information,

$$I(\mu||\gamma) \coloneqq \int_{\mathbb{R}^d}\log\left(\frac{d\mu}{d\gamma}\right)d\gamma,$$

is small for Gaussian mixtures. More explicitly for p_i non-negative numbers summing to 1,

$$\delta\left(\sum_i p_i\gamma_i\right) \le H(p).$$

In the language of Theorem 1 a sharper bound can be achieved.

Corollary 12: When γ_i are translates of the standard Gaussian measure then

$$\delta\left(\sum_i p_i\gamma_i\right) \le \mathcal{T}_fH(p),$$

where $\mathcal{T}_f$ is defined as in Theorem 1.

Proof: By the convexity of the relative Fisher information, the equality $D(\sum_i p_i\gamma_i||\gamma) = \sum_i p_iD(\gamma_i||\gamma)$ $-h(\sum_i p_i\gamma_i) + \sum_i p_ih(\gamma_i)$, $\frac{I(\gamma_i||\gamma)}{2} - D(\gamma_i||\gamma) = 0$, and the application of Theorem 1 we have

$$\delta\left(\sum_i p_i\gamma_i\right) = \frac{I(\sum_i p_i\gamma_i||\gamma)}{2} - D(\sum_i p_i\gamma_i||\gamma)$$
$$\le \sum_i p_i\left(\frac{I(\gamma_i||\gamma)}{2} - D(\gamma_i||\gamma)\right)$$
$$+ h(\sum_i p_i\gamma_i) - \sum_i p_ih(\gamma_i)$$
$$= h(\sum_i p_i\gamma_i) - \sum_i p_ih(\gamma_i)$$
$$\le \mathcal{T}_fH(p).$$

□

VI. Conclusion

The entropy of mixtures of discrete probability distributions has been explored in depth for decades, and many useful bounds have been developed, including some quite recently (see, e.g., [3], which actually treats the quantum setting but includes discrete mixtures as a special case). In the special case of convolutions of discrete probability distributions (which are mixtures of translations of a fixed distribution), even more is known (see, e.g., [47], [50], [51] and references therein). In a different direction, the behavior of differential entropy of infinite (continuous) mixtures of absolutely continuous distributions have also been explored, mainly in the context of convolutions of absolutely continuous distributions (i.e, sums of independent random vectors)– see, e.g., [45], [46], [49] and references therein.

In this article, we focused instead on the differential entropy of mixtures of absolutely continuous distributions, which has received some but comparatively much less attention, and provided tight upper and lower bounds for the same under natural conditions. The efficacy of the bounds is demonstrated, for example, by demonstrating that existing bounds on the conditional entropy $H(X|Z)$ of a random variable

X taking values in a countable set $\mathcal{X}$ conditioned on a continuous random variable Z, become meaningless as the cardinality of the set $\mathcal{X}$ increases while the bounds obtained here remain relevant. Significantly enhanced upper bounds on mutual information of channels that admit discrete input with continuous output are obtained based on the bounds on the entropy of mixture distributions. The technical methodology developed is of interest in its own right whereby connections to existing results either can be derived as corollaries of our more general theorems or are improved upon by our results. These include the reverse Pinsker inequality, bounds on Jensen-Shannon divergence, and bounds that are obtainable via Fano's inequality.

APPENDIX

Proof of Proposition 4: There is nothing to prove in (1), this is exactly the definition of the usual relative entropy from μ to $t\mu+(1-t)\nu$. For (2), by (1) $S_t(\mu||\nu)=0$ iff $D(\mu||t\mu+(1-t)\nu)=0$ which is true iff $\mu=t\mu+(1-t)\nu$ which happens iff $t=1$ or $\mu=\nu$. To prove (3), observe that for a Borel set A

$$\mu(A)\leq\frac{1}{t}(t\mu+(1-t)\nu)(A).$$

This gives the following inequality, from which absolute continuity, and the existence of $\frac{d\mu}{d(t\mu+(1-t)\nu)}$ follow immediately,

$$\frac{d\mu}{d(t\mu+(1-t)\nu)}\leq\frac{1}{t}. \tag{89}$$

Taking logarithms and integrating (89) against μ gives,

$$S_t(\mu||\nu)\leq-\log t.$$

To prove (4), notice that for fixed μ and ν, the map $\Phi_t=t\mu+(1-t)\nu$ is affine, and since the relative entropy is jointly convex [19], convexity in t follows from the computation below.

$$\begin{aligned}S_{(1-\lambda)t_1+\lambda t_2}(\mu||\nu)&=D(\mu||\Phi_{(1-\lambda)t_1+\lambda t_2})\\&=D(\mu||(1-\lambda)\Phi_{t_1}+\lambda\Phi_{t_2})\\&\leq(1-\lambda)D(\mu||\Phi_{t_1})+\lambda D(\mu||\Phi_{t_2})\\&=(1-\lambda)S_{t_1}(\mu||\nu)+\lambda S_{t_2}(\mu||\nu).\end{aligned}$$

