

Relationships between certain f -divergences

James Melbourne
Electrical and Computer Engineering
University of Minnesota
Email: melbo013@umn.edu

Mokshay Madiman
Mathematical Sciences
University of Delaware
Email: madiman@udel.edu

Murti V. Salapaka
Electrical and Computer Engineering
University of Minnesota
Email: murtis@umn.edu

Abstract—We investigate the concavity deficit of the entropy functional. Some properties of the skew-divergence are developed and a “skew” χ^2 divergence is introduced. Various relationships between these f -divergences and others are established, including a reverse Pinsker type inequality for the skew divergence, which in turn yields a sharpening on the upper bound for the entropic concavity deficit.

I. INTRODUCTION

The Boltzmann-Shannon differential entropy h is a concave functional, i.e., for densities f and g on $\mathbb{R}^d$, one has

$$h\left(\frac{f+g}{2}\right) \geq \frac{h(f)+h(g)}{2}.$$

The subject of our investigation for this paper will be the deficit of this concavity inequality. More explicitly, for a sequence of densities $f = \{f_i\}_i$ and numbers $p = \{p_i\}$ satisfying $p_i \geq 0$ such that $\sum_i p_i = 1$, the entropy concavity deficit is the quantity

$$\Phi = \Phi_{f,p} = h\left(\sum_i p_i f_i\right) - \sum_i p_i h(f_i). \quad (1)$$

It is known that this gap can be controlled by the entropy of p (see [2], [20]),

$$h\left(\sum_i p_i f_i\right) - \sum_i p_i h(f_i) \leq H(p), \quad (2)$$

where $H(p)$ is the discrete entropy of the weights $p = (p_i)_i$. (When f_i are probability mass functions on a countable set and all differential entropies are replaced by discrete entropies, this is a classical and easy consequence of the chain rule and the nonnegativity of conditional entropy.)

The density $f = \sum_i p_i f_i$ can be realized as the summation of a discrete variable and continuous random variable. That is, if X and Z are jointly distributed such that $\mathbb{P}(X = x_i) = p_i$ and $\mathbb{P}(Z \in A | X = x_i) = \int f_i(z+x_i)dz$ then the random variable $X+Z$ will have density $\sum_i p_i f_i$ and these inequalities can be interpreted as

$$0 \leq h(X+Z) - h(X+Z|X) \leq H(X),$$

which in the case that X and Z are independent reduce to

$$0 \leq h(X+Z) - h(Z) \leq H(X).$$

This yields one application of the inequalities. The quantity Φ is exactly $I(X; X+Z)$, the mutual information of the discrete input X of a channel, subject to continuous (possibly input dependent) additive noise given by Z , with its output $X+Z$. For more background on the physical motivations of the authors we direct the reader to [11], [17].

In this note, we develop some ideas related to the entropy concavity deficit. These ideas are deeply indebted to the work of Audenaert [1] in Quantum Information Theory, some of whose arguments are adapted directly to the classical setting here. Indeed, the concavity deficit of the entropy can be understood as a classical analog of the Holevo Information [5].

The organization of the paper is the following, in Section II we develop a differential relationship between the skew divergence and a to be putforth skew χ^2 divergence, and use this inequality to prove a reverse Pinsker inequality [14] for the skew-divergence that recovers a result of Verdu [19]. In Section III we develop an upperbound on the concavity deficit. This gives a sharpening of the discrete-continuous entropy inequality (2).

Notation

For a probability density function f on $\mathbb{R}^d$, the entropy is

$$h(f) := - \int f(x) \ln f(x) dx,$$

provided the integral is well defined. In the case that Z is a random variable with density function f , we will write $h(Z) = h(f)$. When p_i is a sequence of non-negative numbers summing to one, we take

$$H(p) := - \sum_i p_i \ln p_i$$

to be the discrete Shannon entropy (in “nats”). When X is a random variable taking distinct values x_i with probability p_i , we write $H(X)$ for $H(p)$.

When Z has conditional densities f_x with respect to the values of a random variable X , in the sense that $\mathbb{P}(Z \in A|X = x) = \int_A f_x$ and $x \mapsto h(f_x)$ is integrable define

$$h(Z|X) := \mathbb{E}h(f_x). \quad (3)$$

When X is continuous, (3) is the usual conditional entropy. When X is discrete, the formula reduces to

$$h(Z|X) = \sum_i p_i h(f_i). \quad (4)$$

Definition I.1. For a convex function f satisfying $f(1) = 0$, and probability measures $\mu \ll \nu$ the f divergence from μ to ν is

$$D_f(\mu||\nu) = \int f\left(\frac{d\mu}{d\nu}\right) d\nu. \quad (5)$$

When the value of a functional of a pair of probability distributions μ, ν is given by (5) we will call the functional an f -divergence. An f -divergence satisfies the following.

