entropy

MDPI

Article

A Two-Moment Inequality with Applications to Rényi Entropy and Mutual Information

Galen Reeves [1,2]

[1] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA; galen.reeves@duke.edu

[2] Department of Statistical Science, Duke University, Durham, NC 27708, USA

Received: 14 September 2020; Accepted: 6 October 2020; Published: 1 November 2020

Abstract: This paper explores some applications of a two-moment inequality for the integral of the rth power of a function, where $0 < r < 1$. The first contribution is an upper bound on the Rényi entropy of a random vector in terms of the two different moments. When one of the moments is the zeroth moment, these bounds recover previous results based on maximum entropy distributions under a single moment constraint. More generally, evaluation of the bound with two carefully chosen nonzero moments can lead to significant improvements with a modest increase in complexity. The second contribution is a method for upper bounding mutual information in terms of certain integrals with respect to the variance of the conditional density. The bounds have a number of useful properties arising from the connection with variance decompositions.

Keywords: information inequalities; mutual information; Rényi entropy; Carlson–Levin inequality

1. Introduction

The interplay between inequalities and information theory has a rich history, with notable examples including the relationship between the Brunn–Minkowski inequality and the entropy power inequality as well as the matrix determinant inequalities obtained from differential entropy [1]. In this paper, the focus is on a "two-moment" inequality that provides an upper bound on the integral of the rth power of a function. Specifically, if f is a nonnegative function defined on $\mathbb{R}^n$ and p, q, r are real numbers satisfying $0 < r < 1$ and $p < 1/r - 1 < q$, then

$$\left(\int f(x)^r \, \mathrm{d}x \right)^{\frac{1}{r}} \leq C_{n,p,q,r} \left(\int \|x\|^{np} f(x) \, \mathrm{d}x \right)^{\frac{qr+r-1}{(q-p)r}} \left(\int \|x\|^{nq} f(x) \, \mathrm{d}x \right)^{\frac{1-r-pr}{(q-p)r}}, \tag{1}$$

where the best possible constant $C_{n,p,q,r}$ is given exactly; see Propositions 2 and 3 ahead. The one-dimensional version of this inequality is a special case of the classical Carlson–Levin inequality [2–4], and the multidimensional version is a special case of a result presented by Barza et al. [5]. The particular formulation of the inequality used in this paper was derived independently in [6], where the proof follows from a direct application of Hölder's inequality and Jensen's inequality.

In the context of information theory and statistics, a useful property of the two-moment inequality is that it provides a bound on a nonlinear functional, namely the r-quasi-norm $\| \cdot \|_r$, in terms of integrals that are linear in f. Consequently, this inequality is well suited to settings where f is a mixture of simple functions whose moments can be evaluated. We note that this reliance on moments to bound a nonlinear functional is closely related to bounds obtained from variational characterizations such as the Donsker–Varadhan representation of Kullback divergence [7] and its generalizations to Rényi divergence [8,9].

The first application considered in this paper concerns the relationship between the entropy of a probability measure and its moments. This relationship is fundamental to the principle of maximum entropy, which originated in statistical physics and has since been applied to statistical inference problems [10]. It also plays a prominent role in information theory and estimation theory where the fact that the Gaussian distribution maximizes differential entropy under second moment constraints ([11], [Theorem 8.6.5]) plays a prominent role. Moment–entropy inequalities for Rényi entropy were studied in a series of works by Lutwak et al. [12–14], as well as related works by Costa et al. [15,16] and Johonson and Vignat [17], in which it is shown that, under a single moment constraint, Rényi entropy is maximized by a family of generalized Gaussian distributions. The connection between these moment–entropy inequalities and the Carlson–Levin inequality was noted recently by Nguyen [18].

In this direction, one of the contributions of this paper is a new family of moment–entropy inequalities. This family of inequalities follows from applying Inequality (1) in the setting where f is a probability density function, and thus there is a one-to-one correspondence between the integral of the rth power and the Rényi entropy of order r. In the special case where one of the moments is the zeroth moment, this approach recovers the moment–entropy inequalities given in previous work. More generally, the additional flexibility provided by considering two different moments can lead to stronger results. For example, in Proposition 6, it is shown that if f is the standard Gaussian density function defined on $\mathbb{R}^n$, then the difference between the Rényi entropy and the upper bound given by the two-moment inequality (equivalently, the ratio between the left- and right-hand sides of (1)) is bounded uniformly with respect to n under the following specification of the moments:

$$p_n = \frac{1-r}{r} - \frac{1}{r}\sqrt{\frac{2(1-r)}{n+1}}, \qquad q_n = \frac{1-r}{r} + \frac{1}{r}\sqrt{\frac{2(1-r)}{n+1}}. \tag{2}$$

Conversely, if one of the moments is restricted to be equal to zero, as is the case in the usual moment–entropy inequalities, then the difference between the Rényi entropy and the upper bound diverges with n.

The second application considered in this paper is the problem of bounding mutual information. In conjunction with Fano's inequality and its extensions, bounds on mutual information play a prominent role in establishing minimax rates of statistical estimation [19] as well as the information-theoretic limits of detection in high-dimensional settings [20]. In many cases, one of the technical challenges is to provide conditions under which the dependence between the observations and an underlying signal or model parameters converges to zero in the limit of high dimension.

This paper introduces a new method for bounding mutual information, which can be described as follows. Let $P_{X,Y}$ be a probability measure on $\mathcal{X} \times \mathcal{Y}$ such that $P_{Y|X=x}$ and P_Y have densities $f(y \mid x)$ and $f(y)$ with respect to the Lebesgue measure on $\mathbb{R}^n$. We begin by showing that the mutual information between X and Y satisfies the upper bound

$$I(X;Y) \le \int \sqrt{\mathsf{Var}(f(y \mid X))}\, \mathrm{d}y, \tag{3}$$

where $\mathsf{Var}(p(y \mid X)) = \int (f(y \mid x) - f(y))^2 \, \mathrm{d}P_X(x)$ is the variance of $f(y \mid X)$; see Proposition 8 ahead. In view of (3), an application of the two-moment Inequality (1) with $r = 1/2$ leads to an upper bound with respect to the moments of the variance of the density:

$$\int \|y\|^{ns} \, \mathsf{Var}(f(y \mid X))\, \mathrm{d}y \tag{4}$$

where this expression is evaluated at $s \in \{p, q\}$ with $p < 1 < q$. A useful property of this bound is that the integrated variance is quadratic in P_X, and thus Expression (4) can be evaluated by swapping the integration over y and with the expectation of over two independent copies of X. For example, when $P_{X,Y}$ is a Gaussian scale mixture, this approach provides closed-form upper bounds in terms of

the moments of the Gaussian density. An early version of this technique is used to prove Gaussian approximations for random projections [21] arising in the analysis of a random linear estimation problem appearing in wireless communications and compressed sensing [22,23].

2. Moment Inequalities

Let $L^p(S)$ be the space of Lebesgue measurable functions from S to $\mathbb{R}$ whose pth power is absolutely integrable, and for $p \neq 0$, define

$$\|f\|_p := \left(\int_S |f(x)|^p \, \mathrm{d}x \right)^{1/p}.$$

Recall that $\|\cdot\|_p$ is a norm for $p \geq 1$ but only a quasi-norm for $0 < p < 1$ because it does not satisfy the triangle inequality. The sth moment of f is defined as

$$\mathcal{M}_s(f) := \int_S \|x\|^s \, |f(x)| \, \mathrm{d}x,$$

where $\|\cdot\|$ denotes the standard Euclidean norm on vectors.

The two-moment Inequality (1) can be derived straightforwardly using the following argument. For $r \in (0,1)$, the mapping $f \mapsto \|f\|_r$ is concave on the subset of nonnegative functions and admits the variational representation

$$\|f\|_r = \inf \left\{ \frac{\|fg\|_1}{\|g\|_{r^*}} \, : \, g \in L^{r^*} \right\}, \tag{5}$$

where $r^* = r/(r-1) \in (-\infty, 0)$ is the Hölder conjugate of r. Consequently, each $g \in L^{r^*}$ leads to an upper bound on $\|f\|_r$. For example, if f has bounded support S, choosing g to be the indicator function of S leads to the basic inequality $\|f\|_r \leq (\mathrm{Vol}(S))^{(1-r)/r} \|f\|_1$. The upper bound on $\|f\|_r$ given in Inequality (1) can be obtained by restricting the minimum in Expression (5) to the parametric class of functions of the form $g(x) = \nu_1 \|x\|^{np} + \nu_2 \|x\|^{nq}$ with $\nu_1, \nu_2 > 0$ and then optimizing over the parameters (ν_1, ν_2). Here, the constraints on p, q are necessary and sufficient to ensure that $g \in L^{r^*}(\mathbb{R}^n)$.

In the following sections, we provide a more detailed derivation, starting with the problem of maximizing $\|f\|_r$ under multiple moment constraints and then specializing to the case of two moments. For a detailed account of the history of the Carlson type inequalities as well as some further extensions, see [4].

2.1. Multiple Moments

Consider the following optimization problem:

$$\begin{array}{lll} \text{maximize} & \|f\|_r & \\ \text{subject to} & f(x) \geq 0 & \text{for all } x \in S \\ & \mathcal{M}_{s_i}(f) \leq m_i & \text{for } 1 \leq i \leq k. \end{array}$$

For $r \in (0,1)$, this is a convex optimization problem because $\|\cdot\|_r^r$ is concave and the moment constraints are linear. By standard theory in convex optimization (e.g., [24]), it can be shown that if the problem is feasible and the maximum is finite, then the maximizer has the form

$$f^*(x) = \left(\sum_{i=1}^k \nu_i^* \, \|x\|^{s_i} \right)^{\frac{1}{r-1}}, \quad \text{for all } x \in S.$$

The parameters $\nu_1^*, \cdots, \nu_k^*$ are nonnegative and the ith moment constraint holds with equality for all i such that ν_i^* is strictly positive—that is, $\nu_i^* > 0 \implies \mu_{s_i}(f^*) = m_i$. Consequently, the maximum can be expressed in terms of a linear combination of the moments:

$$\|f^*\|_r^r = \|(f^*)^r\|_1 = \|f^*(f^*)^{r-1}\|_1 = \sum_{i=1}^k \nu_i^* m_i.$$

For the purposes of this paper, it is useful to consider a relative inequality in terms of the moments of the function itself. Given a number $0 < r < 1$ and vectors $s \in \mathbb{R}^k$ and $\nu \in \mathbb{R}_+^k$, the function $c_r(\nu, s)$ is defined according to

$$c_r(\nu, s) = \left(\int_0^\infty \left(\sum_{i=1}^k \nu_i \, x^{s_i} \right)^{-\frac{r}{1-r}} \mathrm{d}x \right)^{\frac{1-r}{r}},$$

if the integral exists. Otherwise, $c_r(\nu, s)$ is defined to be positive infinity. It can be verified that $c_r(\nu, s)$ is finite provided that there exists i, j such that ν_i and ν_j are strictly positive and $s_i < (1-r)/r < s_j$.

The following result can be viewed as a consequence of the constrained optimization problem described above. We provide a different and very simple proof that depends only on Hölder's inequality.