Since $t\mapsto S_t(\mu||\nu)$ is a non-negative convex function on $(0,1]$ with $S_1(\mu||\nu)=0$ it is necessarily non-increasing. When $\mu\neq\nu$, $\mu\neq t\mu+(1-t)\nu$ so that $S_t(\mu||\nu)>0$ for $t<1$, so that as a function of t the skew divergence is strictly decreasing. To prove that S_t is an f-divergence recall Definition 1. It is straight forward that S_t can be expressed in form (29) with $f(x)=x\log(x/(tx+(1-t))$. Convexity of f follows from the second derivative computation,

$$f''(x)=\frac{(t-1)^2}{x(tx+(1-t))^2}>0.$$

Since $f(1)=0$ the proof is complete. □

In this section we consider $W\sim\varphi_\sigma$ with $\varphi_\sigma(w)=e^{-|w|^2/2\sigma}/(2\pi\sigma^2)^{\frac{d}{2}}$, and use φ to denote φ_1 and use $\mathcal{Z}$ in place of W in this case.

Proposition 8: For $d\geq2$

$$\begin{aligned}&-\int_{B_\lambda^c}\varphi_\sigma(w)\log\varphi_\sigma(w)dw\\&\qquad\leq\left(\frac{d}{2}\log2\pi e^2\sigma^2+\frac{\lambda^2}{\sigma^2}\right)\mathbb{P}(|W|>\lambda)\end{aligned} \tag{90}$$

Proof: We first show the result for general σ follows from the case that $\sigma=1$. Indeed, assuming (90), the substitution $u=w/\sigma$ gives

$$\begin{aligned}&-\int_{B_\lambda^c}\varphi_\sigma(w)\log\varphi_\sigma(w)dw\\&\qquad=\mathbb{P}(|\mathcal{Z}|>\lambda/\sigma)\;d\log\sigma-\int_{B_{\lambda/\sigma}^c}\varphi(u)\log\varphi(u)du\\&\qquad\leq\left[\frac{d}{2}\log2\pi e^2\sigma^2+\frac{\lambda^2}{\sigma^2}\right]\mathbb{P}(|\mathcal{Z}|>\lambda/\sigma),\end{aligned}$$

where we have applied (90) to achieve the inequality. Since W has the same distribution as $\sigma\mathcal{Z}$, the reduction holds. By direct computation,

$$\begin{aligned}&-\int_{B_\lambda^c}\varphi(w)\log\varphi(w)dw\\&=\int_{B_\lambda^c}\varphi(w)\frac{d}{2}\log2\pi dw+\int_{B_\lambda^c}\frac{|w|^2}{2}\varphi(w)\\&=\mathbb{P}(|\mathcal{Z}|>\lambda)\left(\frac{d}{2}\log2\pi+\frac{(2\pi)^{-d/2}\omega_d\int_\lambda^\infty r^{d+1}e^{-r^2/2}dr}{(2\pi)^{-d/2}\omega_d\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr}\right)\\&=\mathbb{P}(|\mathcal{Z}|>\lambda)\left(\frac{d}{2}\log2\pi+\frac{\lambda^de^{-\lambda^2/2}+d\int_\lambda^\infty r^{d-1}e^{-r^2/2}}{\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr}\right)\\&=\mathbb{P}(|\mathcal{Z}|>\lambda)\left(\frac{d}{2}\log2\pi e^2+\frac{\lambda^de^{-\lambda^2/2}}{\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr}\right).\end{aligned}$$

Using $r^{d-1}\geq r\lambda^{d-2}$ for $r\geq\lambda$ when $d\geq2$,

$$\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr\geq\lambda^{d-2}\int_\lambda^\infty re^{-r^2/2}dr=\lambda^{d-2}e^{-r^2/2}.$$

Thus,

$$-\int_{B_\lambda^c}\varphi(w)\log\varphi(w)dw\leq\mathbb{P}(|\mathcal{Z}|>\lambda)\left(\frac{d}{2}\log2\pi e^2+\lambda^2\right).$$

□

Proposition 9: For $d\geq2$,

$$\int_{B_\lambda^c}\varphi(w)|w|dw\leq(\lambda+d\sigma)\mathbb{P}(|W|>\lambda). \tag{91}$$