- Non-negativity, $D_f(\mu||\nu) \geq 0$.
- The map $(\mu, \nu) \mapsto D_f(\mu||\nu)$ is convex.

We direct the reader to [7], [15], [16] for further background on f -divergences and their properties. Let us note the following examples.

- 1) Total Variation

$$\|\mu - \nu\|_{TV} = \int \left(\frac{d\mu}{d\nu} - 1\right)_+ d\nu.$$

where $x_+ := \max\{x, 0\}$, is an f -divergence with $f(x) = (x - 1)_+$. The norm notation is used in reference to the total variation norm for a signed measure γ defined as $\sup_A |\gamma(A)|$.

- 2) Relative entropy,

$$D(\mu||\nu) = \int \log \frac{d\mu}{d\nu} d\mu. \quad (6)$$

Often referred to as the Kullback-Leibler divergence in the literature. Take $f(x) = x \log x$.

- 3) χ^2 -divergence,

$$\chi^2(\mu; \nu) = \int \left(\frac{d\mu}{d\nu} - 1\right)^2 d\nu, \quad (7)$$

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

Several proofs are omitted for brevity here but will be included in a subsequent journal publication.

II. SKEW RELATIVE INFORMATION

Recall that for probability measures $\mu \ll \nu$ we write the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ and the relative entropy or KL divergence,

$$D(\mu||\nu) = \int \log \frac{d\mu}{d\nu} d\mu. \quad (8)$$

In the case that μ is not absolute continuous with respect to ν define $D(\mu||\nu) = \infty$. In the case that μ and ν have a densities u and v with respect to a common reference measure dx , that is $d\mu = udx$ and $d\nu = vdx$,

$$D(\mu||\nu) = \int u(x) \ln \frac{u(x)}{v(x)} dx,$$

and when it is convenient we will also employ the notation $D(u||v) = D(\mu||\nu)$. We will consider the following generalization, first defined in [6].

Definition II.1. For probability measures μ and ν on a common set $\mathcal{X}$ and $t \in [0, 1]$ define their Skew divergence

$$S_t(\mu||\nu) = \int \log \frac{d\mu}{d(t\mu + (1-t)\nu)} d\mu$$

In the case that $d\mu = udx$, and $d\nu = vdx$ we will also write

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

Proposition II.2. 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 $t > 0$ the Radon-Nikodym derivative 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 $f(x) = x \log(x/(tx + (1-t)))$

Proof. 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,

$$\frac{d\mu}{d(t\mu + (1-t)\nu)} \leq \frac{1}{t}. \quad (9)$$

Integrating (9) gives

$$S_t(\mu||\nu) \leq -\log t. \quad (10)$$

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 [3], 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 I.1. It is straight forward that S_t can be expressed in form (5) 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. $\square$

Let us also recall Pearson's χ^2 divergence.

Definition II.3. [13] For measures μ and ν absolutely continuous with respect to a common reference measure dx so that $d\mu = udx$ and $d\nu = vdx$, define

$$\chi^2(\mu; \nu) = \int \left(1 - \frac{d\mu}{d\nu}\right)^2 d\nu = \int \frac{(u-v)^2}{v} dx$$

and $\chi^2(\mu; \nu) = \infty$ when $\frac{d\mu}{d\nu}$ does not exist.

Note that a common reference measure always exists, take $\mu + \nu$ for instance. Formally, the χ^2 divergence of Neyman [12] differs only by a notational convention $\chi_N^2(\nu; \mu) = \chi^2(\mu; \nu)$, see [7] for more modern treatment and [4] 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 II.4. For $t \in [0, 1]$ and measures μ and ν , define

$$\chi_t^2(\mu; \nu) = \int \frac{\left(1 - \frac{d\mu}{d\nu}\right)^2}{t\frac{d\mu}{d\nu} + (1-t)} d\nu.$$

Proposition II.5. The skew χ_t^2 divergence satisfies the following,

- 1) When $d\mu = udx$ and $d\nu = vdx$ with respect to some reference measure dx (not necessarily Lebesgue),

$$\chi_t^2(\mu; \nu) = \int \frac{(u-v)^2}{tu + (1-t)v} dx.$$

- 2) For $t = 0$, $(1-t)^2 \chi_t^2(\mu; \nu) = \chi^2(\mu; t\mu + (1-t)\nu)$.