Proposition 1. *Let f be a nonnegative Lebesgue measurable function defined on the positive reals $\mathbb{R}_+$. For any number $0 < r < 1$ and vectors $s \in \mathbb{R}^k$ and $\nu \in \mathbb{R}_+^k$, we have*

$$\|f\|_r \le c_r(\nu, s) \sum_{i=1}^k \nu_i \, \mathcal{M}_{s_i}(f).$$

Proof. Let $g(x) = \sum_{i=1}^k \nu_i \, x^{s_i}$. Then, we have

$$\|f\|_r^r = \|g^{-r}(fg)^r\|_1 \le \|g^{-r}\|_{\frac{1}{1-r}} \|(gf)^r\|_{\frac{1}{r}} = \|g^{\frac{-r}{1-r}}\|_1^{1-r} \|gf\|_1^r = \left(c_r(\nu, s) \sum_{i=1}^k \nu_i \, \mathcal{M}_{s_i}(f) \right)^r,$$

where the second step is Hölder's inequality with conjugate exponents $1/(1-r)$ and $1/r$. □

2.2. Two Moments

For $a, b > 0$, the beta function $\mathrm{B}(a, b)$ and gamma function $\Gamma(a)$ are given by

$$\mathrm{B}(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1} \, \mathrm{d}t$$
$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \, dt,$$

and satisfy the relation $\mathrm{B}(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, $a, b > 0$. To lighten the notation, we define the normalized beta function

$$\widetilde{B}(a, b) = \mathrm{B}(a, b)(a+b)^{a+b} a^{-a} b^{-b}. \tag{6}$$

Properties of these functions are provided in Appendix A.

The next result follows from Proposition 1 for the case of two moments.

Proposition 2. *Let f be a nonnegative Lebesgue measurable function defined on $[0, \infty)$. For any numbers p, q, r with $0 < r < 1$ and $p < 1/r - 1 < q$,*

$$\|f\|_r \le [\psi_r(p, q)]^{\frac{1-r}{r}} \, [\mathcal{M}_p(f)]^\lambda [\mathcal{M}_q(f)]^{1-\lambda},$$

where $\lambda = (q + 1 - 1/r)/(q - p)$ *and*

$$\psi_r(p,q) = \frac{1}{(q-p)} \widetilde{\mathrm{B}}\left(\frac{r\lambda}{1-r}, \frac{r(1-\lambda)}{1-r}\right), \tag{7}$$

where $\widetilde{B}(\cdot,\cdot)$ *is defined in Equation* (6).

Proof. Letting $s = (p,q)$ and $\nu = (\gamma^{1-\lambda}, \gamma^{-\lambda})$ with $\lambda > 0$, we have

$$[c_r(\nu,s)]^{\frac{r}{1-r}} = \int_0^\infty \left(\gamma^{1-\lambda}\, x^p + \gamma^{-\lambda}\, x^q\right)^{-\frac{r}{1-r}} \mathrm{d}x.$$

Making the change of variable $x \mapsto (\gamma u)^{\frac{1}{q-p}}$ leads to

$$[c_r(\nu,s)]^{\frac{r}{1-r}} = \frac{1}{q-p} \int_0^\infty \frac{u^{b-1}}{(1+u)^{a+b}}\, \mathrm{d}u = \frac{\mathrm{B}\,(a,b)}{q-p},$$

where $a = \frac{r}{1-r}\lambda$ and $b = \frac{r}{1-r}(1-\lambda)$ and the second step follows from recognizing the integral representation of the beta function given in Equation (A3). Therefore, by Proposition 1, the inequality

$$\|f\|_r \leq \left(\frac{\mathrm{B}\,(a,b)}{q-p}\right)^{\frac{1-r}{r}} \left(\gamma^{1-\lambda}\, \mathcal{M}_p(f) + \gamma^{-\lambda}\, \mathcal{M}_q(f)\right),$$

holds for all $\gamma > 0$. Evaluating this inequality with

$$\gamma = \frac{\lambda\, \mathcal{M}_q(f)}{(1-\lambda)\mathcal{M}_p(f)},$$

leads to the stated result. □

The special case $r = 1/2$ admits the simplified expression

$$\psi_{1/2}(p,q) = \frac{\pi \lambda^{-\lambda}(1-\lambda)^{-(1-\lambda)}}{(q-p)\sin(\pi\lambda)}, \tag{8}$$

where we have used Euler's reflection formula for the beta function ([25], [Theorem 1.2.1]).

Next, we consider an extension of Proposition 2 for functions defined on $\mathbb{R}^n$. Given any measurable subset S of $\mathbb{R}^n$, we define

$$\omega(S) = \mathrm{Vol}(B^n \cap \mathrm{cone}(S)), \tag{9}$$

where $B^n = \{u \in \mathbb{R}^n : \|u\| \leq 1\}$ is the n-dimensional Euclidean ball of radius one and

$$\mathrm{cone}(S) = \{x \in \mathbb{R}^n \,:\, tx \in S \text{ for some } t > 0\}.$$

The function $\omega(S)$ is proportional to the surface measure of the projection of S on the Euclidean sphere and satisfies

$$\omega(S) \leq \omega(\mathbb{R}^n) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)}, \tag{10}$$

for all $S \subseteq \mathbb{R}^n$. Note that $\omega(\mathbb{R}_+) = 1$ and $\omega(\mathbb{R}) = 2$.

Proposition 3. *Let f be a nonnegative Lebesgue measurable function defined on a subset S of $\mathbb{R}^n$. For any numbers p, q, r with $0 < r < 1$ and $p < 1/r - 1 < q$,*

$$\|f\|_r \leq [\omega(S)\, \psi_r(p,q)]^{\frac{1-r}{r}}\ [\mathcal{M}_{np}(f)]^{\lambda} [\mathcal{M}_{nq}(f)]^{1-\lambda},$$

where $\lambda = (q + 1 - 1/r)/(q - p)$ and $\psi_r(p,q)$ is given by Equation (7).

Proof. Let f be extended to $\mathbb{R}^n$ using the rule $f(x) = 0$ for all x outside of S and let $g\colon \mathbb{R}_+ \to \mathbb{R}_+$ be defined according to

$$g(y) = \frac{1}{n} \int_{\mathbb{S}^{n-1}} f(y^{1/n} u)\, \mathrm{d}\sigma(u),$$

where $\mathbb{S}^{n-1} = \{u \in \mathbb{R}^n : \|u\| = 1\}$ is the Euclidean sphere of radius one and $\sigma(u)$ is the surface measure of the sphere. In the following, we will show that

$$\|f\|_r \leq (\omega(S))^{\frac{1-r}{r}}\ \|g\|_r \tag{11}$$

$$\mathcal{M}_{ns}(f) = \mathcal{M}_s(g). \tag{12}$$

Then, the stated inequality then follows from applying Proposition 2 to the function g.

To prove Inequality (11), we begin with a transformation into polar coordinates:

$$\|f\|_r^r = \int_0^\infty \int_{\mathbb{S}^{n-1}} |f(tu)|^r\, t^{n-1}\, \mathrm{d}\sigma(u)\, \mathrm{d}t. \tag{13}$$

Letting $\mathbf{1}_{\mathrm{cone}(S)}(x)$ denote the indicator function of the set $\mathrm{cone}(S)$, the integral over the sphere can be bounded using:

$$\begin{aligned}
\int_{\mathbb{S}^{n-1}} |f(tu)|^r\, \mathrm{d}\sigma(u) &= \int_{\mathbb{S}^{n-1}} \mathbf{1}_{\mathrm{cone}(S)}(u)\ |f(tu)|^r\, \mathrm{d}\sigma(u) \\
&\overset{(a)}{\leq} \left(\int_{\mathbb{S}^{n-1}} \mathbf{1}_{\mathrm{cone}(S)}(u)\, \mathrm{d}\sigma(u) \right)^{1-r} \left(\int_{\mathbb{S}^{n-1}} |f(tu)|\, \mathrm{d}\sigma(u) \right)^{r} \\
&\overset{(b)}{=} n\ (\omega(S))^{1-r}\, g^r(t^n),
\end{aligned} \tag{14}$$

where: (a) follows from Hölder's inequality with conjugate exponents $\frac{1}{1-r}$ and $\frac{1}{r}$, and (b) follows from the definition of g and the fact that

$$\begin{aligned}
\omega(S) &= \int_0^1 \int_{\mathbb{S}^{n-1}} \mathbf{1}_{\mathrm{cone}(S)}(u)\, t^{n-1}\, \mathrm{d}\sigma(u)\, \mathrm{d}t \\
&= \frac{1}{n} \int_{\mathbb{S}^{n-1}} \mathbf{1}_{\mathrm{cone}(S)}(u)\, \mathrm{d}\sigma(u).
\end{aligned}$$

Plugging Inequality (14) back into Equation (13) and then making the change of variable $t \to y^{\frac{1}{n}}$ yields

$$\|f\|_r^r \leq n\ (\omega(S))^{1-r} \int_0^\infty g^r(t^n) t^{n-1}\, \mathrm{d}t = (\omega(S))^{1-r}\, \|g\|_r^r.$$

The proof of Equation (12) follows along similar lines. We have

$$\begin{aligned}
\mathcal{M}_{ns}(f) &\overset{(a)}{=} \int_0^\infty \int_{\mathbb{S}^{n-1}} t^{ns} f(tu)\, t^{n-1}\, \mathrm{d}\sigma(u)\, \mathrm{d}t \\
&\overset{(b)}{=} \frac{1}{n} \int_0^\infty \int_{\mathbb{S}^{n-1}} y^s f(y^{\frac{1}{n}} u)\, \mathrm{d}\sigma(u)\, \mathrm{d}y \\
&= \mathcal{M}_s(g)
\end{aligned}$$

where (a) follows from a transformation into polar coordinates and (b) follows from the change of variable $t \mapsto y^{\frac{1}{n}}$.

Having established Inequality (11) and Equation (12), an application of Proposition 2 completes the proof. □

3. Rényi Entropy Bounds

Let X be a random vector that has a density $f(x)$ with respect to the Lebesgue measure on $\mathbb{R}^n$. The differential Rényi entropy of order $r \in (0,1) \cup (1,\infty)$ is defined according to [11]:

$$h_r(X) = \frac{1}{1-r} \log \left(\int_{\mathbb{R}^n} f^r(x)\, \mathrm{d}x \right).$$

Throughout this paper, it is assumed that the logarithm is defined with respect to the natural base and entropy is measured in nats. The Rényi entropy is continuous and nonincreasing in r. If the support set $S = \{x \in \mathbb{R}^n : f(x) > 0\}$ has finite measure, then the limit as r converges to zero is given by $h_0(X) = \log \mathrm{Vol}(S)$. If the support does not have finite measure, then $h_r(X)$ increases to infinity as r decreases to zero. The case $r = 1$ is given by the Shannon differential entropy:

$$h_1(X) = -\int_S f(x) \log f(x)\, \mathrm{d}x.$$

Given a random variable X that is not identical to zero and numbers p, q, r with $0 < r < 1$ and $p < 1/r - 1 < q$, we define the function

$$L_r(X; p, q) = \frac{r\lambda}{1-r} \log \mathbb{E}\left[|X|^p\right] + \frac{r(1-\lambda)}{1-r} \log \mathbb{E}\left[|X|^q\right],$$

where $\lambda = (q + 1 - 1/r)/(q - p)$.