Proof: Again we reduce to the case that $\sigma=1$. Substituting $u=w/\sigma$ gives

$$\begin{aligned}\int_{B_\lambda^c}\varphi_\sigma(w)|w|dw&=\sigma\int_{B_{\lambda/\sigma}^c}\varphi(u)|u|du\\&\leq\left(\frac{\lambda}{\sigma}+d\right)\sigma\mathbb{P}(|\mathcal{Z}|>\lambda/\sigma)\\&=(\lambda+d\sigma)\mathbb{P}(|W|>\lambda),\end{aligned}$$

where we have used (91) for the inequality and $\sigma\mathcal{Z}$ being equidistributed with W for the last equality. We now proceed

in the reduced case. By change of coordinates and integration by parts,

$$\frac{\int_{B_\lambda^c}\varphi(w)|w|dw}{\mathbb{P}(|\mathcal{Z}|>\lambda)} = \frac{\int_\lambda^\infty r^d e^{-r^2/2}dr}{\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr}$$
$$= \frac{\left[-r^{d-1}e^{-r^2/2}\right]_\lambda^\infty}{\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr} + \frac{(d-1)\int_\lambda^\infty r^{d-2}e^{-r^2/2}dr}{\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr}$$

Using $r \geq \lambda$, for $d \geq 2$, $\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr \geq \lambda^{d-2}\int_\lambda^\infty re^{-r^2/2}dr = \lambda^{d-2}e^{-\lambda^2/2}$, so that,

$$\frac{\left[-r^{d-1}e^{-r^2/2}\right]_\lambda^\infty}{\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr}. \leq \lambda.$$

Thus, the result will follow if we prove

$$A(\lambda) := d\int_\lambda^\infty r^{d-1}e^{-r^2/2}dr - (d-1)\int_\lambda^\infty r^{d-2}e^{-r^2/2}dr \geq 0. \tag{92}$$

When $\lambda \geq 1$, the result is obviously true. When $\lambda = 0$, a change of variables gives $A(0) = 2^{\frac{d}{2}-1}\left(d\,\Gamma(\frac{d}{2}) - 2^{\frac{-1}{2}}(d-1)\Gamma(\frac{d-1}{2})\right)$. When $d \geq 4$, $\Gamma(d/2) \geq \Gamma((d-1)/2)$, which follows by induction using $\Gamma(z+1) = z\Gamma(z)$ and $\Gamma(5/2) = 3\sqrt{\pi}/4 \geq 1 = \Gamma(2) \geq \Gamma(3/2) = \sqrt{\pi}/2$. That $A(0) \geq 0$ in the cases $d = 2$ and $d = 3$ can be checked directly, using $\Gamma(3/2) = \sqrt{\pi}/2$ and $\Gamma(1/2) = \sqrt{\pi}$. Finally since $A'(\lambda) = \lambda^{d-2}e^{-\lambda^2/2}(d-1) - d\lambda)$, A is increasing on $(0,(d-1)/d)$ and decreasing on $((d-1)/d,\infty)$. Thus, $A(\lambda) \geq 0$, and hence,

$$\frac{\int_{B_\lambda^c}\varphi(w)|w|dw}{\mathbb{P}(|\mathcal{Z}|>\lambda)} \leq \lambda + d.$$

□

Proposition 10: When $d = 1$, so that $W \sim \varphi_\sigma(w) = e^{-x^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$, we have the following bounds for $\lambda > 0$,

$$\sigma^2\varphi_\sigma(\lambda) = \int_\lambda^\infty w\varphi_\sigma(w)dw \leq \int_\lambda^\infty \varphi_\sigma(w)dw\left(\lambda + \frac{\sigma^2}{\lambda}\right) \tag{93}$$

$$-\int_\lambda^\infty \varphi_\sigma(w)\log\varphi_\sigma(w)\;dw \leq \left(1 + \frac{\lambda^2}{2\sigma^2} + \log(\sqrt{2\pi}\sigma)\right)\int_\lambda^\infty \varphi_\sigma(w)\;dw. \tag{94}$$

Proof: The inequality (93) is standard. The inequality can be reduced to the $\sigma = 1$ by applying (93) after change of variables $u = w/\sigma$. The proof then follows from the $\sigma = 1$ case. Recall $\varphi'(w) = -w\varphi(w)$ and observe that the function

$$g(\lambda) = \int_\lambda^\infty \varphi(w)dw - \frac{\lambda^2}{\lambda^2+1}\varphi(\lambda)$$

satisfies $g(0) > 0$, $\lim_{\lambda\to\infty} g(\lambda) = 0$, and has derivative

$$g'(\lambda) = \frac{-2\varphi(\lambda)}{(\lambda^2+1)^2} < 0,$$

so that $g(\lambda) > 0$ which is equivalent to (93). To prove (94), we again reduce to the $\sigma = 1$ case by the substitution $u = w/\sigma$. Then compute directly using integration by parts,

$$-\int_\lambda^\infty \varphi(w)\log\varphi(w)dw$$
$$= \int_\lambda^\infty \log\sqrt{2\pi}\varphi(w)dw + \frac{1}{2}\int_\lambda^\infty w^2\varphi(w)dw$$
$$= \log\sqrt{2\pi}\int_\lambda^\infty \varphi(w)dw + \frac{1}{2}\left(\lambda\varphi(\lambda) + \int_\lambda^\infty \varphi(w)dw\right)$$
$$\leq \left(1 + \frac{\lambda^2}{2} + \log\sqrt{2\pi}\right)\int_\lambda^\infty \varphi(w)dw.$$