- 3) $\chi_t^2(\mu; \nu) = \chi_{1-t}^2(\nu; \mu)$
- 4) χ_t^2 is an f -divergence with $f(x) = (x-1)^2/(1+t(x-1))$
- 5) The skew χ_t^2 interpolates the divergences of Neyman and Pearson, $\chi_0^2(\mu; \nu) = \chi^2(\mu; \nu)$ and $\chi_1^2(\mu; \nu) = \chi_N^2(\mu; \nu)$.

Proof. For (1), the formula follows in the case $\mu \ll \nu$, from the fact that on the support of ν ,

$$\frac{u}{v} = \frac{d\mu}{d\nu},$$

so that

$$\begin{aligned} \chi_t^2(\mu; \nu) &= \int \frac{\left(1 - \frac{u}{v}\right)^2}{t\frac{u}{v} + (1-t)} v dx \\ &= \int \frac{(u-v)^2}{tu + (1-t)v} dx. \end{aligned}$$

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

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

It is immediate from (1) that (3) holds. To see that χ_t^2 is an f -divergence, compute

$$\frac{d^2}{dx^2} \frac{(x-1)^2}{1+t(x-1)} = \frac{2}{(1+t(x-1))^3} \geq 0$$

for $x > 0$, and (4) follows easily. To prove (5), note that $\chi_0^2(\mu; \nu) = \chi^2(\mu; \nu)$ is immediate from the definition. Applying this and symmetry from (3) we have $\chi_1^2(\mu; \nu) = \chi_0^2(\nu; \mu) = \chi^2(\nu; \mu) = \chi_N^2(\mu; \nu)$. $\square$

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 II.1. For probability measures μ and ν and $t \in (0, 1)$

$$S_t(\mu\|\nu) \leq (1-t)^2 \chi_t^2(\mu; \nu) \quad (11)$$

and

$$\frac{d}{dt} S_t(\mu\|\nu) = (t-1) \chi_t^2(\mu; \nu). \quad (12)$$

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

$$\int \log \left(\frac{d\mu}{d\nu} \right) d\mu \leq \int \left(\frac{d\mu}{d\nu} - 1 \right) d\mu \quad (13)$$

$$= \int \left(\frac{d\mu}{d\nu} - 1 \right)^2 d\nu, \quad (14)$$

giving the classical bound,

$$D(\mu\|\nu) \leq \chi^2(\mu;\nu). \quad (15)$$

Applying (15) to the identities Proposition II.2, (1) and Proposition II.5,(2) gives

$$S_t(\mu\|\nu) = D(\mu\|t\mu + (1-t)\nu) \quad (16)$$

$$\leq \chi^2(\mu; t\mu + (1-t)\nu) \quad (17)$$

$$= (1-t)^2 \chi_t^2(\mu; \nu). \quad (18)$$

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

$$(1-t)\chi_t^2(\mu; \nu) \quad (19)$$

$$= \int \frac{(\frac{d\mu}{d\nu} - 1)(\frac{d\mu}{d\nu} - (t\frac{d\mu}{d\nu} + (1-t)))}{t\frac{d\mu}{d\nu} + (1-t)} d\nu \quad (20)$$

$$= \int \frac{\frac{d\mu}{d\nu} - 1}{t\frac{d\mu}{d\nu} + (1-t)} d\mu - \int (\frac{d\mu}{d\nu} - 1) d\nu \quad (21)$$

$$= \int \frac{\frac{d\mu}{d\nu} - 1}{t\frac{d\mu}{d\nu} + (1-t)} d\mu. \quad (22)$$

Observing the expression

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

we compute directly,

$$\frac{d}{dt} S_t(\mu\|\nu) = - \int \frac{\frac{d\mu}{d\nu} - 1}{t\frac{d\mu}{d\nu} + (1-t)} d\mu. \quad (23)$$

□

Theorem II.2. For μ and ν , and $t \in (0, 1)$,

$$\chi_t^2(\mu; \nu) \leq \frac{\|\mu - \nu\|_{TV}}{t(1-t)} \quad (24)$$

$$S_t(\mu\|\nu) \leq -\log t \|\mu - \nu\|_{TV}. \quad (25)$$

Proof. From the identity in (19) we have

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

Define the function

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

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

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

Applying (24) 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 (25). □

It can be proven that (25) of Theorem II.2 recovers a reverse Pinsker inequality due to Verdu [19].