The next result, which follows directly from Proposition 3, provides an upper bound on the Rényi entropy.

Proposition 4. *Let X be a random vector with a density on $\mathbb{R}^n$. For any numbers p, q, r with $0 < r < 1$ and $p < 1/r - 1 < q$, the Rényi entropy satisfies*

$$h_r(X) \le \log \omega(S) + \log \psi_r(p,q) + L_r(\|X\|^n; p, q), \tag{15}$$

where $\omega(S)$ is defined in Equation (9) and $\psi_r(p,q)$ is defined in Equation (7).

Proof. This result follows immediately from Proposition 3 and the definition of Rényi entropy. □

The relationship between Proposition 4 and previous results depends on whether the moment p is equal to zero:

- *One-moment inequalities:* If $p = 0$, then there exists a distribution such that Inequality (15) holds with equality. This is because the zero-moment constraint ensures that the function that maximizes the Rényi entropy integrates to one. In this case, Proposition 4 is equivalent to previous results that focused on distributions that maximize Rényi entropy subject to a single moment constraint [12,13,15]. With some abuse of terminology, we refer to these bounds as one-moment inequalities. (A more accurate name would be two-moment inequalities under the constraint that one of the moments is the zeroth moment.)
- *Two-moment inequalities:* If $p \neq 0$, then the right-hand side of Inequality (15) corresponds to the Rényi entropy of a nonnegative function that might not integrate to one. Nevertheless, the expression provides an upper bound on the Rényi entropy for any density with the same moments. We refer to the bounds obtained using a general pair (p, q) as two-moment inequalities.

The contribution of two-moment inequalities is that they lead to tighter bounds. To quantify the tightness, we define $\Delta_r(X; p, q)$ to be the gap between the right-hand side and left-hand side of Inequality (15) corresponding to the pair (p, q)—that is,

$$\Delta_r(X; p, q) = \log \omega(S) + \log \psi_r(p, q) + L_r(\|X\|^n; p, q) - h_r(X).$$

The gaps corresponding to the optimal two-moment and one-moment inequalities are defined according to

$$\Delta_r(X) = \inf_{p,q} \Delta_r(X; p, q)$$
$$\widetilde{\Delta}_r(X) = \inf_{q} \Delta_r(X; 0, q).$$

3.1. Some Consequences of These Bounds

By Lyapunov's inequality, the mapping $s \mapsto \frac{1}{s} \log \mathbb{E}\left[|X|^s\right]$ is nondecreasing on $[0, \infty)$, and thus

$$L_r(X; p, q) \leq L_r(X; 0, q) = \frac{1}{q} \log \mathbb{E}\left[|X|^q\right], \quad p \geq 0. \tag{16}$$

In other words, the case $p = 0$ provides an upper bound on $L_r(X; p, q)$ for nonnegative p. Alternatively, we also have the lower bound

$$L_r(X; p, q) \geq \frac{r}{1-r} \log \mathbb{E}\left[|X|^{\frac{1-r}{r}}\right], \tag{17}$$

which follows from the convexity of $\log \mathbb{E}\left[|X|^s\right]$.

A useful property of $L_r(X; p, q)$ is that it is additive with respect to the product of independent random variables. Specifically, if X and Y are independent, then

$$L_r(XY; p, q) = L_r(X; p, q) + L_r(Y; p, q). \tag{18}$$

One consequence is that multiplication by a bounded random variable cannot increase the Rényi entropy by an amount that exceeds the gap of the two-moment inequality with nonnegative moments.

Proposition 5. *Let Y be a random vector on $\mathbb{R}^n$ with finite Rényi entropy of order $0 < r < 1$, and let X be an independent random variable that satisfies $0 < X \leq t$. Then,*

$$h_r(XY) \leq h_r(tY) + \Delta_r(Y; p, q),$$

for all $0 < p < 1/r - 1 < q$.

Proof. Let $Z = XY$ and let S_Z and S_Y denote the support sets of Z and Y, respectively. The assumption that X is nonnegative means that $\mathrm{cone}(S_Z) = \mathrm{cone}(S_Y)$. We have

$$\begin{aligned} h_r(Z) &\overset{(a)}{\leq} \log \omega(S_Z) + \log \psi_r(p, q) + L_r(\|Z\|^n; p, q) \\ &\overset{(b)}{=} h_r(Y) + L_r(|X|^n; p; q) + \Delta_r(Y; p, q) \\ &\overset{(c)}{\leq} h_r(Y) + n \log t + \Delta_r(Y; p, q), \end{aligned}$$

where (a) follows from Proposition 4, (b) follows from Equation (18) and the definition of $\Delta_r(Y; p, q)$, and (c) follows from Inequality (16) and the assumption $|X| \leq t$. Finally, recalling that $h_r(tY) = h_r(Y) + n \log t$ completes the proof. □

3.2. Example with Log-Normal Distribution

If $W \sim \mathcal{N}(\mu, \sigma^2)$, then the random variable $X = \exp(W)$ has a log-normal distribution with parameters (μ, σ^2). The Rényi entropy is given by

$$h_r(X) = \mu + \frac{1}{2}\left(\frac{1-r}{r}\right)\sigma^2 + \frac{1}{2}\log(2\pi r^{\frac{1}{r-1}}\sigma^2),$$

and the logarithm of the sth moment is given by

$$\log \mathbb{E}\left[|X|^s\right] = \mu s + \frac{1}{2}\sigma^2 s^2.$$

With a bit of work, it can be shown that the gap of the optimal two-moment inequality does not depend on the parameters (μ, σ^2) and is given by

$$\Delta_r(X) = \log\left(\widetilde{\mathrm{B}}\left(\frac{r}{2(1-r)}, \frac{r}{2(1-r)}\right)\sqrt{\frac{r}{4(1-r)}}\right) + \frac{1}{2} - \frac{1}{2}\log(2\pi r^{\frac{1}{r-1}}). \tag{19}$$

The details of this derivation are given in Appendix B.1. Meanwhile, the gap of the optimal one-moment inequality is given by

$$\widetilde{\Delta}_r(X) = \inf_q \left[\log\left(\widetilde{\mathrm{B}}\left(\frac{r}{1-r} - \frac{1}{q}, \frac{1}{q}\right)\frac{1}{q}\right) + \frac{1}{2}q\sigma^2\right] - \frac{1}{2}\left(\frac{1-r}{r}\right)\sigma^2 - \frac{1}{2}\log(2\pi r^{\frac{1}{r-1}}\sigma^2). \tag{20}$$

The functions $\Delta_r(X)$ and $\widetilde{\Delta}_r(X)$ are illustrated in Figure 1 as a function of r for various σ^2. The function $\Delta_r(X)$ is bounded uniformly with respect to r and converges to zero as r increases to one. The tightness of the two-moment inequality in this regime follows from the fact that the log-normal distribution maximizes Shannon entropy subject to a constraint on $\mathbb{E}[\log X]$. By contrast, the function $\widetilde{\Delta}_r(X)$ varies with the parameter σ^2. For any fixed $r \in (0,1)$, it can be shown that $\widetilde{\Delta}_r(X)$ increases to infinity if σ^2 converges to zero or infinity.

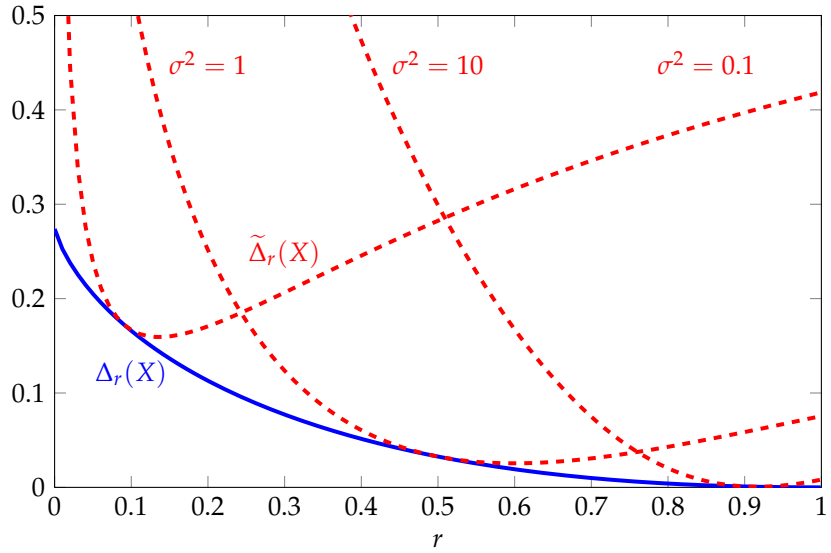

Figure 1. Comparison of upper bounds on Rényi entropy in nats for the log-normal distribution as a function of the order r for various σ^2.

3.3. Example with Multivariate Gaussian Distribution

Next, we consider the case where $Y \sim \mathcal{N}(0, I_n)$ is an n-dimensional Gaussian vector with mean zero and identity covariance. The Rényi entropy is given by

$$h_r(Y) = \frac{n}{2} \log(2\pi r^{\frac{1}{r-1}}),$$

and the sth moment of the magnitude $\|Y\|$ is given by

$$\mathbb{E}\left[\|Y\|^s\right] = \frac{2^{\frac{s}{2}} \Gamma(\frac{n+s}{2})}{\Gamma(\frac{n}{2})}.$$

The next result shows that as the dimension n increases, the gap of the optimal two-moment inequality converges to the gap for the log-normal distribution. Moreover, for each $r \in (0,1)$, the following choice of moments is optimal in the large-n limit:

$$p_n = \frac{1-r}{r} - \frac{1}{r}\sqrt{\frac{2(1-r)}{n+1}}, \qquad q_n = \frac{1-r}{r} + \frac{1}{r}\sqrt{\frac{2(1-r)}{n+1}}. \tag{21}$$

The proof is given in Appendix B.3.

Proposition 6. *If $Y \sim \mathcal{N}(0, I_n)$, then, for each $r \in (0,1)$,*

$$\lim_{n\to\infty} \Delta_r(Y) = \lim_{n\to\infty} \Delta_r(Y; p_n, q_n) = \Delta_r(X),$$

where X has a log-normal distribution and (p_n, q_n) are given by (21).

Figure 2 provides a comparison of $\Delta_r(Y)$, $\Delta_r(Y; p_n, q_n)$, and $\widetilde{\Delta}_r(Y)$ as a function of n for $r = 0.1$. Here, we see that both $\Delta_r(Y)$ and $\Delta_r(Y; p_n, q_n)$ converge rapidly to the asymptotic limit given by the gap of the log-normal distribution. By contrast, the gap of the optimal one-moment inequality $\widetilde{\Delta}_r(Y)$ increases without bound.