The inequality is an application of (93). □

Proposition 11: When X and Z satisfy the conditions of section IV for the one dimensional Gaussian $W \sim \varphi_\sigma(w) = e^{-w^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$,

$$H(X|Z) \leq (M-1)\mathbb{P}(|W| \leq \tau\lambda) + J_1(\varphi)\mathbb{P}(|W| \geq \lambda)$$

with

$$J_1(\varphi) = \left(\frac{\lambda^2}{2\sigma^2} + M + 1\right)\log\gamma(\varphi),$$
$$\gamma(\varphi) \coloneqq \tau M\left(1 + 2\tau + 2\tau^2 + 2\tau^2\frac{\sigma^2}{\lambda^2}\right)\left(1 + \frac{3}{2}\sqrt{\frac{2\pi\sigma^2}{\lambda^2}}\right).$$

Proof: As in the proof of Theorem 8,

$$H(X|Z)$$
$$= \sum_i p_i \int \varphi(w)\log\left(1 + \frac{\sum_{j\neq i} p_j\varphi(T_{ji}(w))\det(T'_{ji}(w))}{p_i\varphi(w)}\right)dw$$
$$\leq \int_{B_\lambda^c}\varphi(w)\log\left(1 + \frac{\sum_j p_j\sum_{i\neq j}\varphi(T_{ji}(w))\det(T'_{ji}(w))}{\varphi(w)}\right)dw$$
$$+ \int_{B_\lambda}\varphi(w)\log\left(1 + \frac{\sum_j p_j\sum_{i\neq j}\varphi(T_{ji}(w))\det(T'_{ji}(w))}{\varphi(w)}\right)dw.$$

with

$$\int_{B_\lambda}\varphi(w)\log\left(1 + \frac{\sum_j p_j\sum_{i\neq j}\varphi(T_{ji}(w))\det(T'_{ji}(w))}{\varphi(w)}\right)dw \leq (M-1)\mathbb{P}(|W| \leq \lambda\tau) + M\mathbb{P}(|W| > \lambda). \tag{95}$$

Splitting the integral,

$$\int_{B_\lambda^c}\varphi(w)\log\left(1 + \frac{\sum_j p_j\sum_{i\neq j}\varphi(T_{ji}(w))\det(T'_{ji}(w))}{\varphi(w)}\right)dw$$
$$= \int_{B_\lambda^c}\varphi(w)\log\left(\sum_j p_j\sum_i\varphi(T_{ji}(w))\det(T'_{ji}(w))\right)dw$$
$$- \int_{B_\lambda^c}\varphi(w)\log\varphi(w)dw.$$

Using Corollary 5, Jensen's inequality, and (93), while writing $\beta = \tau M\left(\frac{1}{\sqrt{2\pi\sigma}} + \frac{3}{2\lambda}\right)$

$$\begin{aligned}
&\int_{B_\lambda^c} \varphi(w) \log\left(\sum_j p_j \sum_i \varphi(T_{ji}(w)) \det(T'_{ji}(w))\right) dw \\
&\le \int_{B_\lambda^c} \varphi(w) \log\left[\tau M\left(1+2\left(\tau+\frac{\tau^2|w|}{\lambda}\right)\right)\left(\|\varphi\|_\infty+\frac{3}{2\lambda}\right)\right] x \\
&\le \mathbb{P}(|W|>\lambda)\log\left[\beta\left(1+2\tau+\frac{2\tau^2\int \mathbb{1}_{\{|w|>\lambda\}}\varphi(w)|w|dw}{\mathbb{P}(|W|>\lambda)\lambda}\right)\right] \\
&\le \mathbb{P}(|W|>\lambda)\log\left[\beta\left(1+2\tau+\frac{2\tau^2(\lambda+\frac{\sigma^2}{\lambda})}{\lambda}\right)\right]. \qquad (96)
\end{aligned}$$

Then applying (94),

$$\begin{aligned}
&-\int_{B_\lambda^c} \varphi(w)\log\varphi(w)dw \\
&\qquad \le \left(1+\frac{\lambda^2}{2\sigma^2}+\log\sqrt{2\pi\sigma^2}\right)\mathbb{P}(|W|>\lambda). \qquad (97)
\end{aligned}$$

Combining (95), (96), and (97) we have

$$H(X|Z) \le (M-1)\mathbb{P}(|W|\le\lambda\tau) + J_1(\varphi)\mathbb{P}(|W|>\lambda).$$

□

ACKNOWLEDGMENT

The authors thank three anonymous reviewers for many insightful comments and suggestions which improved the quality of this paper. In particular, they express their gratitude for the formulation of Lemma 1, due to the reviewers.

