entropy

MDPI

Article

The Cauchy Distribution in Information Theory

Sergio Verdú

Independent Researcher, Princeton, NJ 08540, USA; verdu@informationtheory.org

Abstract: The Gaussian law reigns supreme in the information theory of analog random variables. This paper showcases a number of information theoretic results which find elegant counterparts for Cauchy distributions. New concepts such as that of equivalent pairs of probability measures and the strength of real-valued random variables are introduced here and shown to be of particular relevance to Cauchy distributions.

Keywords: information measures; Cauchy distribution; relative entropy; Kullback–Leibler divergence; differential entropy; Fisher's information; entropy power inequality; f-divergence; Rényi divergence; mutual information; data transmission; lossy data compression

Citation: Verdú, S. The Cauchy Distribution in Information Theory. *Entropy* **2023**, *25*, 346. https://doi.org/10.3390/e25020346

Academic Editor: Sangun Park

Received: 20 September 2022
Revised: 5 November 2022
Accepted: 2 February 2023
Published: 13 February 2023

1. Introduction

Since the inception of information theory [1], the *Gaussian distribution* has emerged as the paramount example of a continuous random variable leading to closed-form expressions for information measures and extremality properties possessing great pedagogical value. In addition, the role of the Gaussian distribution as a ubiquitous model for analog information sources and for additive thermal noise has elevated the corresponding formulas for rate–distortion functions and capacity–cost functions to iconic status in information theory. Beyond discrete random variables, by and large, information theory textbooks confine their coverage and examples to Gaussian random variables.

The *exponential distribution* has also been shown [2] to lead to closed-form formulas for various information measures such as differential entropy, mutual information and relative entropy, as well as rate–distortion functions for Markov processes and the capacity of continuous-time timing channels with memory such as the exponential-server queue [3].

Despite its lack of moments, the *Cauchy distribution* also leads to pedagogically attractive closed-form expressions for various information measures. In addition to showcasing those, we introduce an attribute, which we refer to as the *strength* of a real-valued random variable, under which the Cauchy distribution is shown to possess optimality properties. Along with the stability of the Cauchy law, those properties result in various counterparts to the celebrated fundamental limits for memoryless Gaussian sources and channels.

To enhance readability and ease of reference, the rest of this work is organized in 120 items grouped into 17 sections, plus an appendix.

Section 2 presents the family of Cauchy random variables and their basic properties as well as multivariate generalizations, and the Rider univariate density which includes the Cauchy density as a special case and finds various information theoretic applications.

Section 3 gives closed-form expressions for the differential entropies of the univariate and multivariate densities covered in Section 2.

Introduced previously for unrelated purposes, the Shannon and η-transforms reviewed in Section 4 prove useful to derive several information theoretic results for Cauchy and related laws.

Applicable to any real-valued random variable and inspired by information theory, the central notion of *strength* is introduced in Section 5 along with its major properties. In particular, it is shown that convergence in strength is an intermediate criterion between

convergence in probability and convergence in L_q, $q > 0$, and that differential entropy is continuous with respect to the addition of independent vanishing strength noise.

Section 6 shows that, for any $\rho > 0$ the maximal differential entropy density satisfying $\mathbb{E}[\log(1 + |Z|^\rho)] \leq \theta$ can be obtained in closed form, but its shape (not just its scale) depends on the value of θ. In particular, the Cauchy density is the solution only if $\rho = 2$, and $\theta = \log 4$. In contrast, we show that, among all the random variables with a given strength, the centered Cauchy density has maximal differential entropy, regardless of the value of the constraint. This result suggests the definition of *entropy strength* of Z, as the strength of a Cauchy random variable whose differential entropy is the same as that of Z. Modulo a factor, entropy power is the square of entropy strength. Section 6 also gives a maximal differential entropy characterization of the standard spherical Cauchy multivariate density.

Information theoretic terminology for the logarithm of the Radon–Nikodym derivative, as well as its distribution, the *relative information spectrum* is given in Section 7. The relative information spectrum for Cauchy distributions is found and shown to depend on their location and scale through a single scalar. This is a rare property, not satisfied by most common families such as Gaussian, exponential, Laplace, etc. Section 8 introduces the notion of *equivalent pairs* of probability measures, which plays an important role not only in information theory but in statistical inference. Distinguishing P_1 from Q_1 has the same fundamental limits as distinguishing P_2 from Q_2 if (P_1, Q_1) and (P_2, Q_2) are equivalent pairs. Section 9 studies the interplay between f-divergences and equivalent pairs. A simple formula for the f-divergence between Cauchy distributions results from the explicit expression for the relative information spectrum found in Section 7. These results are then used to easily derive a host of explicit expressions for χ^2-divergence, relative entropy, total variation distance, Hellinger divergence and Rényi divergence in Sections 10–14, respectively.

In addition to the Fisher information matrix of the Cauchy family, Section 15 finds a counterpart of de Bruijn's identity [4] for convolutions with scaled Cauchy random variables, instead of convolutions with scaled Gaussian random variables as in the conventional setting.

Section 16 is devoted to mutual information. The mutual information between a Cauchy random variable and its noisy version contaminated by additive independent Cauchy noise exhibits a pleasing counterpart (modulo a factor of two) with the Gaussian case, in which the signal-to-noise ratio is now given by the ratio of strengths rather than variances. With Cauchy noise, Cauchy inputs maximize mutual information under an output strength constraint. The elementary fact that an output variance constraint translates directly into an input variance constraint does not carry over to input and output strengths, and indeed we identify non-Cauchy inputs that may achieve higher mutual information than a Cauchy input with the same strength. Section 16 also considers the dual setting in which the input is Cauchy, but the additive noise need not be. Lower bounds on the mutual information, attained by Cauchy noise, are offered. However, as the bounds do not depend exclusively on the noise strength, they do not rule out the possibility that a non-Cauchy noise with identical strength may be least favorable. If distortion is measured by strength, the rate–distortion function of a Cauchy memoryless source is shown to admit (modulo a factor of two) the same rate–distortion function as the memoryless Gaussian source with mean–square distortion, replacing the source variance by its strength. Theorem 17 gives a very general continuity result for mutual information that encompasses previous such results. While convergence in probability to zero of the input to an additive-noise transformation does not imply vanishing input-output mutual information, convergence in strength does under very general conditions on the noise distribution.

Some concluding observations about generalizations and open problems are collected in Section 17, including a generalization of the notion of strength.

Those definite integrals used in the main body are collected and justified in the Appendix A.

2. The Cauchy Distribution and Generalizations

In probability theory, the Cauchy (also known as Lorentz and as Breit–Wigner) distribution is the prime example of a real-valued random variable none of whose moments of order one or higher exists, and as such it is not encompassed by either the law of large numbers or the central limit theorem.

1. A real-valued random variable V is said to be *standard Cauchy* if its probability density function is

$$f_V(x) = \frac{1}{\pi}\frac{1}{x^2+1}, \quad x \in \mathbb{R}. \tag{1}$$

Furthermore, X is said to be Cauchy if there exist $\lambda \neq 0$ and $\mu \in \mathbb{R}$ such that $X = \lambda V + \mu$, in which case

$$f_X(x) = \frac{|\lambda|}{\pi}\frac{1}{(x-\mu)^2+\lambda^2}, \quad x \in \mathbb{R}, \tag{2}$$

where μ and $|\lambda|$ are referred to as the *location* (or median) and *scale*, respectively, of the Cauchy distribution. If $\mu = 0$, (2) is said to be centered Cauchy.