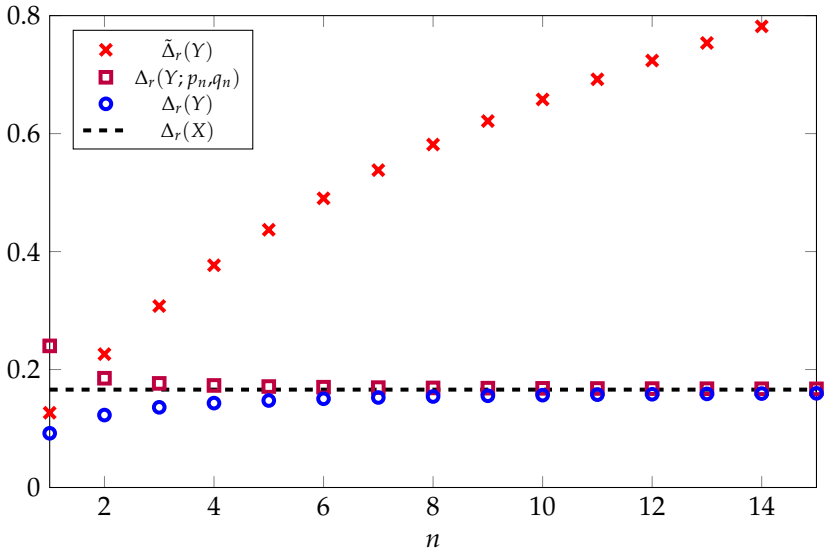

Figure 2. Comparison of upper bounds on Rényi entropy in nats for the multivariate Gaussian distribution $\mathcal{N}(0, I_n)$ as a function of the dimension n with $r = 0.1$. The solid black line is the gap of the optimal two-moment inequality for the log-normal distribution.

3.4. Inequalities for Differential Entropy

Proposition 4 can also be used to recover some known inequalities for differential entropy by considering the limiting behavior as r converges to one. For example, it is well known that the differential entropy of an n-dimensional random vector X with finite second moment satisfies

$$h(X) \leq \frac{1}{2} \log \left(2\pi e \, \mathbb{E} \left[\tfrac{1}{n} \|X\|^2 \right] \right), \tag{22}$$

with equality if and only if the entries of X are i.i.d. zero-mean Gaussian. A generalization of this result in terms of an arbitrary positive moment is given by

$$h(X) \leq \log \frac{\Gamma\left(\frac{n}{s}+1\right)}{\Gamma\left(\frac{n}{2}+1\right)} + \frac{n}{2} \log \pi + \frac{n}{s} \log \left(es \, \mathbb{E} \left[\tfrac{1}{n} \|X\|^s \right] \right), \tag{23}$$

for all $s > 0$. Note that Inequality (22) corresponds to the case $s = 2$.

Inequality (23) can be proved as an immediate consequence of Proposition 4 and the fact that $h_r(X)$ is nonincreasing in r. Using properties of the beta function given in Appendix A, it is straightforward to verify that

$$\lim_{r \to 1} \psi_r(0, q) = (e\,q)^{\frac{1}{q}} \, \Gamma\left(\frac{1}{q} + 1 \right), \quad \text{for all } q > 0.$$

Combining this result with Proposition 4 and Inequality (16) leads to

$$h(X) \leq \log \omega(S) + \log \Gamma\left(\frac{1}{q} + 1 \right) + \frac{1}{q} \log \left(eq \mathbb{E} \left[\|X\|^{nq} \right] \right).$$

Using Inequality (10) and making the substitution $s = nq$ leads to Inequality (23).

Another example follows from the fact that the log-normal distribution maximizes the differential entropy of a positive random variable X subject to constraints on the mean and variance of $\log(X)$, and hence

$$h(X) \leq \mathbb{E}\left[\log(X)\right] + \frac{1}{2} \log \left(2\pi e \, \mathsf{Var}(\log(X)) \right), \tag{24}$$

with equality if and only if X is log-normal. In Appendix B.4, it is shown how this inequality can be proved using our two-moment inequalities by studying the behavior as both p and q converge to zero as r increases to one.

4. Bounds on Mutual Information

4.1. Relative Entropy and Chi-Squared Divergence

Let P and Q be distributions defined on a common probability space that have densities p and q with respect to a dominating measure μ. The relative entropy (or Kullback–Leibler divergence) is defined according to

$$D\left(P \,\|\, Q\right) = \int p \log \left(\frac{p}{q} \right) \mathrm{d}\mu,$$

and the chi-squared divergence is defined as

$$\chi^2(P \,\|\, Q) = \int \frac{(p-q)^2}{q} \mathrm{d}\mu.$$

Both of these divergences can be seen as special cases of the general class of f-divergence measures and there exists a rich literature on comparisons between different divergences [8,26–32]. The chi-squared divergence can also be viewed as the squared L_2 distance between $p/\sqrt{q}$ and $\sqrt{q}$. The chi-square can

also be interpreted as the first non-zero term in the power series expansion of the relative entropy ([26], [Lemma 4]). More generally, the chi-squared divergence provides an upper bound on the relative entropy via

$$D\left(P \,\|\, Q\right) \leq \log(1 + \chi^2(P\|Q)). \tag{25}$$

The proof of this inequality follows straightforwardly from Jensen's inequality and the concavity of the logarithm; see [27,31,32] for further refinements.

Given a random pair (X, Y), the mutual information between X and Y is defined according to

$$I(X;Y) = D\left(P_{X,Y} \,\|\, P_X P_Y\right).$$

From Inequality (25), we see that the mutual information can always be upper bounded using

$$I(X;Y) \leq \log(1 + \chi^2(P_{X,Y}\|P_X P_Y)). \tag{26}$$

The next section provides bounds on the mutual information that can improve upon this inequality.

4.2. Mutual Information and Variance of Conditional Density

Let (X, Y) be a random pair such that the conditional distribution of Y given X has a density $f_{Y|X}(y|x)$ with respect to the Lebesgue measure on $\mathbb{R}^n$. Note that the marginal density of Y is given by $f_Y(y) = \mathbb{E}\left[f_{Y|X}(y|X)\right]$. To simplify notation, we will write $f(y|x)$ and $f(y)$ where the subscripts are implicit. The support set of Y is denoted by S_Y.

The measure of the dependence between X and Y that is used in our bounds can be understood in terms of the variance of the conditional density. For each y, the conditional density $f(y|X)$ evaluated with a random realization of X is a random variable. The variance of this random variable is given by

$$\mathsf{Var}(f(y|X)) = \mathbb{E}\left[(f(y|X) - f(y))^2\right], \tag{27}$$

where we have used the fact that the marginal density $f(y)$ is the expectation of $f(y|X)$. The sth moment of the variance of the conditional density is defined according to

$$V_s(Y|X) = \int_{S_Y} \|y\|^s \,\mathsf{Var}(f(y|X))\,\mathrm{d}y. \tag{28}$$

The variance moment $V_s(Y|X)$ is nonnegative and equal to zero if and only if X and Y are independent.

The function $\kappa(t)$ is defined according to

$$\kappa(t) = \sup_{u \in (0,\infty)} \frac{\log(1+u)}{u^t}, \qquad t \in (0, 1]. \tag{29}$$

The proof of the following result is given in Appendix C. The behavior of $\kappa(t)$ is illustrated in Figure 3.

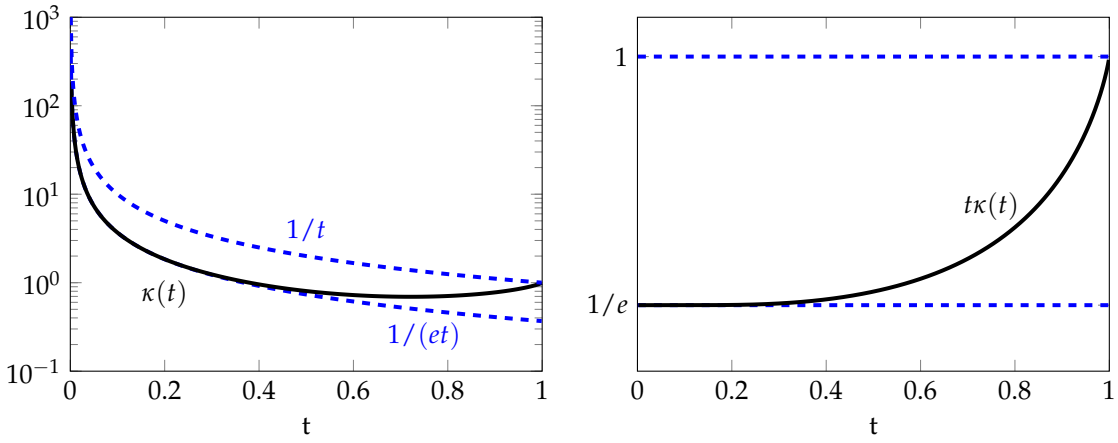

Figure 3. Graphs of $\kappa(t)$ and $t\kappa(t)$ as a function of t.

Proposition 7. *The function $\kappa(t)$ defined in Equation (29) can be expressed as*

$$\kappa(t) = \frac{\log(1+u)}{u^t}, \qquad t \in (0,1]$$

where

$$u = \exp\left(W\left(-\tfrac{1}{t}\exp\left(-\tfrac{1}{t}\right)\right) + \tfrac{1}{t}\right) - 1,$$

and $W(\cdot)$ denotes Lambert's W-function, i.e., $W(z)$ is the unique solution to the equation $z = w\exp(w)$ on the interval $[-1,\infty)$. Furthermore, the function $g(t) = t\kappa(t)$ is strictly increasing on $(0,1]$ with $\lim_{t\to 0} g(t) = 1/e$ and $g(1) = 1$, and thus

$$\frac{1}{et} \le \kappa(t) \le \frac{1}{t}, \qquad t \in (0,1],$$

where the lower bound $1/(et)$ is tight for small values of $t \in (0,1)$ and the upper bound $1/t$ is tight for values of t close to 1.

We are now ready to give the main results of this section, which are bounds on the mutual information. We begin with a general upper bound in terms of the variance of the conditional density.

Proposition 8. *For any $0 < t \le 1$, the mutual information satisfies*

$$I(X;Y) \le \kappa(t) \int_{S_Y} [f(y)]^{1-2t} \, [\mathsf{Var}(f(y \mid X))]^t \, \mathrm{d}y.$$

Proof. We use the following series of inequalities:

$$\begin{aligned} I(X;Y) &\overset{(a)}{=} \int f(y)\, D\left(P_{X|Y=y} \,\middle\|\, P_X\right) \mathrm{d}y \\ &\overset{(b)}{\le} \int f(y) \log\left(1 + \chi^2(P_{X|Y=y} \| P_X)\right) \mathrm{d}y \\ &\overset{(c)}{=} \int f(y) \log\left(1 + \frac{\mathsf{Var}(f(y \mid X))}{f^2(y)}\right) \mathrm{d}y \\ &\overset{(d)}{\le} \kappa(t) \int f(y) \left(\frac{\mathsf{Var}(f(y \mid X))}{f^2(y)}\right)^t \mathrm{d}y, \end{aligned}$$

where (a) follows from the definition of mutual information, (b) follows from Inequality (25), and (c) follows from Bayes' rule, which allows us to write the chi-square in terms of the variance of the conditional density:

$$\chi^2(P_{X|Y=y}\|P_X) = \mathbb{E}\left[\left(\frac{f(y|X)}{f(y)} - 1\right)^2\right] = \frac{\mathsf{Var}(f(y|X))}{f^2(y)}.$$

Inequality (d) follows from the nonnegativity of the variance and the definition of $\kappa(t)$. □

Evaluating Proposition 8 with $t = 1$ recovers the well-known inequality $I(X;Y) \le \chi^2(P_{X,Y}\|P_X P_Y)$. The next two results follow from the cases $0 < t < \frac{1}{2}$ and $t = \frac{1}{2}$, respectively.