REFERENCES

[1] E. Abbe and A. Barron, "Polar coding schemes for the AWGN channel," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2011, pp. 194–198.
[2] I. C. Abou-Faycal, M. D. Trott, and S. Shamai (Shitz), "The capacity of discrete-time memoryless Rayleigh-fading channels," *IEEE Trans. Inf. Theory*, vol. 47, no. 4, pp. 1290–1301, May 2001.
[3] K. M. R. Audenaert, "Quantum skew divergence," *J. Math. Phys.*, vol. 55, no. 11, Nov. 2014, Art. no. 112202.
[4] E. Aurell, K. Gawedzki, C. Mejía-Monasterio, R. Mohayaee, and P. Muratore-Ginanneschi, "Refined second law of thermodynamics for fast random processes," *J. Stat. Phys.*, vol. 147, no. 3, pp. 487–505, May 2012.
[5] T. Austin, "Measure concentration and the weak Pinsker property," *Mathématiques l'IHÉS*, vol. 128, no. 1, pp. 1–119, Nov. 2018.
[6] C. H. Bennett, "Logical reversibility of computation," *IBM J. Res. Develop.*, vol. 17, no. 6, pp. 525–532, Nov. 1973.
[7] L. Birge, "A new lower bound for multiple hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1611–1615, Apr. 2005.
[8] S. Bobkov and M. Madiman, "Concentration of the information in data with log-concave distributions," *Ann. Probab.*, vol. 39, no. 4, pp. 1528–1543, 2011.
[9] S. Bobkov and M. Madiman, "The entropy per coordinate of a random vector is highly constrained under convexity conditions," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 4940–4954, Aug. 2011.
[10] S. G. Bobkov, N. Gozlan, C. Roberto, and P.-M. Samson, "Bounds on the deficit in the logarithmic Sobolev inequality," *J. Funct. Anal.*, vol. 267, no. 11, pp. 4110–4138, 2014.
[11] S. G. Bobkov and A. Marsiglietti, "Entropic CLT for smoothed convolutions and associated entropy bounds," *Int. Math. Res. Notices*, vol. 2020, no. 21, pp. 8057–8080, Nov. 2020.
[12] S. Bobkov and J. Melbourne, "Hyperbolic measures on infinite dimensional spaces," *Probab. Surv.*, vol. 13, pp. 57–88, Jan. 2016.
[13] S. G. Bobkov and J. Melbourne, "Localization for infinite-dimensional hyperbolic measures," in *Doklady Mathematics*, vol. 91, no. 3. New York, NY, USA: Springer, May 2015, pp. 297–299.
[14] C. Borell, "Complements of Lyapunov's inequality," *Mathematische Annalen*, vol. 205, no. 4, pp. 323–331, Dec. 1973.
[15] J. Briët and P. Harremoës, "Properties of classical and quantum Jensen–Shannon divergence," *Phys. Rev. A, Gen. Phys.*, vol. 79, no. 5, pp. 283–304, May 2009.
[16] T. H. Chan, S. Hranilovic, and F. R. Kschischang, "Capacity-achieving probability measure for conditionally Gaussian channels with bounded inputs," *IEEE Trans. Inf. Theory*, vol. 51, no. 6, pp. 2073–2088, May 2005.
[17] R. Chetrite, G. Falkovich, and K. Gawedzki, "Fluctuation relations in simple examples of non-equilibrium steady states," *J. Stat. Mech., Theory Exp.*, vol. 2008, no. 8, Aug. 2008, Art. no. P08005.
[18] T. A. Courtade, M. Fathi, and A. Pananjady, "Quantitative stability of the entropy power inequality," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5691–5703, Feb. 2018.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[20] N. Cressie and T. R. Read, "Multinomial goodness-of-fit tests," *J. Roy. Stat. Soc., B (Methodol.)*, vol. 46, no. 3, pp. 440–464, 1984.
[21] A. Dytso, M. Goldenbaum, H. V. Poor, and S. S. Shitz, "A generalized Ozarow-Wyner capacity bound with applications," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1058–1062.
[22] A. Dytso, D. Tuninetti, and N. Devroye, "Interference as noise: Friend or foe?" *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3561–3596, Jun. 2016.
[23] R. Eldan, J. Lehec, and Y. Shenfeld, "Stability of the logarithmic Sobolev inequality via the Föllmer process," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 56, no. 3, pp. 2253–2269, Aug. 2020.
[24] R. Eldan and D. Mikulincer, "Stability of the Shannon–Stam inequality via the Föllmer process," *Probab. Theory Rel. Fields*, vol. 77, pp. 1–32, Mar. 2020.
[25] M. Fathi, E. Indrei, and M. Ledoux, "Quantitative logarithmic Sobolev inequalities and stability estimates," *Discrete Continuous Dyn. Syst.-A*, vol. 36, no. 12, p. 6835, 2016.
[26] A. Figalli, F. Maggi, and A. Pratelli, "Sharp stability theorems for the anisotropic Sobolev and log-Sobolev inequalities on functions of bounded variation," *Adv. Math.*, vol. 242, pp. 80–101, Aug. 2013.
[27] M. Fradelizi, J. Li, and M. Madiman, "Concentration of information content for convex measures," *Electron. J. Probab.*, vol. 25, pp. 1–22, Jan. 2020.
[28] M. Fradelizi, M. Madiman, and L. Wang, "Optimal concentration of information content for log-concave densities," in *High Dimensional Probability VII, The Cargèse Volume* (Progress in Probability) C. Houdré, D. Mason, P. Reynaud-Bouret, and J. Rosinski, Eds. Basel, Switzerland: Birkhäuser, 2016.
[29] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 529–536.
[30] O. Guédon, "Kahane-Khinchine type inequalities for negative exponent," *Mathematika*, vol. 46, no. 1, pp. 165–173, Jun. 1999.
[31] A. Guntuboyina, "Lower bounds for the minimax risk using f-divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, Apr. 2011.
[32] L. Györfi and I. Vajda, "A class of modified Pearson and Neyman statistics," *Statist. Risk Model.*, vol. 19, no. 3, pp. 239–251, Jan. 2001.
[33] A. S. Holevo, "*Quantum Systems, Channels, Information: A Mathematical Introduction*, vol. 16. Berlin, Germany: Walter de Gruyter, 2012.
[34] M. F. Huber, T. Bailey, H. Durrant-Whyte, and U. D. Hanebeck, "On entropy approximation for Gaussian mixture random vectors," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst.*, Aug. 2008, pp. 181–188.
[35] W. Huleihel, Z. Goldfeld, T. Koch, M. Madiman, and M. Medard, "Design of discrete constellations for peak-power-limited complex Gaussian channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 556–560.
[36] C. Jarzynski, "Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 56, no. 5, p. 5018, 1997.
[37] K. Kampa, E. Hasanbelliu, and J. C. Principe, "Closed-form Cauchy-Schwarz PDF divergence for mixture of Gaussians," in *Proc. Int. Joint Conf. Neural Netw.*, Jul. 2011, pp. 2578–2585.
[38] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM J. Res. Develop.*, vol. 5, no. 3, pp. 183–191, Jul. 1961.
[39] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York, NY, USA: Springer, 2012.