2. Since $\mathbb{E}[\max\{0, V\}] = \mathbb{E}[\max\{0, -V\}] = \infty$, the mean of a Cauchy random variable does not exist. Furthermore, $\mathbb{E}[|V|^q] = \infty$ for $q \geq 1$, and the moment generating function of V does not exist (except, trivially, at 0). The characteristic function of the standard Cauchy random variable is

$$\mathbb{E}\left[\mathrm{e}^{i\omega V}\right] = \mathrm{e}^{-|\omega|}, \quad \omega \in \mathbb{R}. \tag{3}$$

3. Using (3), we can verify that a Cauchy random variable has the curious property that adding an independent copy to it has the same effect, statistically speaking, as adding an identical copy. In addition to the Gaussian and Lévy distributions, the Cauchy distribution is *stable*: a linear combination of independent copies remains in the family, and is *infinitely divisible*: it can be expressed as an n-fold convolution for any n. It follows from (3) that if $\{V_1, V_2, \ldots\}$ are independent, standard Cauchy, and $\mathbf{a}$ is a deterministic sequence with finite ℓ_1-norm $\|\mathbf{a}\|_1$, then $\sum_{i=1}^{\infty} a_i V_i$ has the same distribution as $\|\mathbf{a}\|_1 V$. In particular, the time average of independent identically distributed Cauchy random variables has the same distribution as any of the random variables. The families $\{\lambda V, \lambda \in \mathcal{I}\}$ and $\{V + \mu, \mu \in \mathcal{I}\}$, with $\mathcal{I}$ any interval of the real line, are some of the simplest parametrized random variables that are not an *exponential family*.
4. If Θ is uniformly distributed on $[-\frac{\pi}{2}, \frac{\pi}{2}]$, then $\tan\Theta$ is standard Cauchy. This follows since, in view of (1) and (A1), the standard Cauchy cumulative distribution function is

$$F_V(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x), \quad x \in \mathbb{R}. \tag{4}$$

Therefore, V has unit semi-interquartile length. The functional inverse of (4) is the standard Cauchy *quantile function* given by

$$\mathbb{Q}_V(t) = \tan\left(\pi\left(t - \tfrac{1}{2}\right)\right), \quad t \in (0,1). \tag{5}$$

5. If X_1 and X_2 are standard Gaussian with correlation coefficient $\rho \in (-1, 1)$, then X_1/X_2 is Cauchy with scale $\sqrt{1-\rho^2}$ and location ρ. This implies that the reciprocal of a standard Cauchy random variable is also standard Cauchy.
6. Taking the cue from the Gaussian case, we say that a random vector is multivariate Cauchy if any linear combination of its components has a Cauchy distribution. Necessary and sufficient conditions for a characteristic function to be that of a multivariate

Cauchy were shown by Ferguson [5]. Unfortunately, no general expression is known for the corresponding probability density function. This accounts for the fact that one aspect, in which the Cauchy distribution does not quite reach the wealth of information theoretic results attainable with the Gaussian distribution, is in the study of multivariate models of dependent random variables. Nevertheless, special cases of multivariate Cauchy distribution do admit some interesting information theoretic results as we will see below. The standard spherical multivariate Cauchy probability density function on $\mathbb{R}^n$ is (e.g., [6])

$$f_{V^n}(\mathbf{x}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\pi^{\frac{n+1}{2}}}\left(1+\|\mathbf{x}\|^2\right)^{-\frac{n+1}{2}}, \tag{6}$$

where $\Gamma(\cdot)$ is the Gamma function. Therefore, $V^n = (V_1, \ldots, V_n)$ are exchangeable random variables. If $X_0, X_1, \ldots, X_n$ are independent standard normal, then the vector $X_0^{-1}X^n$ has the density in (6). With the aid of (A10), we can verify that any subset of $k \in \{1, \ldots, n-1\}$ components of V^n is distributed according to V^k. In particular, the marginals of (6) are given by (1). Generalizing (3), the characteristic function of (6) is

$$\mathbb{E}\left[\mathrm{e}^{i\,\mathbf{t}^\top V^n}\right] = \mathrm{e}^{-\|\mathbf{t}\|}, \quad \mathbf{t} \in \mathbb{R}^n. \tag{7}$$

7. In parallel to Item 1, we may generalize (6) by dropping the restriction that it be centered at the origin and allowing ellipsoidal deformation, i.e., letting $Z^n = \mathbf{\Lambda}^{\frac{1}{2}}V^n + \boldsymbol{\mu}$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and a positive definite $n \times n$ matrix $\mathbf{\Lambda}$. Therefore,

$$f_{Z^n}(\mathbf{x}) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\pi^{\frac{n+1}{2}}\det^{\frac{1}{2}}(\mathbf{\Lambda})}\left(1+(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)^{-\frac{n+1}{2}}. \tag{8}$$

While $\boldsymbol{\rho}^\top Z^n$ is a Cauchy random variable for $\boldsymbol{\rho} \in \mathbb{R}^n - \{0\}$, (8) fails to encompass every multivariate Cauchy distribution—in particular, the important case of independent Cauchy random variables. Another reason the usefulness of the model in (8) is limited is that it is not closed under independent additions: if V^n and $\bar{V}^n$ are independent, each distributed according to (6); then, $\mathbf{\Lambda}^{\frac{1}{2}}V^n + \bar{\mathbf{\Lambda}}^{\frac{1}{2}}\bar{V}^n$, while multivariate Cauchy, does not have a density of the type in (8) unless $\mathbf{\Lambda} = \alpha\,\bar{\mathbf{\Lambda}}$ for some $\alpha > 0$.

8. Another generalization of the (univariate) Cauchy distribution, which comes into play in our analysis, was introduced by Rider in 1958 [7]. With $\rho > 0$ and $\beta\rho > 1$,

$$f_{V_{\beta,\rho}}(x) = \frac{\kappa_{\beta,\rho}}{(1+|x|^\rho)^\beta}, \quad x \in \mathbb{R}, \tag{9}$$

$$\kappa_{\beta,\rho} = \frac{\rho\,\Gamma(\beta)}{2\,\Gamma\left(\frac{1}{\rho}\right)\Gamma\left(\beta-\frac{1}{\rho}\right)}. \tag{10}$$

In addition to the (β, ρ) parametrization in (9), we may introduce scale and location parameters by means of $\lambda V_{\beta,\rho} + \mu$, just as we did in the Cauchy case $(\beta, \rho) = (1, 2)$. Another notable special case is $\sqrt{\nu}\,V_{\frac{\nu+1}{2},2}$, which is the centered Student-t random variable, itself equivalent to a Pearson type VII distribution.

3. Differential Entropy

9. The differential entropy of a Cauchy random variable is

$$h(\lambda V + \mu) = \log|\lambda| + h(V), \tag{11}$$

$$h(V) = -\int_{-\infty}^{\infty} f_V(t)\,\log f_V(t)\,\mathrm{d}t = \log(4\pi), \tag{12}$$

using (A3). Throughout this paper, unless the logarithm base is explicitly shown, it can be chosen by the reader as long as it is the same on both sides of the equation. For natural logarithms, the information measure unit is the *nat*.

10. An alternative, sometimes advantageous, expression for the differential entropy of a real-valued random variable is feasible if its cumulative distribution function F_X is continuous and strictly monotonic. Then, the quantile function is its functional inverse, i.e., $F_X(\mathbb{Q}_X(t)) = t$ for all $t \in (0,1)$, which implies that $\dot{\mathbb{Q}}_X(t) f_X(\mathbb{Q}_X(t)) = 1$ for all $t \in (0,1)$. Moreover, since X and $\mathbb{Q}_X(U)$ with U uniformly distributed on $[0,1]$ have identical distributions, we obtain

$$h(X) = \mathbb{E}[-\log f_X(X)] = \mathbb{E}[-\log f_X(\mathbb{Q}_X(U))] = \int_0^1 \log \dot{\mathbb{Q}}_X(t)\, \mathrm{d}t. \tag{13}$$

Since (4) is indeed continuous and strictly monotonic, we can verify that we recover (12) by means of (5), (13) and (A2).

11. Despite not having finite moments, an independent identically distributed sequence of Cauchy random variables $\{Z_i\}$ is *information stable* in the sense that

$$\frac{1}{n}\sum_{i=1}^{n} \log f_Z(Z_i) \to -h(Z), \quad \text{a.s.} \tag{14}$$

because of the strong law of large numbers.