Corollary II.6 (Verdu [19] Theorem 7). *For probability measures μ and γ such that $\frac{d\mu}{d\gamma} \leq \frac{1}{\beta}$ with $\beta \in (0, 1)$*

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

III. CONSEQUENCES

Given a sequence of densities $\{f_i\}$ and weights $\{p_i\}$, define $\tilde{f}_i$ to be the density constructed with f_i removed, explicitly,

$$\tilde{f}_i = \frac{\sum_{j \neq i} p_j f_j}{(1 - p_i)}.$$

Furthermore define

$$\mathcal{T} = \mathcal{T}_{f,p} = \sup_i \|f_i - \tilde{f}_i\|_{TV}$$

Proposition III.1. *The concavity deficit satisfies the following*

$$\Phi_{f,p} = \sum_i p_i D(f_i\|f) = \sum_i p_i S_{p_i}(f_i\|\tilde{f}_i). \quad (26)$$

Proof. Recalling our convention $f = \sum_i p_i f_i$, the expression of Φ is a direct computation

$$\begin{aligned} \Phi &= - \int \left(\sum_i p_i f_i \right) \ln f + \sum_i p_i \int f_i \ln f_i \\ &= \sum_i p_i \int f_i \ln f_i / f \\ &= \sum_i p_i D(f_i\|f). \end{aligned}$$

This gives the first equality, the second equality follows from Proposition II.2.1. □

Theorem III.1. *For a sequence of densities $f = \{f_i\}$, and weights $p = \{p_i\}$ the concavity deficit satisfies the following*

$$\Phi(f, p) \leq \mathcal{T}_{f,p} H(p).$$

Proof. Combining Proposition III.1 and Theorem II.2 we have,

$$\Phi(f, p) \leq - \sum_i p_i \log p_i \|f_i - \tilde{f}_i\|_{TV}$$

By the definition of $\mathcal{T}_{f,p}$ as $\sup_i \|f_i - \tilde{f}_i\|_{TV}$ the result follows. □

Since the total variation of any two measures is bounded above by 1 this gives a sharpening of (2),

$$h(X + Z) \leq \mathcal{T}H(X) + h(Z|X).$$

Taking the case that $n = 2$ and $p_1 = p_2 = \frac{1}{2}$ in

$$\Phi = h\left(\sum_{i=1}^n p_i f_i\right) - \sum_{i=1}^n p_i h(f_i)$$

gives the Jensen-Shannon Divergence.

Definition III.2. For probability measures μ and ν define the Jensen-Shannon divergence,

$$JSD(\mu||\nu) = \frac{1}{2} (D(\mu||2^{-1}(\mu + \nu)) + D(\nu||2^{-1}(\mu + \nu))). \quad (27)$$

With this definition we have an immediate corollary.

Corollary III.3 (Lin [8], Topsøe [18]). For μ and ν probability measures,

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

Corollary III.4. When $f = \{f_0, f_1\}$, where $f_0(x) = g(x + \mu)$, $f_1(x) = g(x - \mu)$ for g unimodal and symmetric and $p = \{1/2, 1/2\}$. Then

$$h(f) \leq h(g) + (1 - 2\mathcal{E}) \ln 2$$

Where

$$\mathcal{E} = \int_{\mu}^{\infty} g(x) dx$$

Let us note the following interpretation of $\mathcal{E}$. First, consider a bit $B_1 = \mathbb{1}_{(0, \infty)}(X)$ where X taking values $\pm\mu$ with equal probability. Second take, $B_2 = \mathbb{1}_{(0, \infty)}(X + Z)$ where X is as previous, and Z is an independent noise distributed according to g . Then

$$\mathcal{E} = \mathbb{P}(B_1 \neq B_2).$$

Indeed, using independence and then symmetry

$$\begin{aligned} \mathbb{P}(B_1 \neq B_2) &= \mathbb{P}(X = \mu; X + Z \leq 0) \\ &\quad + \mathbb{P}(X = -\mu; X + Z > 0) \\ &= \frac{1}{2} \mathbb{P}(Z \leq -\mu) + \frac{1}{2} \mathbb{P}(Z > \mu) \\ &= \int_{\mu}^{\infty} g(x) dx. \end{aligned}$$