[40] L. Lee, "Measures of distributional similarity," in *Proc. 37th Annu. Meeting Assoc. Comput. Linguistics*, 1999, pp. 25–32.
[41] J. Li, A. Marsiglietti, and J. Melbourne, "Further investigations of Rényi entropy power inequalities and an entropic characterization of s-concave densities," *Geometric Aspects Funct. Anal.*, vol. 2266, no. 4, p. 95, 2020.
[42] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.
[43] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[44] L. Lovász and M. Simonovits, "Random walks in a convex body and an improved volume algorithm," *Random Struct. Algorithms*, vol. 4, no. 4, pp. 359–412, 1993.
[45] M. Madiman and F. Ghassemi, "Combinatorial entropy power inequalities: A preliminary study of the Stam region," *IEEE Trans. Inf. Theory*, vol. 65, no. 3, pp. 1375–1386, 2018.
[46] M. Madiman and I. Kontoyiannis, "Entropy bounds on abelian groups and the Ruzsa divergence," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 77–92, Jan. 2018.
[47] M. Madiman, A. W. Marcus, and P. Tetali, "Information-theoretic inequalities in additive combinatorics," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Cairo, Egypt, Jan. 2010, pp. 1–4.
[48] M. Madiman, J. Melbourne, and P. Xu, "Forward and reverse entropy power inequalities in convex geometry," in *Convexity Concentration*. New York, NY, USA: Springer, 2017, pp. 427–485.
[49] M. Madiman, P. Nayar, and T. Tkocz, "Two remarks on generalized entropy power inequalities," in *Geometric Aspects of Functional Analysis* (Lecture Notes in Mathematics), vol. 2266, B. Klartag and E. Milman, Eds. Cham, Switzerland: Springer, 2020, pp. 169–185.
[50] M. Madiman, L. Wang, and J. O. Woo, "Majorization and Rényi entropy inequalities via Sperner theory," *Discrete Math.*, vol. 342, no. 10, pp. 2911–2923, Oct. 2019.
[51] M. Madiman, L. Wang, and J. O. Woo, "Rényi entropy inequalities for sums in prime cyclic groups," *SIAM J. Discrete Math.*, vol. 35, no. 3, pp. 1628–1649, Jan. 2021.
[52] C. Maes, K. Netočný, and B. Wynants, "Steady state statistics of driven diffusions," *Phys. A, Stat. Mech. Appl.*, vol. 387, no. 12, pp. 2675–2689, May 2008.
[53] J. Melbourne, "Strongly convex divergences," *Entropy*, vol. 22, no. 11, p. 1327, 2020.
[54] J. Melbourne, M. Madiman, and M. V. Salapaka, "Relationships between certain f-divergences," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, 2019, pp. 1068–1073.
[55] J. Melbourne, S. Talukdar, S. Bhaban, and M. Salapaka, "Error bounds for a mixed entropy inequality," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1973–1977.
[56] J. Melbourne, S. Talukdar, and M. V. Salapaka, "Realizing information erasure in finite time," in *Proc. IEEE Conf. Decis. Control (CDC)*, Dec. 2018, pp. 4135–4140.
[57] D. Mikulincer, "Stability of Talagrand's Gaussian transport-entropy inequality via the Föllmer process," *Isr. J. Math.*, vol. 242, pp. 1–27, Apr. 2021.
[58] K. Moshksar and A. K. Khandani, "Arbitrarily tight bounds on differential entropy of Gaussian mixtures," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3340–3354, Jun. 2016.
[59] E. Nelson, "*Dynamical Theories of Brownian Motion*, vol. 3. Princeton, NJ, USA: Princeton Univ. Press, 1967.
[60] J. Neyman, "Contribution to the theory of the χ^2 test," in *Proc. Berkeley Symp. Math. Statist. Probab.*, vol. 1. Berkeley, CA, USA: Univ. California Press Berkeley, 1949, pp. 239–273.
[61] V. H. Nguyen, "Inégalités fonctionnelles et convexité," Ph.D. dissertation, Dept. Math., Université Pierre et Marie Curie (Paris VI), Paris, France, Oct. 2013.
[62] F. Nielsen, "On a generalization of the Jensen–Shannon divergence and the Jensen–Shannon centroid," *Entropy*, vol. 22, no. 2, p. 221, Feb. 2020.
[63] F. Nielsen and R. Nock, "MaxEnt upper bounds for the differential entropy of univariate continuous distributions," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 402–406, Apr. 2017.