12. With V^n distributed according to the standard spherical multivariate Cauchy density in (6), it is shown in [8] that

$$\mathbb{E}\Big[\log_{\mathrm{e}}\Big(1 + \|V^n\|^2\Big)\Big] = \psi\left(\frac{n+1}{2}\right) + \log_{\mathrm{e}} 4 + \gamma, \tag{15}$$

where γ is the Euler–Mascheroni constant and $\psi(\cdot)$ is the digamma function. Therefore, the differential entropy of (6) is, in nats, (see also [9])

$$h(V^n) = \frac{n+1}{2}\mathbb{E}\Big[\log_{\mathrm{e}}\Big(1 + \|V^n\|^2\Big)\Big] + \frac{n+1}{2}\log_{\mathrm{e}} \pi - \log_{\mathrm{e}} \Gamma\left(\frac{n+1}{2}\right) \tag{16}$$

$$= \frac{n+1}{2}\left(\log_{\mathrm{e}}(4\pi) + \gamma + \psi\left(\frac{n+1}{2}\right)\right) - \log_{\mathrm{e}} \Gamma\left(\frac{n+1}{2}\right), \tag{17}$$

whose growth is essentially linear with n: the conditional differential entropy $h(V_{n+1}|V^n) = h(V^{n+1}) - h(V^n)$ is monotonically decreasing with

$$h(V_2|V_1) = \tfrac{3}{2}(\gamma + \psi(\tfrac{3}{2})) + \log_{\mathrm{e}} 4 = 2.306... \tag{18}$$

$$\lim_{n\to\infty} h(V_{n+1}|V^n) = \tfrac{1}{2}(1 + \gamma + \log_{\mathrm{e}}(4\pi)) = 2.054... \tag{19}$$

13. By the scaling law of differential entropy and its invariance to location, we obtain

$$h\Big(\mathbf{\Lambda}^{\frac{1}{2}} V^n + \boldsymbol{\mu}\Big) = h(V^n) + \frac{1}{2}\log|\det(\mathbf{\Lambda})|. \tag{20}$$

14. Invoking (A6), we obtain a closed-form formula for the differential entropy, in nats, of the generalized Cauchy distribution (9) as

$$h(V_{\beta,\rho}) = -\log_{\mathrm{e}} \kappa_{\beta,\rho} + \beta\, \mathbb{E}\big[\log_{\mathrm{e}}\big(1 + |V_{\beta,\rho}|^{\rho}\big)\big] \tag{21}$$

$$= -\log_{\mathrm{e}} \kappa_{\beta,\rho} + \beta\, \psi(\beta) - \beta\, \psi\Big(\beta - \tfrac{1}{\rho}\Big), \tag{22}$$

with $\kappa_{\beta,\rho}$ defined in (10).

15. The Rényi differential entropy of order $\alpha \in (0,1) \cup (1,\infty)$ of an absolutely continuous random variable with probability density function f_X is

$$h_\alpha(X) = \frac{1}{1-\alpha} \log \int f_X^\alpha(t)\, \mathrm{d}t. \tag{23}$$

For Cauchy random variables, we obtain, with the aid of (A12),

$$h_\alpha(\lambda V + \mu) = \log|\lambda| + h_\alpha(V), \tag{24}$$

$$h_\alpha(V) = \frac{\frac{1}{2}-\alpha}{1-\alpha} \log \pi + \frac{1}{1-\alpha} \log \frac{\Gamma(\alpha - \frac{1}{2})}{\Gamma(\alpha)}, \quad \alpha > \frac{1}{2}, \tag{25}$$

which is infinite for $\alpha \in (0, \frac{1}{2}]$, converges to $\log(4\pi)$ (cf. (12)) as $\alpha \to 1$, and to $\log \pi$, the reciprocal of the mode height, as $\alpha \to \infty$.

16. Invoking (A13), the Rényi differential entropy of order $\alpha \in \left(\frac{1}{\beta\rho}, 1\right) \cup (1,\infty)$ of the generalized Cauchy distribution (9) is

$$h_\alpha(V_{\beta,\rho}) = \frac{\alpha}{1-\alpha} \log \kappa_{\beta,\rho} + \frac{1}{1-\alpha} \log \frac{2\,\Gamma\left(\beta\alpha - \frac{1}{\rho}\right)\Gamma\left(\frac{1}{\rho}\right)}{\rho\,\Gamma(\beta\alpha)}. \tag{26}$$

4. The Shannon- and η-Transforms

In this section, we recall the definitions of two notions introduced in [10] for the unrelated purpose of expressing the asymptotic singular value distribution of large random matrices.

17. The Shannon transform of a nonnegative random variable X is the function $\mathcal{V}_X\colon [0,\infty) \to [0,\infty)$, defined by

$$\mathcal{V}_X(\theta) = \mathbb{E}\big[\log_{\mathrm{e}}(1+\theta X)\big]. \tag{27}$$

Unless $\mathcal{V}_X(\theta) = \infty$ for all $\theta > 0$ (e.g., if X has the log-Cauchy density $\frac{1}{\pi x}\frac{1}{1+\log^2 x}, x > 0$), or $\mathcal{V}_X(\theta) = 0$, $\theta \geq 0$, (which occurs if $X = 0$ a.s.), the Shannon transform is a strictly concave continuous function from $\mathcal{V}_X(0) = 0$, which grows without bound as $\theta \to \infty$.

18. If V is standard Cauchy, then (A4) results in

$$\mathcal{V}_{V^2}(\theta^2) = 2\log_{\mathrm{e}}(1+|\theta|), \tag{28}$$

and the handy relationship

$$\mathbb{E}\left[\log\left(\beta^2 + \lambda^2 V^2\right)\right] = 2\log(|\beta| + |\lambda|). \tag{29}$$

19. For the distribution in (9) with $(\beta,\rho) = (2,2)$, (A7) results in

$$\mathcal{V}_{V_{2,2}^2}(\theta^2) = 2\log_{\mathrm{e}}(1+|\theta|) - \frac{2|\theta|}{1+|\theta|}. \tag{30}$$

20. The η-transform $\eta_X\colon [0,\infty) \to (0,1]$ of a non-negative random variable is defined as the function

$$\eta_X(\theta) = \mathbb{E}\left[\frac{1}{1+\theta X}\right] = 1 - \theta\,\dot{\mathcal{V}}_X(\theta), \tag{31}$$

which is intimately related to the Cauchy–Stieltjes transform [11]. For example,

$$\eta_{V^2}(\theta^2) = \frac{1}{1+|\theta|}, \tag{32}$$

$$\eta_{V_{2,2}^2}(\theta^2) = \frac{1+2|\theta|}{(1+|\theta|)^2}. \tag{33}$$

5. Strength

The purpose of this section is to introduce an attribute which is particularly useful to compare random variables that do not have finite moments.

21. The *strength* $\varsigma(Z) \in [0, +\infty]$ of a real-valued random variable Z is defined as

$$\varsigma(Z) = \inf\left\{\varsigma > 0 \colon \mathbb{E}\left[\log\left(1+\frac{Z^2}{\varsigma^2}\right)\right] \le \log 4\right\}. \tag{34}$$

It follows that the only random variable with zero strength is $Z = 0$, almost surely. If the inequality in (34) is not satisfied for any $\varsigma > 0$, then $\varsigma(Z) = \infty$. Otherwise, $\varsigma(Z)$ is the unique positive solution $\varsigma > 0$ to

$$\mathbb{E}\left[\log\left(1+\frac{Z^2}{\varsigma^2}\right)\right] = \log 4. \tag{35}$$

If $\varsigma(Z) \le \varsigma$, then (35) holds with $\le$.

22. The set of probability measures whose strength is upper bounded by a given finite nonnegative constant,

$$\mathcal{A}_\varsigma = \{P_Z \colon \varsigma(Z) \le \varsigma\}, \tag{36}$$

is convex: The set $\mathcal{A}_0$ is a singleton as seen in Item 21, while, for $0 < \varsigma < \infty$, we can express (36) as

$$\mathcal{A}_\varsigma = \left\{P_Z \colon \mathbb{E}\left[\log\left(1+\frac{Z^2}{\varsigma^2}\right)\right] \le \log 4\right\}. \tag{37}$$

Therefore, if $P_{Z_0} \in \mathcal{A}_\varsigma$ and $P_{Z_1} \in \mathcal{A}_\varsigma$, we must have $\alpha P_{Z_1} + (1-\alpha)P_{Z_0} \in \mathcal{A}_\varsigma$.

23. The peculiar constant in the definition of strength is chosen so that if V is standard Cauchy, then its strength is $\varsigma(V) = 1$ because, in view of (29),

$$\mathbb{E}\left[\log\left(1+V^2\right)\right] = \log 4. \tag{38}$$

24. If $Z = k \in \mathbb{R}$, a.s., then its strength is

$$\varsigma(Z) = \frac{|k|}{\sqrt{3}}. \tag{39}$$

25. The left side of (35) is the *Shannon transform* of Z^2 evaluated at ς^{-2}, which is continuous in ς^2. If $\varsigma(Z) \in (0, \infty)$ then, (35) can be written as

$$\varsigma^2(Z) = \frac{1}{\mathcal{V}_{Z^2}^{-1}(\log_{\mathrm{e}} 4)}, \tag{40}$$

where, on the right side, we have denoted the functional inverse of the Shannon transform. Clearly, the square root of the right side of (40) cannot be expressed as

the expectation with respect to Z of any $\mathsf{b}\colon \mathbb{R} \to \mathbb{R}$ that does not depend on P_Z. Nevertheless, thanks to (37), (36) can be expressed as

$$\mathcal{A}_\varsigma = \left\{ P_Z \colon \mathbb{E}\left[\mathsf{b}_{\varsigma^2}(Z)\right] \le 1 \right\}, \text{ with } \mathsf{b}_{\varsigma^2}(x) = \log_4\left(1 + \frac{x^2}{\varsigma^2}\right). \tag{41}$$

26.