Proposition 9. *For any $0 < r < 1$, the mutual information satisfies*

$$I(X;Y) \le \kappa(t)\left(e^{h_r(Y)}\, V_0(Y|X)\right)^t,$$

where $t = (1-r)/(2-r)$.

Proof. Starting with Proposition 8 and applying Hölder's inequality with conjugate exponents $1/(1-t)$ and $1/t$ leads to

$$I(X;Y) \le \kappa(t)\left(\int f^r(y)\,\mathrm{d}y\right)^{1-t}\left(\int \mathsf{Var}(f(y \mid X))\,\mathrm{d}y\right)^t = \kappa(t)\, e^{t\, h_r(Y)}\, V_0^t(Y|X),$$

where we have used the fact that $r = (1-2t)/(1-t)$. □

Proposition 10. *For any $p < 1 < q$, the mutual information satisfies*

$$I(X;Y) \le C(\lambda)\sqrt{\frac{\omega(S_Y)\, V_{np}^{\lambda}(Y|X)\, V_{nq}^{1-\lambda}(Y|X)}{(q-p)}},$$

where $\lambda = (q-1)/(q-p)$ and

$$C(\lambda) = \kappa(\tfrac{1}{2})\sqrt{\frac{\pi\lambda^{-\lambda}(1-\lambda)^{-(1-\lambda)}}{\sin(\pi\lambda)}},$$

with $\kappa(\frac{1}{2}) = 0.80477\ldots$.

Proof. Evaluating Proposition 8 with $t = 1/2$ gives

$$I(X;Y) \le \kappa(\tfrac{1}{2})\int_{S_Y}\sqrt{\mathsf{Var}(f(y \mid X))}\,\mathrm{d}y.$$

Evaluating Proposition 3 with $r = \frac{1}{2}$ leads to

$$\left(\int_{S_Y}\sqrt{\mathsf{Var}(f(y \mid X))}\,\mathrm{d}y\right)^2 \le \omega(S_Y)\,\psi_{1/2}(p,q)\, V_{np}^{\lambda}(Y|X)\, V_{nq}^{1-\lambda}(Y|X).$$

Combining these inequalities with the expression for $\psi_{1/2}(p,q)$ given in Equation (8) completes the proof. □

The contribution of Propositions 9 and 10 is that they provide bounds on the mutual information in terms of quantities that can be easy to characterize. One application of these bounds is to establish

conditions under which the mutual information corresponding to a sequence of random pairs (X_k, Y_k) converges to zero. In this case, Proposition 9 provides a sufficient condition in terms of the Rényi entropy of Y_n and the function $V_0(Y_n|X_n)$, while Proposition 10 provides a sufficient condition in terms of $V_s(Y_n|X_n)$ evaluated with two difference values of s. These conditions are summarized in the following result.

Proposition 11. *Let (X_k, Y_k) be a sequence of random pairs such that the conditional distribution of Y_k given X_k has a density on $\mathbb{R}^n$. The following are sufficient conditions under which the mutual information of $I(X_k; Y_k)$ converges to zero as k increases to infinity:*

1. *There exists $0 < r < 1$ such that*

$$\lim_{k\to\infty} e^{h_r(Y_k)} V_0(Y_k|X_k) = 0.$$

2. *There exists $p < 1 < q$ such that*

$$\lim_{k\to\infty} V_{np}^{q-1}(Y_k|X_k) V_{nq}^{1-p}(Y_k|X_k) = 0.$$

4.3. Properties of the Bounds

The variance moment $V_s(Y|X)$ has a number of interesting properties. The variance of the conditional density can be expressed in terms of an expectation with respect to two independent random variables X_1 and X_2 with the same distribution as X via the decomposition:

$$\mathsf{Var}(f(y|X)) = \mathbb{E}\left[f(y|X)f(y|X) - f(y|X_1)f(y|X_2)\right].$$

Consequently, by swapping the order of the integration and expectation, we obtain

$$V_s(Y|X) = \mathbb{E}\left[K_s(X, X) - K_s(X_1, X_2)\right], \tag{30}$$

where

$$K_s(x_1, x_2) = \int \|y\|^s f(y|x_1) f(y|x_2)\, \mathrm{d}y.$$

The function $K_s(x_1, x_2)$ is a positive definite kernel that does not depend on the distribution of X. For $s = 0$, this kernel has been studied previously in the machine learning literature [33], where it is referred to as the expected likelihood kernel.

The variance of the conditional density also satisfies a data processing inequality. Suppose that $U \to X \to Y$ forms a Markov chain. Then, the square of the conditional density of Y given U can be expressed as

$$f_{Y|U}^2(y|u) = \mathbb{E}\left[f_{Y|X}(y|X_1') f_{Y|X}(y|X_2') \mid U = u\right],$$

where $(U, X_1', X_2') \sim P_U P_{X_1|U} P_{X_2|U}$. Combining this expression with Equation (30) yields

$$V_s(Y|U) = \mathbb{E}\left[K_s(X_1', X_2') - K_s(X_1, X_2)\right], \tag{31}$$

where we recall that (X_1, X_2) are independent copies of X.

Finally, it is easy to verify that the function $V_s(Y)$ satisfies

$$V_s(aY|X) = |a|^{s-n} V_s(Y|X), \quad \text{for all } a \neq 0.$$

Using this scaling relationship, we see that the sufficient conditions in Proposition 11 are invariant to scaling of Y.

4.4. Example with Additive Gaussian Noise

We now provide a specific example of our bounds on the mutual information. Let $X \in \mathbb{R}^n$ be a random vector with distribution P_X and let Y be the output of a Gaussian noise channel

$$Y = X + W, \tag{32}$$

where $W \sim \mathcal{N}(0, I_n)$ is independent of X. If $\|X\|$ has finite second moment, then the mutual information satisfies

$$I(X;Y) \leq \frac{n}{2} \log \left(1 + \frac{1}{n} \mathbb{E}\left[\|X\|^2 \right] \right), \tag{33}$$

where equality is attained if and only if X has zero-mean isotropic Gaussian distribution. This inequality follows straightforwardly from the fact that the Gaussian distribution maximizes differential entropy subject to a second moment constraint [11]. One of the limitations of this bound is that it can be loose when the second moment is dominated by events that have small probability. In fact, it is easy to construct examples for which $\|X\|$ does not have a finite second moment, and yet $I(X;Y)$ is arbitrarily close to zero.

Our results provide bounds on $I(X;Y)$ that are less sensitive to the effects of rare events. Let $\phi_n(x) = (2\pi)^{-n/2} \exp(-\|x\|^2/2)$ denote the density of the standard Gaussian distribution on $\mathbb{R}^n$. The product of the conditional densities can be factored according to

$$\begin{aligned} f(y \mid x_1) f(y \mid x_2) &= \phi_{2n}\left(\begin{bmatrix} y - x_1 \\ y - x_2 \end{bmatrix} \right) = \phi_{2n}\left(\begin{bmatrix} \sqrt{2}y - (x_1 + x_2)/\sqrt{2} \\ (x_1 - x_2)/\sqrt{2} \end{bmatrix} \right) \\ &= \phi_n\left(\sqrt{2}\, y - \frac{x_1 + x_2}{\sqrt{2}} \right) \phi_n\left(\frac{x_1 - x_2}{\sqrt{2}} \right), \end{aligned}$$

where the second step follows because $\phi_{2n}(\cdot)$ is invariant to orthogonal transformations. Integrating with respect to y leads to

$$K_s(x_1, x_2) = 2^{-\frac{n+s}{2}} \mathbb{E}\left[\left\| W + \frac{x_1 + x_2}{\sqrt{2}} \right\|^s \right] \phi_n\left(\frac{x_1 - x_2}{\sqrt{2}} \right),$$

where we recall that $W \sim \mathcal{N}(0, I_n)$. For the case $s = 0$, we see that $K_0(x_1, x_2)$ is a Gaussian kernel, thus

$$V_0(Y|X) = (4\pi)^{-\frac{n}{2}} \left[1 - \mathbb{E}\left[e^{-\frac{1}{4}\|X_1 - X_2\|^2} \right] \right]. \tag{34}$$

A useful property of $V_0(Y|X)$ is that the conditions under which it converges to zero are weaker than the conditions needed for other measures of dependence. Observe that the expectation in Equation (34) is bounded uniformly with respect to (X_1, X_2). In particular, for every $\epsilon > 0$ and $x \in \mathbb{R}$, we have

$$1 - \mathbb{E}\left[e^{-\frac{1}{4}(X_1 - X_2)^2} \right] \leq \epsilon^2 + 2\mathbb{P}\left[|X - x| \geq \epsilon \right],$$

where we have used the inequality $1 - e^{-x} \leq x$ and the fact that $\mathbb{P}\left[|X_1 - X_2| \geq 2\epsilon \right] \leq 2\mathbb{P}\left[|X - x| \geq \epsilon \right]$. Consequently, $V_0(Y|X)$ converges to zero whenever X converges to a constant value x in probability.

To study some further properties of these bounds, we now focus on the case where X is a Gaussian scalar mixture generated according to

$$X = A\sqrt{U}, \quad A \sim \mathcal{N}(0,1), \quad U \geq 0, \tag{35}$$

with A and U independent. In this case, the expectations with respect to the kernel $K_s(x_1, x_2)$ can be computed explicitly, leading to

$$V_s(Y|X) = \frac{\Gamma(\frac{1+s}{2})}{2\pi}\mathbb{E}\left[(1+2U)^{\frac{s}{2}} - \frac{(1+U_1)^{\frac{s}{2}}(1+U_2)^{\frac{s}{2}}}{(1+\frac{1}{2}(U_1+U_2))^{\frac{s+1}{2}}}\right], \tag{36}$$

where (U_1, U_2) are independent copies of U. It can be shown that this expression depends primarily on the magnitude of U. This is not surprising given that X converges to a constant if and only if U converges to zero.