[64] T. Nishiyama and I. Sason, "On relations between the relative entropy and χ^2-divergence, generalizations and applications," *Entropy*, vol. 22, no. 5, p. 563, May 2020.
[65] L. H. Ozarow and A. D. Wyner, "On the capacity of the Gaussian channel with a finite number of input levels," *IEEE Trans. Inf. Theory*, vol. 36, no. 6, pp. 1426–1428, Nov. 1990.
[66] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, vol. 50, no. 302, pp. 157–175, 1900.
[67] I. Sason, "On f-divergences: Integral representations, local behavior, and inequalities," *Entropy*, vol. 20, no. 5, p. 383, May 2018.
[68] I. Sason and S. Verdu, "Upper bounds on the relative entropy and Rényi divergence as a function of total variation distance for finite alphabets," in *Proc. IEEE Inf. Theory Workshop-Fall (ITW)*, Oct. 2015, pp. 214–218.
[69] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.
[70] A. Saumard and J. A. Wellner, "Log-concavity and strong log-concavity: A review," *Statist. Surv.*, vol. 8, p. 45, Apr. 2014.
[71] P. Schlattmann," *Medical Applications of Finite Mixture Models*. Berlin, Germany: Springer, 2009.
[72] U. Seifert, "Entropy production along a stochastic trajectory and an integral fluctuation theorem," *Phys. Rev. Lett.*, vol. 95, no. 4, Jul. 2005, Art. no. 040602.
[73] S. Shamai (Shitz) and I. Bar-David, "The capacity of average and peak-power-limited quadrature Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 41, no. 4, pp. 1060–1071, Jul. 1995.
[74] J. G. Smith, "The information capacity of amplitude- and variance-constrained scalar Gaussian channels," *Inf. Control*, vol. 18, no. 3, pp. 203–219, 1971.
[75] S. Talukdar, S. Bhaban, and M. Salapaka, "Beating Landauer's bound by memory erasure using time multiplexed potentials," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7645–7650, Jul. 2017.
[76] S. Talukdar, S. Bhaban, J. Melbourne, and M. Salapaka, "Analysis of heat dissipation and reliability in information erasure: A Gaussian mixture approach," *Entropy*, vol. 20, no. 10, p. 749, Sep. 2018.
[77] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1602–1609, Jul. 2000.
[78] F. Topsøe, "Jenson–Shannon divergence and norm-based measures of discrimination and variation," Dept. Math., Univ. Copenhagen, Copenhagen, Denmark, Tech. Rep., 2003. [Online]. Available: http://web.math.ku.dk/~topsoe/sh.ps
[79] G. Toscani, "A strengthened entropy power inequality for log-concave densities," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6550–6559, Dec. 2015.
[80] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 1, pp. 55–67, Jan. 1982.
[81] L. R. Varshney, "Transporting information and energy simultaneously," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2008, pp. 1612–1616.
[82] S. Verdu, "Total variation distance and the distribution of relative information," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2014, pp. 1–3.
[83] I. Vincze, "On the concept and measure of information contained in an observation," in *Contributions to Probability*. Amsterdam, The Netherlands: Elsevier, 1981, pp. 207–214.
[84] L. Wang, "Heat capacity bound, energy fluctuations convexity," Ph.D. dissertation, Dept. Phys., Yale University, New Haven, CT, USA, May 2014.
[85] L. Wang and M. Madiman, "Beyond the entropy power inequality, via rearrangements," *IEEE Trans. Inf. Theory*, vol. 60, no. 9, pp. 5116–5137, Sep. 2014.
[86] Y. Wu and S. Verdu, "Functional properties of MMSE," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2010, pp. 1453–1457.
[87] Y. Wu and S. Verdu, "The impact of constellation cardinality on Gaussian channel capacity," in *Proc. 48th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2010, pp. 620–628.
[88] Y. Wu, C. Xiao, Z. Ding, X. Gao, and S. Jin, "A survey on MIMO transmission with finite input signals: Technical challenges, advances, and future trends," *Proc. IEEE*, vol. 106, no. 10, pp. 1779–1833, Oct. 2018.
[89] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, 1999.