Theorem 1. *The strength of a real-valued random variable satisfies the following properties:*

(a)

$$\varsigma(\lambda Z) = |\lambda|\, \varsigma(Z). \tag{42}$$

(b)

$$\varsigma^2(Z) \le \frac{1}{3}\, \mathbb{E}[Z^2], \tag{43}$$

with equality if and only if $|Z|$ is deterministic.

(c) *If $0 < q < 2$, and $\|Z\|_q = \mathbb{E}^{\frac{1}{q}}[|Z|^q] < \infty$, then*

$$\varsigma(Z) \le \kappa_q^{\frac{1}{q}}\, \|Z\|_q, \quad \textit{with } \kappa_q = \max_{x>0} \frac{\log_4(1+x^2)}{x^q}. \tag{44}$$

(d) *If V is standard Cauchy, independent of X, then $\varsigma(X+V)$ is the solution to*

$$\mathcal{V}_{X^2}\left((\varsigma+1)^{-2}\right) = 2\log\frac{2}{1+\varsigma^{-1}}, \tag{45}$$

if it exists, otherwise, $\varsigma(X+V) = \infty$. Moreover, $\le$ holds in (45) if $\varsigma(X+V) \le \varsigma$.

(e)

$$2\log(2\min\{1, \varsigma(Z)\}) \le \mathbb{E}\left[\log\left(1+Z^2\right)\right] \le 2\log(2\max\{1, \varsigma(Z)\}). \tag{46}$$

(f) *If $0 < \varsigma(Z) < \infty$, then*

$$h(Z) = \log(4\pi\, \varsigma(Z)) - D(Z \,\|\, \varsigma(Z)V), \tag{47}$$

where V is standard Cauchy, and $D(X\,\|\,Y)$ stands for the relative entropy with reference probability measure P_Y and dominated measure P_X.

(g)

$$h(Z) < \infty \iff \varsigma(Z) < \infty \iff \mathbb{E}\left[\log\left(1+Z^2\right)\right] < \infty. \tag{48}$$

(h) *If V is standard Cauchy, then*

$$\varsigma(Z) < \infty \text{ and } h(Z) \in \mathbb{R} \iff D(Z \,\|\, \lambda\, V) < \infty, \textit{ for all } \lambda > 0. \tag{49}$$

(i) *The finiteness of strength is sufficient for the finiteness of the entropy of the integer part of the random variable, i.e.,*

$$H(\lfloor Z \rfloor) = \infty \implies \varsigma(Z) = \infty.$$

(j) *If $Z_n \to Z$ in L_q for any $q \in (0,1]$, then $\varsigma(Z_n) \to \varsigma(Z)$.*

(k)

$$Z_n \to 0 \text{ i.p.} \iff \mathbb{E}\left[\log\left(1+Z_n^2\right)\right] \to 0 \iff \varsigma(Z_n) \to 0. \tag{50}$$

(l) *If* $\varsigma(X_n) \to 0$, *then* $\varsigma(Z + X_n) \to \varsigma(Z)$.
(m) *If* $\varsigma(X_n) \to 0$, $\varsigma(Z) < \infty$ *and* Z *is independent of* X_n, *then* $h(Z + X_n) \to h(Z)$.

Proof. For the first three properties, it is clear that they are satisfied if $\varsigma(Z) = 0$, i.e., $Z = 0$ almost surely.

(a) If $\varsigma^2 \in (0, \infty)$ is the solution to (35), then $\lambda^2 \varsigma^2$ is a solution to (35) with λZ taking the role of Z. If (35) has no solution, neither does its version in which λZ takes the role of Z.

(b) Jensen's inequality applied to the left side of (35) results in $3\varsigma^2 \leq \mathbb{E}[Z^2]$. The strict concavity of $\log(1 + t)$ implies that equality holds if and only if Z^2 is deterministic. If (35) has no solution, the same reasoning implies that $\mathbb{E}[Z^2] = \infty$.

(c) First, it is easy to check that, for $q \in (0, 2)$, the function $f_q \colon (0, \infty) \to (0, \infty)$ given by $f_q(t) = t^{-q} \log_4(1 + t^2)$ attains its maximum κ_q at a unique point. Assume $\varsigma(Z) \in (0, \infty)$. Since $\kappa_q \, t^q \geq \log_4(1 + t^2)$ for all $t > 0$, letting $t = |Z|/\varsigma(Z)$ and taking expectations, (35) (choosing 4 as the logarithm base) results in

$$\frac{\kappa_q}{\varsigma^q(Z)} \mathbb{E}[|Z|^q] \geq 1, \tag{51}$$

which is the same as (44). If $\varsigma(Z) = \infty$, then $\infty = \mathbb{E}[\log(1 + Z^2)] \leq \kappa_q \, \mathbb{E}[|Z|^q]$.

(d) Invoking (A4) with $\alpha^2 = \varsigma^2 + x^2$ and $|\sin\beta| = \frac{\varsigma}{\sqrt{x^2+\varsigma^2}}$, we obtain

$$\mathbb{E}\left[\log\left(1 + \frac{(x+V)^2}{\varsigma^2}\right)\right] = \log\left(\frac{(1+\varsigma)^2 + x^2}{\varsigma^2}\right) \tag{52}$$

$$= \log\left(1 + \frac{x^2}{(1+\varsigma)^2}\right) - 2\log\frac{\varsigma}{\varsigma+1}. \tag{53}$$

Substituting x by X and averaging over X, the result follows from the definition of strength.

(e) The result holds trivially if either $\varsigma(Z) = 0$ or $\varsigma(Z) = \infty$. Otherwise, we simply rewrite (35) as

$$2\log(2\varsigma(Z)) = \mathbb{E}\Big[\log\Big(\varsigma^2(Z) + Z^2\Big)\Big], \tag{54}$$

and upper/lower bound the right side by $\mathbb{E}\big[\log\big(1 + Z^2\big)\big]$.

(f)

$$D(Z \,\|\, \varsigma(Z)V) = -h(Z) + \log(\varsigma(Z)\,\pi) + \mathbb{E}\left[\log\left(1 + \frac{Z^2}{\varsigma^2(Z)}\right)\right] \tag{55}$$

$$= \log(4\pi\,\varsigma(Z)) - h(Z), \tag{56}$$

where (55) and (56) follow from (2) and (35), respectively.

(g)

- If $\varsigma(Z) < \infty$, then $\mathbb{E}\big[\log\big(1 + Z^2\big)\big] < \infty$ and $h(Z) < \infty$ follow from (46) and (47), respectively.
- If $\mathbb{E}\big[\log\big(1 + Z^2\big)\big] < \infty$, the dominated convergence theorem implies

$$\lim_{\varsigma \to \infty} \mathbb{E}\left[\log\left(1 + \left(\frac{Z}{\varsigma}\right)^2\right)\right] = 0. \tag{57}$$

Excluding the case $Z = 0$ a.s. for which both $\mathbb{E}\big[\log(1+Z^2)\big]$ and $\varsigma(Z)$ are zero, we have

$$\lim_{\varsigma \downarrow 0} \mathbb{E}\left[\log\left(1+\left(\frac{Z}{\varsigma}\right)^2\right)\right] = \lim_{\varsigma \downarrow 0} \mathcal{V}_{Z^2}\left(\frac{1}{\varsigma^2}\right) = \infty. \tag{58}$$

Since (35) is continuous in ς, it must have a finite solution in view of (57) and (58).