Our results can also be used to bound the mutual information $I(U;Y)$ by noting that $U \to X \to Y$ forms a Markov chain, and taking advantage of the characterization provided in Equation (31). Letting $X_1' = A_1\sqrt{U}$ and $X_2' = A_2\sqrt{U}$ with (A_1, A_2, U) be mutually independent leads to

$$V_s(Y|U) = \frac{\Gamma(\frac{1+s}{2})}{2\pi}\mathbb{E}\left[(1+U)^{\frac{s-1}{2}} - \frac{(1+U_1)^{\frac{s}{2}}(1+U_2)^{\frac{s}{2}}}{(1+\frac{1}{2}(U_1+U_2))^{\frac{s+1}{2}}}\right], \tag{37}$$

In this case, $V_s(Y|U)$ is a measure of the variation in U. To study its behavior, we consider the simple upper bound

$$V_s(Y|U) \leq \frac{\Gamma(\frac{1+s}{2})}{2\pi}\mathbb{P}\left[U_1 \neq U_2\right]\mathbb{E}\left[(1+U)^{\frac{s-1}{2}}\right], \tag{38}$$

which follows from noting that the term inside the expectation in Equation (37) is zero on the event $U_1 = U_2$. This bound shows that if $s \leq 1$ then $V_s(Y|U)$ is bounded uniformly with respect to distributions on U, and if $s > 1$, then $V_s(Y|U)$ is bounded in terms of the $(\frac{s-1}{2})$th moment of U.

In conjunction with Propositions 9 and 10, the function $V_s(Y|U)$ provides bounds on the mutual information $I(U;Y)$ that can be expressed in terms of simple expectations involving two independent copies of U. Figure 4 provides an illustration of the upper bound in Proposition 10 for the case where U is a discrete random variable supported on two points, and X and Y are generated according to Equations (32) and (35). This example shows that there exist sequences of distributions for which our upper bounds on the mutual information converge to zero while the chi-squared divergence between P_{XY} and $P_X P_Y$ is bounded away from zero.

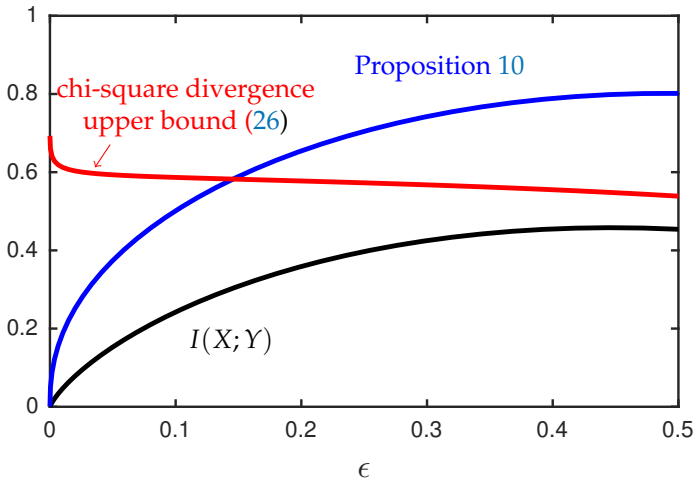

Figure 4. Bounds on the mutual information $I(U;Y)$ in nats when $U \sim (1-\epsilon)\delta_1 + \epsilon\delta_{a(\epsilon)}$, with $a(\epsilon) = 1 + 1/\sqrt{\epsilon}$, and X and Y are generated according to Equations (32) and (35). The bound from Proposition 10 is evaluated with $p = 0$ and $q = 2$.

5. Conclusions

This paper provides bounds on Rényi entropy and mutual information that are based on a relatively simple two-moment inequality. Extensions to inequalities with more moments are worth exploring. Another potential application is to provide a refined characterization of the "all-or-nothing" behavior seen in a sparse linear regression problem [34,35], where the current methods of analysis depend on a complicated conditional second moment method.

Funding: This research was supported in part by the National Science Foundation under Grant 1750362 and in part by the Laboratory for Analytic Sciences (LAS). Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author and do not necessarily reflect the views of the sponsors.

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. The Gamma and Beta Functions

This section reviews some properties of the gamma and beta functions. For $x > 0$, the gamma function is defined according to $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t$. Binet's formula for the logarithm for the gamma function ([25], [Theorem 1.6.3]) gives

$$\log \Gamma(x) = \left(x - \frac{1}{2}\right) \log x - x + \frac{1}{2} \log(2\pi) + \theta(x), \tag{A1}$$

where the remainder term $\theta(x)$ is convex and nonincreasing with $\lim_{x\to 0} \theta(x) = \infty$ and $\lim_{x\to\infty} \theta(x) = 0$. Euler's reflection formula ([25], [Theorem 1.2.1]) gives

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin(\pi x)}, \quad 0 < x < 1. \tag{A2}$$

For $x, y > 0$, the beta function can be expressed as follows

$$\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \int_0^1 t^{x-1}(1-t)^{y-1} \, \mathrm{d}t = \int_0^\infty \frac{u^{a-1}}{(1+u)^{a+b}} \, \mathrm{d}u, \tag{A3}$$

where the second integral expression follows from the change of variables $t \mapsto u/(1+u)$. Recall that $\widetilde{\mathrm{B}}(x,y) = \mathrm{B}(x,y)(x+y)^{x+y} x^{-x} y^{-y}$. Using Equation (A1) leads to

$$\log\left(\widetilde{\mathrm{B}}(x,y)\sqrt{\frac{xy}{2\pi(x+y)}}\right) = \theta(x) + \theta(y) - \theta(x+y). \tag{A4}$$

It can also be shown that ([36], [Equation (2), pg. 2])

$$\widetilde{\mathrm{B}}(x,y) \geq \frac{x+y}{xy}. \tag{A5}$$

Appendix B. Details for Rényi Entropy Examples

This appendix studies properties of the two-moment inequalities for Rényi entropy described in Section 3.

Appendix B.1. Log-Normal Distribution

Let X be a log-normal random variable with parameters (μ, σ^2) and consider the parametrization

$$p = \frac{1-r}{r} - (1-\lambda)\sqrt{\frac{(1-r)\,u}{r\lambda(1-\lambda)}}$$
$$q = \frac{1-r}{r} + \lambda\sqrt{\frac{(1-r)\,u}{r\lambda(1-\lambda)}}.$$

where $\lambda \in (0,1)$ and $u \in (0,\infty)$. Then, we have

$$\psi_r(p,q) = \widetilde{\mathrm{B}}\left(\frac{r\lambda}{1-r}, \frac{r(1-\lambda)}{1-r}\right)\sqrt{\frac{r\lambda(1-\lambda)}{(1-r)\,u}}$$
$$L_r(X;p,q) = \mu + \frac{1}{2}\left(\frac{1-r}{r}\right)\sigma^2 + \frac{1}{2}u\sigma^2.$$

Combining these expressions with Equation (A4) leads to

$$\Delta_r(X;p,q) = \theta\Big(\frac{r\lambda}{1-r}\Big) + \theta\Big(\frac{r(1-\lambda)}{1-r}\Big) - \theta\Big(\frac{r}{1-r}\Big) + \frac{1}{2}u\sigma^2 - \frac{1}{2}\log\left(u\sigma^2\right) - \frac{1}{2}\log(r^{\frac{1}{r-1}}). \tag{A6}$$

We now characterize the minimum with respect to the parameters (λ, u). Note that the mapping $\lambda \mapsto \theta(\frac{r\lambda}{1-r}) + \theta(\frac{r(1-\lambda)}{1-r})$ is convex and symmetric about the point $\lambda = 1/2$. Therefore, the minimum with respect to λ is attained at $\lambda = 1/2$. Meanwhile, mapping $u \mapsto u\sigma^2 - \log(u\sigma^2)$ is convex and attains its minimum at $u = 1/\sigma^2$. Evaluating Equation (A6) with these values, we see that the optimal two-moment inequality can be expressed as

$$\Delta_r(X) = 2\theta\left(\frac{r}{2(1-r)}\right) - \theta\left(\frac{r}{1-r}\right) + \frac{1}{2}\log\left(e\,r^{\frac{1}{1-r}}\right).$$

By Equation (A4), this expression is equivalent to Equation (A1). Moreover, the fact that $\Delta_r(X)$ decreases to zero as r increases to one follows from the fact that $\theta(x)$ decreases to zero and x increases to infinity.

Next, we express the gap in terms of the pair (p,q). Comparing the difference between $\Delta_r(X;p,q)$ and $\Delta_r(X)$ leads to

$$\Delta_r(X;p,q) = \Delta_r(X) + \frac{1}{2}\varphi\left(\frac{r\lambda(1-\lambda)}{1-r}(q-p)^2\sigma^2\right) + \theta\Big(\frac{r\lambda}{1-r}\Big) + \theta\Big(\frac{r(1-\lambda)}{1-r}\Big) - 2\theta\Big(\frac{r}{2(1-r)}\Big),$$

where $\varphi(x) = x - \log(x) - 1$. In particular, if $p = 0$, then we obtain the simplified expression

$$\Delta_r(X;0,q) = \Delta_r(X) + \frac{1}{2}\varphi\left(\Big(q - \frac{1-r}{r}\Big)\sigma^2\right) + \theta\Big(\frac{r}{1-r} - \frac{1}{q}\Big) + \theta\Big(\frac{1}{q}\Big) - 2\theta\Big(\frac{r}{2(1-r)}\Big).$$

This characterization shows that the gap of the optimal one-moment inequality $\widetilde{\Delta}_r(X)$ increases to infinity in the limit as either $\sigma^2 \to 0$ or $\sigma^2 \to \infty$.

Appendix B.2. Multivariate Gaussian Distribution

Let $Y \sim \mathcal{N}(0, I_n)$ be an n-dimensional Gaussian vector and consider the parametrization

$$p = \frac{1-r}{r} - \frac{1-\lambda}{r}\sqrt{\frac{2(1-r)\,z}{\lambda(1-\lambda)\,n}}$$
$$q = \frac{1-r}{r} + \frac{\lambda}{r}\sqrt{\frac{2(1-r)\,z}{\lambda(1-\lambda)\,n}}.$$

where $\lambda \in (0,1)$ and $z \in (0,\infty)$. We can write

$$\log \omega(S_Y) = \frac{n}{2}\log \pi - \log\left(\frac{n}{2}\right) - \log\Gamma\left(\frac{n}{2}\right)$$
$$\psi_r(p,q) = \widetilde{\mathrm{B}}\left(\frac{r\lambda}{1-r}, \frac{r(1-\lambda)}{1-r}\right)\sqrt{\frac{r\lambda(1-\lambda)}{(1-r)}}\sqrt{\frac{nr}{2z}}.$$

Furthermore, if

$$(1-\lambda)\sqrt{\frac{2(1-r)z}{\lambda(1-\lambda)n}} < 1, \tag{A7}$$

then $L_r(\|Y\|^n; p, q)$ is finite and is given by

$$L_r(\|Y\|^n; p, q) = Q_{r,n}(\lambda, z) + \frac{n}{2}\log 2 + \frac{r}{1-r}\left[\log\Gamma\left(\frac{n}{2r}\right) - \log\Gamma\left(\frac{n}{2}\right)\right],$$

where

$$Q_{r,n}(\lambda,z) = \frac{r\lambda}{1-r}\log\Gamma\left(\frac{n}{2r} - \frac{1-\lambda}{r}\sqrt{\frac{(1-r)nz}{2\lambda(1-\lambda)}}\right) + \frac{r(1-\lambda)}{1-r}\log\Gamma\left(\frac{n}{2r} + \frac{\lambda}{r}\sqrt{\frac{(1-r)nz}{2\lambda(1-\lambda)}}\right)$$
$$- \frac{r}{1-r}\log\Gamma\left(\frac{n}{2r}\right). \tag{A8}$$