Thus $\mathcal{E}$ can be understood as the probability of error in communicating X when subject to an additive independent noise Z . In this setting the statement of Corollary III.4 can be formulated as

$$I(X; X + Z) = h(X + Z) - h(Z + X|X) \quad (28)$$

$$= h(f) - h(g) \quad (29)$$

$$\leq (1 - 2\mathcal{E}) \ln 2 \quad (30)$$

Alternatively, if θ is a uniform prior on $\{0, 1\}$, and $\ell(x, y) = 1 - \delta_x(y)$ the risk function, then it can be shown that $\mathcal{E} = R_{\theta}$ is the associated Bayes risk. This gives the following upper bound,

$$R_{\theta} \leq \frac{h(g) - h(f)}{2 \log 2} - \frac{1}{2}. \quad (31)$$

Proof of Corollary III.4. Note that $\tilde{f}_1 = f_0$ and $\tilde{f}_0 = f_1$ so that

$$\mathcal{T} = \max \|f_i - \tilde{f}_i\|_{TV} = \|f_1 - f_0\|_{TV}.$$

Since g is a unimodal and symmetric probability density,

$$\mathcal{T} = \int (f_1 - f_0)_+ \quad (32)$$

$$= \int_0^{\infty} g(x - \mu) - g(x + \mu) dx \quad (33)$$

$$= \int_{-\mu}^{\infty} g - \int_{\mu}^{\infty} g \quad (34)$$

$$= 1 - 2 \int_{\mu}^{\infty} g. \quad (35)$$

□

IV. CONCLUSION

The investigation of the concavity of entropy is facilitated by a careful study $D(f||g)$ where g is a density that can be decomposed into a convex combination of f with another density. Such studies are equivalent to, and motivate the manipulation of “skew” divergences. Several new relationships between skew divergences are established. Their improved behavior, is exemplified in the reverse Pinsker type inequality that the skew divergence is proven to satisfy. In turn, the reversal of Pinsker’s inequality is leveraged to arrive at a sharpening of the general concavity upper bound for entropy. An expanded version of this paper, with further development of these ideas and the complementary lower bounds for the Holevo information initiated in [10] can be found in [9].

REFERENCES

- [1] K.M.R. Audenaert. Quantum skew divergence. *Journal of Mathematical Physics*, 55(11):112202, 2014.
- [2] S. G. Bobkov and A. Marsiglietti. Entropic CLT for smoothed convolutions and associated entropy bounds. *arXiv preprint arXiv:1903.03666*, 2019.
- [3] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. J. Wiley, New York, 1991.
- [4] N. Cressie and T.R.C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):440–464, 1984.
- [5] A.S. Holevo. Bounds for the quantity of information transmitted by a quantum communication channel. *Problemy Peredachi Informatsii*, 9(3):3–11, 1973.
- [6] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999.

- [7] F. Liese and I. Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.
- [8] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.
- [9] J. Melbourne, M. Madiman, S. Talukdar, S. Bhaban, and M.V. Salapaka. The concavity deficit for entropy. *arXiv preprint arXiv:1805.11257*, 2018.
- [10] J. Melbourne, S. Talukdar, S. Bhaban, and M. Salapaka. Error bounds for a mixed entropy inequality. In *Information Theory (ISIT), 2018 IEEE International Symposium on*, 2018.
- [11] J. Melbourne, S. Talukdar, and M.V. Salapaka. Realizing information erasure in finite time. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 4135–4140. IEEE, 2018.
- [12] J. Neyman. Contribution to the theory of the χ^2 test. In *Proceedings of the Berkeley symposium on mathematical statistics and probability*, volume 1, pages 239–273. University of California Press Berkeley, 1949.
- [13] 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. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
- [14] M.S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Fransisco, 1964.
- [15] I. Sason. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, 2018.
- [16] I. Sason and S. Verdu. *f*-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.
- [17] S. Talukdar, S. Bhaban, J. Melbourne, and M. Salapaka. Analysis of heat dissipation and reliability in information erasure: A gaussian mixture approach. *Entropy*, 20(10):749, 2018.
- [18] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on information theory*, 46(4):1602–1609, 2000.
- [19] S. Verdú. Total variation distance and the distribution of relative information. In *2014 Information Theory and Applications Workshop (ITA)*, pages 1–3. IEEE, 2014.
- [20] L. Wang and M. Madiman. Beyond the entropy power inequality, via rearrangements. *IEEE Transactions on Information Theory*, 60(9):5116–5137, 2014.