(h) It is sufficient to assume $\lambda = 1$ for the condition on the right of (49) because the condition on the left holds if and only if it holds for αZ, for any $\alpha > 0$ and $D(\alpha Z \,\|\, \alpha V) = D(Z \,\|\, V)$. If $h(Z) < \infty$, then

$$D(Z \,\|\, V) = -h(Z) + \log \pi + \mathbb{E}\Big[\log\Big(1+Z^2\Big)\Big], \tag{59}$$

which is finite unless either $h(Z) = -\infty$ or $\mathbb{E}[\log(1+Z^2)] = \infty$. This establishes $\Longrightarrow$ in view of (48). To establish $\Longleftarrow$, it is enough to show that

$$D(Z \,\|\, V) < \infty \Longrightarrow \mathbb{E}\Big[\log\Big(1+Z^2\Big)\Big] < \infty, \tag{60}$$

in view of (48) and the fact that, according to (59), $h(Z) > -\infty$ if both $D(Z \,\|\, V)$ and $\mathbb{E}\big[\log(1+Z^2)\big]$ are finite. To show (60), we invoke the following variational representation of relative entropy (first noted by Kullback [12] for absolutely continuous random variables): If $P_Z \ll P_V$, then

$$D(Z \,\|\, V) = \max_{Q:\, Q \ll P_V} \mathbb{E}\left[\log \frac{\mathrm{d}Q}{\mathrm{d}P_V}(Z)\right], \tag{61}$$

attained only at $Q = P_Z$. Let Q be the absolutely continuous random variable with probability density function

$$q(x) = \frac{\log_{\mathrm{e}} 2}{4|x| \log_{\mathrm{e}}^2 |x|}\, 1\{|x| \geq 2\} + \frac{1}{8} 1\{|x| < 2\}. \tag{62}$$

Then,

$$\begin{aligned}
\infty > D(Z \,\|\, V) &> \mathbb{E}\left[\log \frac{q(Z)}{f_V(Z)}\right] && (63)\\
&= \mathbb{E}\left[1\{|Z| \geq 2\}\left(\log \frac{\pi \log_{\mathrm{e}} 2}{4} + \log\Big(1+Z^2\Big) - \log\Big(|Z| \log_{\mathrm{e}}^2 |Z|\Big)\right)\right] \\
&\quad + \mathbb{E}\left[1\{|Z| < 2\}\left(\log \frac{\pi}{8} + \log\Big(1+Z^2\Big)\right)\right] && (64)\\
&> \tfrac{1}{5}\mathbb{E}\Big[1\{Z \geq 2\} \log\Big(1+Z^2\Big)\Big] + \log \tfrac{5\pi}{8} && (65)\\
&\geq \tfrac{1}{5}\mathbb{E}\Big[\log\Big(1+Z^2\Big)\Big] + \tfrac{4}{5}\log 5 - \log \tfrac{8}{\pi}, && (66)
\end{aligned}$$

where (65) holds since

$$\tfrac{4}{5}\log(1+x^2) \geq -\log(\pi \log_{\mathrm{e}} 2) + 2 \log\log_{\mathrm{e}} |x| + \log |x|, \quad |x| > 2. \tag{67}$$

(i)

$$\begin{aligned}
\varsigma(Z) < \infty &\Longrightarrow \mathbb{E}\Big[\log(1+Z^2)\Big] < \infty && (68)\\
&\Longrightarrow \mathbb{E}[\log(1+|Z|)] < \infty && (69)\\
&\Longrightarrow H(\lfloor Z \rfloor) < \infty, && (70)
\end{aligned}$$

where (68)–(70) follow from (48), $\log(1+x^2) \le 2\log(1+|x|)$, and p. 3743 in [13], respectively.

(j) If $\varsigma(Z)=0$, then $Z=0$ a.e., and the result follows from (44). For all $(x,z)\in\mathbb{R}^2$,

$$\left|\log_{\mathrm{e}}\left(\frac{1+(x+z)^2}{1+z^2}\right)\right| \le \log_{\mathrm{e}}\left(1+\frac{1}{2}(x^2+|x|\sqrt{4+x^2})\right) \tag{71}$$

$$\le \frac{2}{q}|x|^q, \tag{72}$$

where (71) follows by maximizing the left side over $z\in\mathbb{R}$. Denote the difference between the right side and the left side of (72) by $f_q(x)$, an even function which satisfies $f_q(0)=0$, and

$$\dot{f}_q(x) = 2x^{q-1} - \frac{2}{\sqrt{4+x^2}} > 0, \quad x>0,\ 0<q\le 1. \tag{73}$$

Therefore, (72) follows. Assuming $0<\varsigma(Z)<\infty$, we have

$$\left|\mathbb{E}\left[\log(1+Z_n^2)\right] - \mathbb{E}\left[\log(1+Z^2)\right]\right| \le \mathbb{E}\left[\left|\log(1+Z_n^2)-\log(1+Z^2)\right|\right] \tag{74}$$

$$\le \frac{2}{q}\mathbb{E}[|Z_n-Z|^q]\,\log \mathrm{e}. \tag{75}$$

Now, because of the scaling property in (42), we may assume without loss of generality that $\varsigma(Z)=1$. Thus, (74) and (75) result in

$$\left|\mathbb{E}\left[\log(1+Z_n^2)\right] - \log 4\right| \le \frac{2}{q}\mathbb{E}[|Z_n-Z|^q]\,\log \mathrm{e}, \tag{76}$$

which requires that $\varsigma(Z_n)\to 1$, since, by assumption, the right side vanishes. Assume now that $\varsigma(Z)=\infty$, and therefore, $\mathbb{E}\left[\log(1+Z^2)\right]=\infty$. Inequality (75) remains valid in this case, implying that, as soon as the right side is finite (which it must be for all sufficiently large n), $\mathbb{E}\left[\log(1+Z_n^2)\right]=\infty$, and therefore, $\varsigma(Z_n)=\infty$ in view of (48).

(k)

1st $\Longleftarrow$ For any $\epsilon>0$, Markov's inequality results in

$$\mathbb{P}[|Z_n|>\epsilon] = \mathbb{P}\left[\log\left(1+Z_n^2\right) > \log\left(1+\epsilon^2\right)\right] \le \frac{\mathbb{E}\left[\log\left(1+Z_n^2\right)\right]}{\log(1+\epsilon^2)}. \tag{77}$$

$\Longrightarrow$ First, we show that, for any $\alpha>0$, we have

$$\mathbb{E}\left[\log\left(1+Z_n^2\right)\right]\to 0 \implies \mathbb{E}\left[\log\left(1+\alpha\, Z_n^2\right)\right]\to 0. \tag{78}$$

The case $0<\alpha<1$ is trivial. The case $\alpha>1$ follows because $\mathbb{E}\left[\log\left(1+Z_n^2\right)\right] \to 0$ implies

$$\mathbb{E}\left[\log\left(1+\alpha\, Z_n^2\right)\right] = \mathbb{E}\left[\log\left(1+(\alpha-1)\, Z_n^2\right)\right], \tag{79}$$

where $\ge$ is obvious, and $\le$ holds because

$$\log\left(1+\alpha\, t^2\right) = \log\left(1+t^2\right) + \log\left(1+(\alpha-1)\frac{t^2}{1+t^2}\right) \tag{80}$$

$$\le \log\left(1+t^2\right) + \log\left(1+(\alpha-1)\, t^2\right). \tag{81}$$

If $\varsigma(Z_n) = \infty$ infinitely often, so is $\mathbb{E}\left[\log\left(1+Z_n^2\right)\right]$ in view of (48). Assume that $\limsup \varsigma(Z_n) = \varsigma \in (0,\infty]$, and $\varsigma(Z_n)$ is finite for all sufficiently large. Then, there is a subsequence such that $\varsigma(Z_{n_i}) \to \varsigma$, and

$$\log 4 = \mathbb{E}\left[\log\left(1+\left(\frac{Z_{n_i}}{\varsigma(Z_{n_i})}\right)^2\right)\right] \le \mathbb{E}\left[\log\left(1+\left(\frac{Z_{n_i}}{\lambda}\right)^2\right)\right], \quad (82)$$

for all sufficiently large i and $\lambda < \varsigma$. Consequently, (78) implies that $\mathbb{E}\left[\log\left(1+Z_n^2\right)\right] \not\to 0$.

2nd $\Longleftarrow$ Suppose that $\mathbb{E}\left[\log\left(1+Z_n^2\right)\right] \not\to 0$. Therefore, there is a subsequence along which $\mathbb{E}\left[\log\left(1+Z_{n_i}^2\right)\right] > \eta > 0$. If $\eta \ge \log 4$, then $\varsigma(Z_{n_i}) > 1$ along the subsequence. Because of the continuity of the Shannon transform and the fact that it grows without bound as its argument goes to infinity (Item 25), if $0 < \eta < \log 4$, we can find $1 < \alpha < \infty$ such that $\mathbb{E}\left[\log\left(1+\alpha Z_{n_i}^2\right)\right] > \log 4$, which implies $\varsigma(Z_{n_i}) > \alpha^{-1/2}$. Therefore, $\varsigma(Z_n) \not\to 0$ as we wanted to show.