Here, we note that the scaling in Equation (21) corresponds to $\lambda = 1/2$ and $z = n/(n+1)$, and thus the condition Inequality (A7) is satisfied for all $n \geq 1$. Combining the above expressions and then using Equations (A1) and (A4) leads to

$$\Delta_r(Y; p, q) = \theta\left(\frac{r\lambda}{1-r}\right) + \theta\left(\frac{r(1-\lambda)}{1-r}\right) - \theta\left(\frac{r}{1-r}\right) + Q_{r,n}(z,\lambda) - \frac{1}{2}\log z - \frac{1}{2}\log\left(r^{\frac{1}{r-1}}\right)$$
$$+ \frac{r}{1-r}\theta\left(\frac{n}{2r}\right) - \frac{1}{1-r}\theta\left(\frac{n}{2}\right). \tag{A9}$$

Next, we study some properties of $Q_{r,n}(\lambda, z)$. By Equation (A1), the logarithm of the gamma function can be expressed as the sum of convex functions:

$$\log\Gamma(x) = \varphi(x) + \frac{1}{2}\log\left(\frac{1}{x}\right) + \frac{1}{2}\log(2\pi) - 1 + \theta(x),$$

where $\varphi(x) = x\log x + 1 - x$. Starting with the definition of $Q(\lambda, z)$ and then using Jensen's inequality yields

$$\begin{aligned} Q_{r,n}(z,\lambda) &\geq \frac{r\lambda}{1-r}\varphi\left(\frac{n}{2r} - \frac{1-\lambda}{r}\sqrt{\frac{(1-r)nz}{2\lambda(1-\lambda)}}\right) \\ &\quad + \frac{r(1-\lambda)}{1-r}\varphi\left(\frac{n}{2r} + \frac{\lambda}{r}\sqrt{\frac{(1-r)nz}{2\lambda(1-\lambda)}}\right) - \frac{r}{1-r}\varphi\left(\frac{n}{2r}\right) \\ &= \frac{\lambda}{a}\varphi\left(1 - \sqrt{\left(\tfrac{1-\lambda}{\lambda}\right)az}\right) + \frac{(1-\lambda)}{a}\varphi\left(1 + \sqrt{\left(\tfrac{\lambda}{1-\lambda}\right)az}\right), \end{aligned}$$

where $a = 2(1-r)/n$. Using the inequality $\varphi(x) \geq (3/2)(x-1)^2/(x+2)$ leads to

$$\begin{aligned} Q_{r,n}(\lambda, z) &\geq \frac{z}{2}\left[\left(1 - \sqrt{\left(\tfrac{1-\lambda}{\lambda}\right)bz}\right)\left(1 + \sqrt{\left(\tfrac{\lambda}{1-\lambda}\right)bz}\right)\right]^{-1} \\ &\geq \frac{z}{2}\left(1 + \sqrt{\left(\tfrac{\lambda}{1-\lambda}\right)bz}\right)^{-1}, \end{aligned} \tag{A10}$$

where $b = 2(1-r)/(9n)$.

Observe that the right-hand side of Inequality (A10) converges to $z/2$ as n increases to infinity. It turns out this limiting behavior is tight. Using Equation (A1), it is straightforward to show that $Q_n(\lambda, z)$ converges pointwise to $z/2$ as n increases to infinity—that is,

$$\lim_{n\to\infty} Q_{r,n}(\lambda, z) = \frac{1}{2}z, \tag{A11}$$

for any fixed pair $(\lambda, z) \in (0,1) \times (0,\infty)$.

Appendix B.3. Proof of Proposition 6

Let $D = (0,1) \times (0,\infty)$. For fixed $r \in (0,1)$, we use $Q_n(\lambda, z)$ to denote the function $Q_{r,n}(\lambda, z)$ defined in Equation (A8) and we use $G_n(\lambda, z)$ to denote the right-hand side of Equation (A9). These functions are defined to be equal to positive infinity for any pair $(\lambda, z) \in D$ such that Inequality (A7) does not hold.

Note that the terms $\theta(n/(2r))$ and $\theta(n/2)$ converge to zero in the limit as n increases to infinity. In conjunction with Equation (A11), this shows that $G_n(\lambda, z)$ converges pointwise to a limit $G(\lambda, z)$ given by

$$G(\lambda, z) = \theta\left(\frac{r\lambda}{1-r}\right) + \theta\left(\frac{r(1-\lambda)}{1-r}\right) - \theta\left(\frac{r}{1-r}\right) + \frac{1}{2}z - \frac{1}{2}\log(z) - \frac{1}{2}\log(r^{\frac{1}{r-1}}).$$

At this point, the correspondence with the log-normal distribution can be seen from the fact that $G(\lambda, z)$ is equal to the right-hand side of Equation (A6) evaluated with $u\sigma^2 = z$.

To show that the gap corresponding to the log-normal distribution provides an upper bound on the limit, we use

$$\begin{aligned} \limsup_{n\to\infty} \Delta_r(Y) &= \limsup_{n\to\infty} \inf_{(\lambda,z)\in D} G_n(\lambda, z) \\ &\leq \inf_{(\lambda,z)\in D} \limsup_{n\to\infty} G_n(\lambda, z) \\ &= \inf_{(\lambda,z)\in D} G(\lambda, z) \\ &= \Delta_r(X). \end{aligned} \tag{A12}$$

Here, the last equality follows from the analysis in Appendix B.1, which shows that the minimum of $G(\lambda, z)$ is a attained at $\lambda = 1/2$ and $z = 1$.

To prove the lower bound requires a bit more work. Fix any $\epsilon \in (0,1)$ and let $D_\epsilon = (0, 1-\epsilon] \times (0, \infty)$. Using the lower bound on $Q_n(\lambda, z)$ given in Inequality (A10), it can be verified that

$$\liminf_{n\to\infty} \inf_{(\lambda,z)\in D_\epsilon} \left[Q_{r,n}(z,\lambda) - \frac{1}{2}\log z \right] \geq \frac{1}{2}.$$

Consequently, we have

$$\liminf_{n\to\infty} \inf_{(\lambda,z)\in D_\epsilon} G_n(\lambda, z) = \inf_{(\lambda,z)\in D_\epsilon} G(\lambda, z) \geq \Delta_r(X). \tag{A13}$$

To complete the proof we will show that for any sequence λ_n that converges to one as n increases to infinity, we have

$$\liminf_{n\to\infty} \inf_{z\in(0,\infty)} G_n(\lambda_n, z) = \infty. \tag{A14}$$

To see why this is the case, note that by Equation (A4) and Inequality (A5),

$$\theta\left(\tfrac{r\lambda}{1-r}\right) + \theta\left(\tfrac{r(1-\lambda)}{1-r}\right) - \theta\left(\tfrac{r}{1-r}\right) \geq \frac{1}{2}\log\left(\frac{1-r}{2\pi r\lambda(1-\lambda)}\right).$$

Therefore, we can write

$$G_n(\lambda, z) \geq Q_n(\lambda, z) - \frac{1}{2}\log\left(\lambda(1-\lambda)z\right) + c_n, \tag{A15}$$

where c_n is bounded uniformly for all n. Making the substitution $u = \lambda(1-\lambda)z$, we obtain

$$\inf_{z>0} G_n(\lambda, z) \geq \inf_{u>0} \left[Q_n\left(\lambda, \frac{u}{\lambda(1-\lambda)}\right) - \frac{1}{2}\log u \right] + c_n.$$

Next, let $b_n = 2(1-r)/(9n)$. The lower bound in Inequality (A10) leads to

$$\inf_{u>0} \left[Q_n\left(\lambda, \frac{u}{\lambda(1-\lambda)}\right) - \frac{1}{2}\log u \right] \geq \inf_{u>0} \left[\frac{u}{2\lambda}\left(\frac{1}{1-\lambda+\sqrt{b_n u}}\right) - \frac{1}{2}\log u \right]. \tag{A16}$$

The limiting behavior in Equation (A14) can now be seen as a consequence of Inequality (A15) and the fact that, for any sequence λ_n converging to one, the right-hand side of Inequality (A16) increases without bound as n increases. Combining Inequality (A12), Inequality (A13), and Equation (A14) establishes that the large n limit of $\Delta_r(Y)$ exists and is equal to $\Delta_r(X)$. This concludes the proof of Proposition 6.

Appendix B.4. Proof of Inequality (24)

Given any $\lambda \in (0,1)$ and $u \in (0, \infty)$ let

$$p(r) = \frac{1-r}{r} - \sqrt{\frac{1-r}{r}\left(\frac{1-\lambda}{\lambda}\right)u}$$

$$q(r) = \frac{1-r}{r} + \sqrt{\frac{1-r}{r}\left(\frac{\lambda}{1-\lambda}\right)u}.$$

We need the following results, which characterize the terms in Proposition 4 in the limit as r increases to one.

Lemma A1. *The function $\psi_r(p(r), q(r))$ satisfies*

$$\lim_{r\to 1} \psi_r(p(r), q(r)) = \sqrt{\frac{2\pi}{u}}.$$

Proof. Starting with Equation (A4), we can write

$$\psi_r(p,q) = \frac{1}{q-p}\sqrt{\frac{2\pi(1-r)}{r\lambda(1-\lambda}} \exp\left(\theta\Big(\frac{r\lambda}{1-r}\Big) + \theta\Big(\frac{r(1-\lambda)}{1-r}\Big) - \theta\Big(\frac{r}{1-r}\Big)\right).$$

As r converges to one, the terms in the exponent converge to zero. Note that $q(r) - p(r) = \sqrt{r\lambda(1-\lambda)/(1-r)}$ completes the proof. □

Lemma A2. *If X is a random variable such that $s \mapsto \mathbb{E}\left[|X|^s\right]$ is finite in a neighborhood of zero, then $\mathbb{E}\left[\log(X)\right]$ and $\mathsf{Var}(\log(X))$ are finite, and*

$$\lim_{r\to 1} L_r(X; p(r), q(r)) = \mathbb{E}\left[\log|X|\right] + \frac{u}{2}\,\mathsf{Var}(\log|X|).$$

Proof. Let $\Lambda(s) = \log(\mathbb{E}\left[|X|^s\right])$. The assumption that $\mathbb{E}\left[|X|^s\right]$ is finite in a neighborhood of zero means that $\mathbb{E}\left[(\log|X|)^m\right]$ is finite for all positive integers m, and thus $\Lambda(s)$ is real analytic in a neighborhood of zero. Hence, there exist constants $\delta > 0$ and $C < \infty$, depending on the distribution of X, such that

$$\left|\Lambda(s) - as + bs^2\right| \le C\,|s|^3, \quad \text{for all } |s| \le \delta,$$

where $a = \mathbb{E}\left[\log|X|\right]$ and $b = \frac{1}{2}\,\mathsf{Var}(|X|)$. Consequently, for all r such that $1-\delta < p(r) < (1-r)/r < q(r) < 1+\delta$, it follows that

$$\left|L_r(X; p(r), q(r)) - a - \left(\tfrac{1-r}{r} + u\right) b\right| \quad \le C\frac{r}{1-r}\left(\lambda|p(r)|^3 + (1-\lambda)|q(r)|^3\right).$$

Taking the limit as r increases to one completes the proof. □

We are now ready to prove Inequality (24). Combining Proposition 4 with Lemma A1 and Lemma A2 yields

$$\limsup_{r\to\infty} h_r(X) \le \frac{1}{2}\log\left(\frac{2\pi}{u}\right) + \mathbb{E}\left[\log X\right] + \frac{u}{2}\,\mathsf{Var}(\log X).$$

The stated inequality follows from evaluating the right-hand side with $u = 1/\,\mathsf{Var}(\log X)$, recalling that $h(X)$ corresponds to the limit of $h_r(X)$ as r increases to one.