James Melbourne (Member, IEEE) received the Ph.D. degree in mathematics from the University of Minnesota in 2015. He was a Post-Doctoral Researcher at the Mathematics Department, University of Delaware, from 2015 to 2017, and a Post-Doctoral Researcher in electrical and computer engineering at the University of Minnesota from 2017 to 2020. He is currently an Assistant Professor at the Department of Probability and Statistics, CIMAT. His research focuses on convexity theory in the context of probabilistic, geometric, and information theoretic inequalities as well as their applications.

Saurav Talukdar received the B.Tech. and M.Tech. degrees in mechanical engineering from the Indian Institute of Technology, Mumbai, India, in 2013, and the Ph.D. degree in mechanical engineering from the University of Minnesota, Minneapolis, USA, in 2018. He was a Battery Algorithm Engineer with Apple working on system identification and thermal management of Lithium ion batteries in iPhones. In 2019, he joined Google, where he is currently a Control Systems and Machine Learning Engineer and focuses on energy optimization for the Google Data Bioadjust Center.

Shreyas Bhaban received the B.Tech. degree in instrumentation and control engineering from the College of Engineering, Pune, Maharashtra, India, in 2011, and the master's and Ph.D. degrees in electrical engineering from the University of Minnesota—Twin Cities, Minneapolis, MN, USA, in 2014 and 2018, respectively. In 2018, he joined Becton Dickinson Biosciences, where he was a Systems Engineer working on system identification and fault tolerance for flow cytometers. In 2021, he joined Abbott Diagnostics Division, where he is currently a Research and Development Systems Engineer working on molecular image analyzer instruments for haematology applications. His research interests include systems and controls engineering, biophysics, nanodynamics, and its applications to opto-mechanical systems and medical devices.

Mokshay Madiman (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, in 1999, and the Sc.M. and Ph.D. degrees in applied mathematics from Brown University, Providence, RI, USA, in 2001 and 2005, respectively. From 2005 to 2012, he was with the Department of Statistics, Yale University, New Haven, CT, USA, first as a Gibbs Assistant Professor, then as an Assistant Professor, and finally as an Associate Professor of statistics and applied mathematics. He has spent a semester each visiting the School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India; the National Mathematics Initiative, Indian Institute of Science, Bengaluru; the Department of Operations Research and Financial Engineering, Princeton University; and the Institute for Mathematics and Its Applications, Minneapolis. Since January 2013, he has been an Associate Professor with the Department of Mathematical Sciences, University of Delaware. From 2014 to 2017, he was an Adjunct Professor of mathematics with the Tata Institute of Fundamental Research. His research interests include probability theory, information theory, combinatorics, functional analysis, and convex geometry. He was awarded the NSF CAREER Award in 2011.

Murti V. Salapaka (Fellow, IEEE) received the B.Tech. degree in mechanical engineering from the Indian Institute of Technology, Madras, in 1991, and the M.S. and Ph.D. degrees in mechanical engineering from the University of California at Santa Barbara in 1993 and 1997, respectively. He was a Faculty Member with the Electrical and Computer Engineering Department, Iowa State University, Ames, from 1997 to 2007. He is currently the Director of graduate studies and the Vincentine Hermes Luh Chair Professor with the Electrical and Computer Engineering Department, University of Minnesota, Minneapolis. His research interests include control and network science, nanoscience, single molecule physics, and sustainable energy. He has received the 1997 National Science Foundation CAREER Award.