(l) We start by showing that

$$\mathbb{E}\left[\log\left(1+X_n^2\right)\right] \to 0 \iff \mathbb{E}[f(X_n)] \to 0, \quad (83)$$

where we have denoted the right side of (71) with arbitrary logarithm base by $f(x)$. Since $\dot{f}(x) = \frac{2\log \mathrm{e}}{\sqrt{4+x^2}}$, it is easy to verify that

$$0 \le f(x) - \log(1+x^2) \le \log\frac{4}{3}, \quad x \in \mathbb{R}, \quad (84)$$

where the lower and upper bounds are attained uniquely at $x = 0$ and $|x| = \frac{1}{\sqrt{2}}$, respectively. The lower bound results in $\Longleftarrow$ in (83). To show $\Longrightarrow$, decompose, for arbitrary $\epsilon > 0$,

$$\begin{aligned}
\mathbb{E}[f(X_n)] &= \mathbb{E}[f(X_n)1\{|X_n| < \epsilon\}] + \mathbb{E}[f(X_n)1\{|X_n| \ge \epsilon\}] && (85)\\
&\le f(\epsilon) + \mathbb{E}[f(X_n)1\{|X_n| \ge \epsilon\}] && (86)\\
&\le f(\epsilon) + A_\epsilon\, \mathbb{E}\left[\log\left(1+X_n^2\right)1\{|X_n| \ge \epsilon\}\right] && (87)\\
&\le f(\epsilon) + A_\epsilon\, \epsilon^3, && (88)
\end{aligned}$$

where

$$A_\epsilon = 1 + \frac{\log\frac{4}{3}}{\log(1+\epsilon^2)}, \quad (89)$$

(87) holds from the upper bound in (84), and the fact that (89) is decreasing in ϵ, and (88) holds for all sufficiently large n if $\mathbb{E}\left[\log\left(1+X_n^2\right)\right] \to 0$. Since the right side of (88) goes to 0 as $\epsilon \to 0$, (83) is established. Assume $0 < \varsigma(Z) < \infty$. From the linearity property (42), we have $\varsigma(Z+X_n) = \varsigma(Z)\cdot\varsigma(\bar{Z}+\bar{X}_n)$ with $\bar{Z} = \varsigma^{-1}(Z)\,Z$ and $\bar{X}_n = \varsigma^{-1}(Z)\,X_n$ which satisfies $\varsigma(\bar{X}_n) \to 0$. Therefore, we may restrict attention to $\varsigma(Z) = 1$ without loss of generality. Following (71) and (74), and abbreviating $Z_n = Z + X_n$, we obtain

$$\begin{aligned}
\left|\mathbb{E}\left[\log(1+Z_n^2)\right] - \log 4\right| &\le \mathbb{E}\left[\left|\log(1+Z_n^2) - \log(1+Z^2)\right|\right] && (90)\\
&\le \mathbb{E}[f(X_n)]. && (91)
\end{aligned}$$

Thus, the desired result follows in view of (50) and (83). To handle the case $\varsigma(Z) = \infty$, we use the same reasoning as in the proof of (i) since (83) remains valid in that case.

(m) If $\varsigma(Z) = 0$, then $Z = 0$ a.s., $h(Z) = -\infty$ and $h(X_n) \to -\infty$ in view of Part (f). Assume henceforth that $\varsigma(Z) > 0$. Since $h(Z + X_n) \geq h(Z)$, it suffices to show

$$\limsup_{n\to\infty} h(X_n + Z) \leq h(Z). \tag{92}$$

Under the assumptions, Part (l) guarantees that

$$\varsigma(X_n + Z) \to \varsigma(Z). \tag{93}$$

If V is a standard Cauchy random variable, then $\varsigma(Z + X_n)V \to \varsigma(Z)V$ in distribution as the characteristic function converges: $\mathrm{e}^{-\varsigma(Z+X_n)|t|} \to \mathrm{e}^{-\varsigma(Z)|t|}$ for all t. Analogously, according to Part (k), $Z + X_n \xrightarrow{D} Z$ since $X_n \to 0$ in probability. Since the strength of $X_n + Z$ is finite for all sufficiently large n, we may invoke (47) to express, for those n,

$$\begin{aligned} h(X_n + Z) - h(Z) = \log \frac{\varsigma(Z + X_n)}{\varsigma(Z)} \\ - D(Z + X_n \,\|\, \varsigma(Z + X_n)\, V) + D(Z \,\|\, \varsigma(Z)\, V). \end{aligned} \tag{94}$$

The lower semicontinuity of relative entropy under weak convergence (which, in turn, is a corollary to the Donsker–Varadhan [14,15] variational representation of relative entropy) results in

$$\liminf_{n\to\infty} D(Z + X_n \,\|\, \varsigma(Z + X_n)V) \geq D(Z \,\|\, \varsigma(Z)V), \tag{95}$$

because $Z + X_n \xrightarrow{D} Z$ and $\varsigma(Z + X_n)V \xrightarrow{D} \varsigma(Z)V$. Therefore, (92) follows from (94) and (95).

□

27. In view of (42) and Item 23, $\varsigma(\lambda V) = |\lambda|$ if V is standard Cauchy. Furthermore, if X_1 and X_2 are centered independent Cauchy random variables, then their sum is centered Cauchy with

$$\varsigma(X_1 + X_2) = \varsigma(X_1) + \varsigma(X_2). \tag{96}$$

More generally, it follows from Theorem 1-(d) that, if X_1 is centered Cauchy, and (96) holds for $X_2 = \alpha X$ and all $\alpha \in \mathbb{R}$, then X must be centered Cauchy. Invoking (52), we obtain

$$\varsigma(\lambda V + \mu) = \frac{|\lambda|}{3} + \frac{1}{3}\sqrt{4\lambda^2 + 3\mu^2}, \tag{97}$$

which is also valid for $\lambda = 0$ as we saw in Item 24.

28. If X is standard Gaussian, then $\varsigma^2(X) = 0.171085\ldots$, and $\varsigma^2(\sigma X) = \sigma^2\varsigma^2(X)$. Therefore, if X_1 and X_2 are zero-mean independent Gaussian random variables, then

$$\varsigma^2(X_1 + X_2) = \varsigma^2(X_1) + \varsigma^2(X_2). \tag{98}$$

Thus, in this case, $\varsigma(X_1 + X_2) < \varsigma(X_1) + \varsigma(X_2)$.

29. It follows from Theorem 1-(d) that, with X independent of standard Cauchy V, we obtain $\varsigma(X+V) > \varsigma(X) + \varsigma(V)$ whenever X is such that

$$\mathcal{V}_{X^2}\left((2+\varsigma(X))^{-2}\right) > 2\log_{\mathrm{e}}\left(1+\frac{\varsigma(X)}{2+\varsigma(X)}\right). \tag{99}$$

An example is the heavy-tailed probability density function

$$f_X(x) = \frac{1}{\pi}\frac{\log_4(1+x^2)}{1+x^2}, \tag{100}$$

for which $7.0158\ldots = \varsigma(X+V) > \varsigma(X) + \varsigma(V) = 6.8457\ldots$.

30. Using (A8), we can verify that, if X is zero-mean uniform with variance σ^2, then

$$\varsigma^2(X) = \frac{3}{c^2}\sigma^2 = 0.221618\ldots\sigma^2, \tag{101}$$

where c is the solution to $\log_{\mathrm{e}}(1+c^2) + \frac{2}{c}\arctan(c) = 2 + \log_{\mathrm{e}} 4$.

31. We say that $Z_n \to 0$ in strength if $\varsigma(Z_n) \to 0$. Parts (j) and (k) of Theorem 1 show that this convergence criterion is intermediate between the traditional in probability and L_q criteria. It is not equivalent to either one: If

$$Z_n = \begin{cases} 0, & \text{with probability } 1-\frac{1}{n}; \\ 2^n, & \text{with probability } \frac{1}{n}, \end{cases} \tag{102}$$

then $\varsigma(Z_n) \to 1$, while $Z_n \to 0$ in probability. If, instead, $Z_n = \frac{3}{2}^n$, with probability $\frac{1}{n}$, then $Z_n \to 0$ in strength, but not in L_q for any $0 < q$.

32. The assumption in Theorem 1-(m) that $X_n \to 0$ in strength cannot be weakened to convergence in probability. Suppose that X_n is absolutely continuous with probability density function

$$f_{X_n}(t) = \begin{cases} n-1, & t \in \left[0, \frac{1}{n}\right]; \\ 0, & t \in (-\infty, 0) \cup \left(\frac{1}{n}, 2\right); \\ \frac{1}{n}\frac{\log_{\mathrm{e}} 2}{t\log_{\mathrm{e}}^2 t}, & t \in [2, \infty). \end{cases} \tag{103}$$

We have $X_n \to 0$ in probability since, regardless of how small $\epsilon > 0$, $\mathbb{P}[X_n > \epsilon] = \frac{1}{n}$ for all $n \geq \frac{1}{\epsilon}$. Furthermore,

$$h(X_n + Z) \geq h(X_n) = \infty, \tag{104}$$

because (103) is the mixture of a uniform and an infinite differential entropy probability density function, and differential entropy is concave. We conclude that $h(X_n + Z) \not\to h(Z)$, since $h(Z) < \infty$.