Appendix C. Proof of Proposition 7

The function $\kappa\colon (0,1] \to \mathbb{R}_+$ can be expressed as

$$\kappa(t) = \sup_{u\in(0,\infty)} \rho_t(u), \tag{A17}$$

where $\rho_t(u) = \log(1+u)/u^t$. For $t = 1$, the bound $\log(1+u) \le u$ implies that $\rho_1(u) \le 1$. Noting that $\lim_{u\to 0}\rho_1(u) = 1$, we conclude that $\kappa(1) = 1$.

Next, we consider the case $t \in (0,1)$. The function ρ_t is continuously differentiable on $(0,\infty)$ with

$$\mathrm{sgn}(\rho_t'(u)) = \mathrm{sgn}\left(u - t(1+u)\log(1+u)\right). \tag{A18}$$

Under the assumption $t \in (0,1)$, we see that $\rho_t(u)$ is increasing for all u sufficiently close to zero and decreasing for all u sufficiently large, and thus the supremum is attained at a stationary point of $\rho_t(u)$ on $(0,\infty)$. Making the substitution $w = \log(1+u) - 1/t$ leads to

$$\rho_t'(u) = 0 \quad \Longleftrightarrow \quad we^w = -\frac{1}{t}e^{-\frac{1}{t}}.$$

For $t \in (0,1)$, it follows that $-\frac{1}{t}e^{-\frac{1}{t}} \in (-e^{-1}, 0)$, and thus $\rho_t'(u)$ has a unique root that can be expressed as

$$u_t^* = \exp\left(W\left(-\tfrac{1}{t}\exp\left(-\tfrac{1}{t}\right)\right) + \tfrac{1}{t}\right) - 1,$$

where Lambert's function $W(z)$ is the solution to the equation $z = we^w$ on the interval on $[-1,\infty)$.

Lemma A3. *The function $g(t) = t\kappa(t)$ is strictly increasing on $(0,1]$ with $\lim_{t\to 0} g(t) = 1/e$ and $g(1) = 1$.*

Proof. The fact that $g(1) = 1$ follows from $\kappa(1) = 1$. By the envelope theorem [37], the derivative of $g(t)$ can be expressed as

$$g'(t) = \frac{\mathrm{d}}{\mathrm{d}t}\, t\rho_t(u)\Big|_{u=u_t^*} = \frac{\log(1+u_t^*)}{(u_t^*)^t} - t\log(u_t^*)\frac{\log(1+u_t^*)}{(u_t^*)^t}$$

In view of Equation (A18), it follows that $\rho_t'(u_t^*) = 0$ can be expressed equivalently as

$$\frac{u_t^*}{(1+u_t^*)\log(1+u_t^*)} = t, \tag{A19}$$

and thus

$$\operatorname{sgn}(g'(t)) = \operatorname{sgn}\left(1 - \frac{u_t^*\log u_t^*}{(1+u_t^*)\log(1+u_t^*)}\right). \tag{A20}$$

Noting that $u\log u < (1+u)\log(1+u)$ for all $u \in (0,\infty)$, it follows that $g'(t) > 0$ is strictly positive, and thus $g(t)$ is strictly increasing.

To prove the small t limit, we use Equation (A19) to write

$$\log(g(t)) = \log\left(\frac{u_t^*}{1+u_t^*}\right) - \frac{u_t^*\log u_t^*}{(1+u_t^*)\log(1+u_t^*)}. \tag{A21}$$

Now, as t decreases to zero, Equation (A19) shows that u_t^* increases to infinity. By Equation (A21), it then follows that $\log(g(t))$ converges to negative one, which proves the desired limit. □

References

1. Dembo, A.; Cover, T.M.; Thomas, J.A. Information Theoretic Inequalities. *IEEE Trans. Inf. Theory* **1991**, *37*, 1501–1518. [CrossRef]
2. Carslon, F. Une inégalité. *Ark. Mat. Astron. Fys.* **1934**, *25*, 1–5.
3. Levin, V.I. Exact constants in inequalities of the Carlson type. *Doklady Akad. Nauk. SSSR (N. S.)* **1948**, *59*, 635–638.
4. Larsson, L.; Maligranda, L.; Persson, L.E.; Pečarić, J. *Multiplicative Inequalities of Carlson Type and Interpolation*; World Scientific Publishing Company: Singapore, 2006.
5. Barza, S.; Burenkov, V.; Pečarić, J.E.; Persson, L.E. Sharp multidimensional multiplicative inequalities for weighted L_p spaces with homogeneous weights. *Math. Inequalities Appl.* **1998**, *1*, 53–67. [CrossRef]
6. Reeves, G. Two-Moment Inequailties for Rényi Entropy and Mutual Information. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 664–668.

7. Gray, R.M. *Entropy and Information Theory*; Springer-Verlag: Berlin/Heidelberg, Germany, 2013.
8. van Erven, T.; Harremoës, P. Rényi Divergence and Kullback–Liebler Divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3937–3820. [CrossRef]
9. Atar, R.; Chowdharry, K.; Dupuis, P. Abstract. Robust Bounds on Risk-Sensitive Functionals via Rényi Divergence. *SIAM/ASA J. Uncertain. Quantif.* **2015**, *3*, 18–33. [CrossRef]
10. Rosenkrantz, R. (Ed.) *E. T. Jaynes: Papers on Probability, Staistics and Statistical Physics*; Springer: Berlin/Heidelberg, Germany, 1989.
11. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
12. Lutwak, E.; Yang, D.; Zhang, G. Moment-entropy inequalities. *Ann. Probab.* **2004**, *32*, 757–774. [CrossRef]
13. Lutwak, E.; Yang, D.; Zhang, G. Moment-Entropy Inequalities for a Random Vector. *IEEE Trans. Inf. Theory* **2007**, *53*, 1603–1607. [CrossRef]
14. Lutwak, E.; Lv, S.; Yang, D.; Zhang, G. Affine Moments of a Random Vector. *IEEE Trans. Inf. Theory* **2013**, *59*, 5592–5599. [CrossRef]
15. Costa, J.A.; Hero, A.O.; Vignat, C. A Characterization of the Multivariate Distributions Maximizing Rényi Entropy. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Lausanne, Switzerland, 30 June–5 July 2002. [CrossRef]
16. Costa, J.A.; Hero, A.O.; Vignat, C. A Geometric Characterization of Maximum Rényi Entropy Distributions. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Seattle, WA, USA, 9–14 July 2006; pp. 1822–1826.
17. Johnson, O.; Vignat, C. Some results concerning maximum Rényi entropy distributions. *Ann. de l'Institut Henri Poincaré (B) Probab. Stat.* **2007**, *43*, 339–351. [CrossRef]
18. Nguyen, V.H. A simple proof of the Moment-Entropy inequalities. *Adv. Appl. Math.* **2019**, *108*, 31–44. [CrossRef]
19. Barron, A.; Yang, Y. Information-theoretic determination of minimax rates of convergence. *Ann. Stat.* **1999**, *27*, 1564–1599. [CrossRef]
20. Wu, Y.; Xu, J. Statistical problems with planted structures: Information-theoretical and computational limits. In *Information-Theoretic Methods in Data Science*; Rodrigues, M.R.D.; Eldar, Y.C., Eds.; Cambridge University Press: Cambridge, UK, 2020; Chapter 13.
21. Reeves, G. Conditional Central Limit Theorems for Gaussian Projections. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Aachen, Germany, 25–30 June 2017; pp. 3055–3059.
22. Reeves, G.; Pfister, H.D. The Replica-Symmetric Prediction for Random Linear Estimation with Gaussian Matrices is Exact. *IEEE Trans. Inf. Theory* **2019**, *65*, 2252–2283. [CrossRef]
23. Reeves, G.; Pfister, H.D. Understanding Phase Transitions via Mutual Information and MMSE. In *Information-Theoretic Methods in Data Science*; Rodrigues, M.R.D.; Eldar, Y.C., Eds.; Cambridge University Press: Cambridge, UK, 2020; Chapter 7.
24. Rockafellar, R.T. *Convex Analysis*; Princeton University Press: Princeton, NJ, USA, 1970.
25. Andrews, G.E.; Askey, R.; Roy, R. *Special Functions; Vol. 71, Encyclopedia of Mathematics and its Applications*, Cambridge University Press: Cambridge, UK, 1999.
26. Nielsen, F.; Nock, R. On the Chi Square and Higher-Order Chi Distrances for Approximationg f-Divergences. *IEEE Signal Process. Lett.* **1014**, *1*, 10–13.
27. Sason, I.; Verdú, S. f-Divergence Inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
28. Sason, I. On the Rényi Divergence, Joint Range of Relative Entropy, and a Channel Coding Theorem. *IEEE Trans. Inf. Theory* **2016**, *62*, 23–34. [CrossRef]
29. Sason, I.; Verdú, S. Improved Bounds on Lossless Source Coding and Guessing Moments via Rényi Measures. *IEEE Trans. Inf. Theory* **2018**, *64*, 4323–4326. [CrossRef]
30. Sason, I. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy* **2018**, *20*, 383. [CrossRef]
31. Melbourne, J.; Madiman, M.; Salapaka, M.V. Relationships between certain f-divergences. In Proceedings of the Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 24–27 September 2019; pp. 1068–1073.
32. Nishiyama, T.; Sason, I. On Relations Between the Relative Entropy and χ^2-Divergence, Generalizations and Applications. *Entropy* **2020**, *22*, 563. [CrossRef]
33. Jebara, T.; Kondor, R.; Howard, A. Probability Product Kernels. *J. Mach. Learn. Res.* **2004**, *5*, 818–844.

34. Reeves, G.; Xu, J.; Zadik, I. The All-or-Nothing Phenomenon in Sparse Linear Regression. In Proceedings of the Conference On Learning Theory (COLT), Phoenix, AZ, USA, 25–28 June 2019.
35. Reeves, G.; Xu, J.; Zadik, I. All-or-nothing phenomena from single-letter to high dimensions. In Proceedings of the IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Guadeloupe, France, 15–18 December 2019.
36. Grenié, L.; Molteni, G. Inequalities for the beta function. *Math. Inequalities Appl.* **2015**, *18*, 1427–1442. [CrossRef]
37. Milgrom, P.; Segal, I. Envelope Theorems for Arbitrary Choice Sets. *Econometrica* **2002**, *70*, 583–601. [CrossRef]