33. The following result on the continuity of differential entropy is shown in [16]: if X and Z are independent, $\mathbb{E}[|Z|] < \infty$ and $\mathbb{E}[|X|] < \infty$, then

$$\lim_{\epsilon \downarrow 0} h(\epsilon X + Z) = h(Z). \tag{105}$$

This result is weaker than Theorem 1-(m) because finite first absolute moment implies finite strength as we saw in (44), and $\epsilon X \to 0$ in L_1 if $\epsilon \to 0$, and therefore, it vanishes in strength too.

34. If Z and V are centered and standard Cauchy, respectively, then $\min_\lambda D(Z \,\|\, \lambda V)$ is achieved by $\lambda = \varsigma(Z)$. Otherwise, in general, this does not hold. Since $D(Z \,\|\, \lambda V) = \mathcal{V}_{Z^2}\left(\lambda^{-2}\right) - h(Z) + \log_{\mathrm{e}}(\pi\lambda)$, the minimum is attained at the solution to

$$\eta_{Z^2}\left(\frac{1}{\lambda_*^2}\right) = \frac{1}{2}, \tag{106}$$

where we have used the η-transform in (31). If $Z = V_{2,2}$, recalling (33), (106), results in $\lambda_* = \sqrt{2} - 1$, while $\varsigma(V_{2,2}) = 0.302\ldots$.

35. Using (28) and the concavity of $\log(1 + x)$, we can verify that

$$\varsigma(X_\alpha) \le \alpha\varsigma(X_1) + (1-\alpha)\varsigma(X_0), \quad X_\alpha \sim \alpha P_{X_1} + (1-\alpha)P_{X_0}, \tag{107}$$

if X_0 and X_1 are centered Cauchy, or, more generally, if $X_0 = \lambda_0 X$, $X_1 = \lambda_1 X$ and $\mathcal{V}_{X^2}(\theta^2)$ is concave on θ. Not only is this property not satisfied if $X = 1$ but (107) need not hold in that case, as we can verify numerically for $\alpha = 0.1$, $\lambda_1 = 1$ and $\lambda_0 > 20$.

6. Maximization of Differential Entropy

36. Among random variables with a given second moment (resp. first absolute moment), differential entropy is maximized by the zero-mean Gaussian (resp. Laplace) distribution. More generally, among random variables with a given p-absolute moment μ, differential entropy is maximized by the parameter-p Subbotin (or generalized normal) distribution with p-absolute moment μ [17]

$$f_X(x) = \frac{p^{1-\frac{1}{p}}}{2\,\Gamma(\frac{1}{p})\,\mu^{\frac{1}{p}}} \mathrm{e}^{-\frac{|x|^p}{p\,\mu}}, \quad x \in \mathbb{R}. \tag{108}$$

Among nonnegative random variables with a given mean, differential entropy is maximized by the exponential distribution. In those well-known solutions, the cost function is an affine function of the negative logarithm of the maximal differential entropy probability density function. Is there a cost function such that, among all random variables with a given expected cost, the Cauchy distribution is the maximal differential entropy solution? To answer this question, we adopt a more general viewpoint. Consider the following result, whose special case $\rho = 2$ was solved in [18] using convex optimization:

Theorem 2. *Fix $\rho > 0$ and $\theta > 0$.*

$$\max_{Z\colon \mathbb{E}\left[\log_{\mathrm{e}}(1+|Z|^\rho)\right] \le \theta} h(Z) = h(V_{\beta,\rho}), \tag{109}$$

where $V_{\beta,\rho}$ is defined in (9), the right side of (109) is given in (22), and $\beta > \rho^{-1}$ is the solution to

$$\theta = \psi(\beta) - \psi\left(\beta - \tfrac{1}{\rho}\right). \tag{110}$$

Therefore, the standard Cauchy distribution is the maximal differential entropy distribution provided that $\rho = 2$ and $\theta = \log_{\mathrm{e}} 4$.

Proof.

(a) For every $\rho > 0$ and $\theta > 0$, there is a unique $\beta > \rho^{-1}$ that satisfies (110) because the function of β on the right side is strictly monotonically decreasing, grows without bound as $\beta \downarrow \frac{1}{\rho}$, and goes to zero as $\beta \to \infty$.

(b) For any Z which satisfies $\mathbb{E}\big[\log_{\mathrm{e}}(1+|Z|^{\rho})\big] \leq \theta$, its relative entropy, in nats, with respect to $V_{\beta,\rho}$ is

$$\begin{aligned} D(Z \,\|\, V_{\beta,\rho}) &= -h(Z) - \log_{\mathrm{e}} \kappa_{\beta,\rho} + \beta\, \mathbb{E}\big[\log_{\mathrm{e}}(1+|Z|^{\rho})\big] && (111)\\ &\leq -h(Z) - \log_{\mathrm{e}} \kappa_{\beta,\rho} + \beta\, \theta && (112)\\ &= -h(Z) - \log_{\mathrm{e}} \kappa_{\beta,\rho} + \beta\, \psi(\beta) - \beta\, \psi\Big(\beta - \tfrac{1}{\rho}\Big) && (113)\\ &= h(V_{\beta,\rho}) - h(Z), && (114) \end{aligned}$$

where (113) and (114) follow from (110) and (22), respectively. Since relative entropy is nonnegative, and zero only if both measures are identical, not only does (2) hold but any random variable other than $V_{\beta,\rho}$ achieves strictly lower differential entropy.

□

37. An unfortunate consequence stemming from Theorem 2 is that, while we were able to find out a cost function such that the Cauchy distribution is the maximal differential entropy distribution under an average cost constraint, this holds only for a specific value of the constraint. This behavior is quite different from the classical cases discussed in Item 36 for which the solution is, modulo scale, the same regardless of the value of the cost constraint. As we see next, this deficiency is overcome by the notion of strength introduced in Section 5.

38.

Theorem 3. *Strength constraint. The differential entropy of a real-valued random variable with strength $\varsigma(Z)$ is upper bounded by*

$$h(Z) \leq \log(4\pi\, \varsigma(Z)). \tag{115}$$

If $0 < \varsigma(Z) < \infty$, equality holds if and only if Z has a centered Cauchy density, i.e., $Z = \lambda V$ for some $\lambda > 0$.

Proof.

(a) If Z is not an absolutely continuous random variable, or more generally, $h(Z) = -\infty$ such as in the case $\varsigma(Z) = 0$ in which $Z = 0$ with probability one, then (115) is trivially satisfied.

(b) If $0 < \varsigma(Z) < \infty$ and $h(Z) > -\infty$, then we invoke (47) to conclude that not only does (115) hold, but it is satisfied with equality if and only if $Z = \varsigma(Z)V$.

□

39. The entropy power of a random variable Z is the variance of a Gaussian random variable whose differential entropy is $h(Z)$, i.e.,

$$N(Z) = \frac{1}{2\pi \mathrm{e}} \exp(2\, h(Z)). \tag{116}$$

While the power of a Cauchy random variable is infinite, its entropy power is given by

$$N(\lambda V + \mu) = \frac{1}{2\pi \mathrm{e}} \exp(2\, h(\lambda V + \mu)) = \frac{8\, \pi\, \lambda^2}{\mathrm{e}}. \tag{117}$$

In the same spirit as the definition of entropy power, Theorem 3 suggests the definition of $N_{\mathsf{C}}(Z)$, the *entropy strength* of Z, as the strength of a centered Cauchy random variable whose differential entropy is $h(Z)$, i.e., $h(Z) = h(N_{\mathsf{C}}(Z)V)$. Therefore,

$$N_{\mathsf{C}}(Z) = \frac{1}{4\pi}\exp(h(Z)) \tag{118}$$

$$= \varsigma(Z)\,\exp(-D(Z\,\|\,\varsigma(Z)V)) \tag{119}$$

$$\leq \varsigma(Z), \tag{120}$$

where (119) follows from (56), and (120) holds with equality if and only if Z is centered Cauchy. Note that, for all $(\alpha, \mu) \in \mathbb{R}^2$,

$$N_{\mathsf{C}}(\alpha\, Z + \mu) = |\alpha|\, N_{\mathsf{C}}(Z). \tag{121}$$

Comparing (116) and (118), we see that entropy power is simply a scaled version of the entropy strength squared,

$$N(Z) = \frac{8\pi}{\mathrm{e}} N_{\mathsf{C}}^2(Z). \tag{122}$$

The entropy power inequality (e.g., [19,20]) states that, if X_1 and X_2 are independent real-valued random variables, then

$$N(X_1 + X_2) \geq N(X_1) + N(X_2), \tag{123}$$

regardless of whether they have moments. According to (122), we may rewrite the entropy power inequality (123) replacing each entropy power by the corresponding squared entropy strength. Therefore, the squared entropy strength of the sum of independent random variables satisfies

$$N_{\mathsf{C}}^2(X_1 + X_2) \geq N_{\mathsf{C}}^2(X_1) + N_{\mathsf{C}}^2(X_2). \tag{124}$$

It is well-known that equality holds in (123), and hence (124), if and only if both random variables are Gaussian. Indeed, if X_1 and X_2 are centered Cauchy with respective strengths $\varsigma_1 > 0$ and $\varsigma_2 > 0$, then (124) becomes $(\varsigma_1 + \varsigma_2)^2 > \varsigma_1^2 + \varsigma_2^2$.

40. Theorem 3 implies that any random variable with infinite differential entropy has infinite strength. There are indeed random variables with finite differential entropy and infinite strength. For example, let $Z \in [2, \infty)$ be an absolutely continuous random variable with probability density function

$$f_Z(t) = \begin{cases} 0.473991\ldots \log_{\mathrm{e}}^{-2} n, & t \in \left[n, n + \frac{1}{n}\right],\ n \in \{2, 3, \ldots\}; \\ 0, & \text{elsewhere.} \end{cases} \tag{125}$$

Then, $h(Z) = 1.99258\ldots$ nats, while the entropy of the quantized version as well as the strength satisfy $H(\lfloor Z \rfloor) = \infty = \varsigma(Z)$.

41. With the same approach, we may generalize Theorem 3 to encompass the full slew of the generalized Cauchy distributions in (9). To that end, fix $\rho > 0$ and define the (ρ, θ)-*strength* of a random variable as

$$\varsigma_{\rho,\theta}(Z) = \inf\left\{\varsigma > 0 \colon \mathbb{E}\left[\log_{\mathrm{e}}\left(1 + \left|\frac{Z}{\varsigma}\right|^{\rho}\right)\right] \leq \theta\right\}. \tag{126}$$

Therefore, $\varsigma_{\rho,\theta}(Z) = \varsigma(Z)$ for $(\rho,\theta) = (2, \log_{\mathrm{e}} 4)$, and if (β,ρ,θ) satisfy (110), then $\varsigma_{\rho,\theta}(V_{\beta,\rho}) = 1$. As in Item 25, if $\varsigma_{\rho,\theta}(Z) \in (0,\infty)$, we have

$$\varsigma^{\rho}_{\rho,\theta}(Z) = \frac{1}{\mathcal{V}^{-1}_{|Z|^\rho}(\theta)}. \tag{127}$$

42.

Theorem 4. *Generalized strength constraint. Fix $\rho > 0$ and $\theta > 0$. The differential entropy of a real-valued random variable with (ρ,θ)-strength $\varsigma_{\rho,\theta}(Z)$ is upper bounded by*

$$h(Z) \le \log\big(\varsigma_{\rho,\theta}(Z)\big) + h(V_{\beta,\rho}), \tag{128}$$

where β is given by the solution to (110), $V_{\beta,\rho}$ has the generalized Cauchy density (9), and $h(V_{\beta,\rho})$ is given in (22). If $\varsigma_{\rho,\theta}(Z) < \infty$, equality holds if and only if Z is a constant times $V_{\beta,\rho}$.

Proof. As with Theorem 3, in the proof, we may assume $0 < \varsigma_{\rho,\theta}(Z) < \infty$ to avoid trivialities. Then,

$$\mathbb{E}\left[\log_{\mathrm{e}}\left(1 + \left|\frac{Z}{\varsigma_{\rho,\theta}(Z)}\right|^{\rho}\right)\right] = \theta, \tag{129}$$

and, in nats,

$$\begin{aligned}
D\left(\frac{Z}{\varsigma_{\rho,\theta}(Z)}\,\middle\|\, V_{\beta,\rho}\right) &= -h(Z) - \log_{\mathrm{e}}\frac{\kappa_{\beta,\rho}}{\varsigma_{\rho,\theta}(Z)} + \beta\,\mathbb{E}\left[\log_{\mathrm{e}}\left(1 + \left|\frac{Z}{\varsigma_{\rho,\theta}(Z)}\right|^{\rho}\right)\right] && (130)\\
&= -h(Z) - \log_{\mathrm{e}}\frac{\kappa_{\beta,\rho}}{\varsigma_{\rho,\theta}(Z)} + \beta\,\theta && (131)\\
&= -h(Z) - \log_{\mathrm{e}}\frac{\kappa_{\beta,\rho}}{\varsigma_{\rho,\theta}(Z)} + \beta\,\psi(\beta) - \beta\,\psi\Big(\beta - \tfrac{1}{\rho}\Big) && (132)\\
&= -h(Z) + \log_{\mathrm{e}}\big(\varsigma_{\rho,\theta}(Z)\big) + h(V_{\beta,\rho}), && (133)
\end{aligned}$$

where (130), (131), (132), and (133) follow from (9), (129), (110), and (22), respectively. □

43. In the multivariate case, we may find a simple upper bound on differential entropy based on the strength of the norm of the random vector.

Theorem 5. *The differential entropy of a random vector Z^n is upper bounded by*

$$h(Z^n) \le n\,\log(\varsigma(\|Z^n\|)) + \frac{n+1}{2}\log(4\pi) - \log\Gamma\left(\frac{n+1}{2}\right). \tag{134}$$

Proof. As in the proof of Theorem 3, we may assume that $0 < \varsigma(\|Z^n\|) < \infty$. As usual, V^n denotes the standard spherical multivariate Cauchy density in (6). Since for $\alpha \neq 0$, $f_{\alpha V^n}(x^n) = |\alpha|^{-n} f_{V^n}(\alpha^{-1}x^n)$, we have

$$\begin{aligned}
D(Z^n \,\|\, \varsigma(\|Z^n\|)\, V^n) &= -h(Z^n) - \mathbb{E}\Big[\log f_{\varsigma(\|Z^n\|)V^n}(Z^n)\Big] && (135)\\
&= -h(Z^n) + n\,\log(\varsigma(\|Z^n\|)) - \log\frac{\Gamma\left(\frac{n+1}{2}\right)}{\pi^{\frac{n+1}{2}}} + \frac{n+1}{2}\mathbb{E}\left[\log\left(1 + \frac{\|Z^n\|^2}{\varsigma^2(\|Z^n\|)}\right)\right] && (136)\\
&= -h(Z^n) + n\,\log(\varsigma(\|Z^n\|)) - \log\Gamma\left(\frac{n+1}{2}\right) + \frac{n+1}{2}\log(4\pi), && (137)
\end{aligned}$$

where (136) and (137) follow from (6) and the definition of strength, respectively. □

For $n = 1$, Theorem 5 becomes the bound in (115). For $n = 2, 3, \ldots$, the right side of (15) is greater than $\log_{\mathrm{e}} 4$, and, therefore, $\varsigma(\|Z^n\|) > 1$. Consequently, in the multivariate case, there is no Z^n such that (134) is tight.

44. To obtain a full generalization of Theorem 3 in the multivariate case, it is advisable to define the strength of a random n-vector as

$$\varsigma(Z^n) = \inf\left\{\varsigma > 0 \colon -\mathbb{E}\left[\log f_{V^n}\left(\varsigma^{-1} Z^n\right)\right] \le h(V^n)\right\} \tag{138}$$

$$= \varsigma_{2,\theta_n}(\|Z^n\|) \tag{139}$$

for $\theta_n = \psi\left(\frac{n+1}{2}\right) + \gamma + \log_{\mathrm{e}} 4$. To verify (139), note (15)–(17). Notice that $\varsigma(\lambda V^n) = |\lambda|$ and for $n = 1$, (138) is equal to (34). The following result provides a maximal differential entropy characterization of the standard spherical multivariate Cauchy density.

Theorem 6. *Let V^n have the standard multivariate Cauchy density (6), Then,*

$$h(Z^n) \le n \, \log \varsigma(Z^n) + h(V^n), \tag{140}$$

where $h(V^n)$ is given in (17). If $0 < \varsigma(Z^n) < \infty$, equality holds in (140) if and only if $Z^n = \lambda V^n$ for some $\lambda \neq 0$.

Proof. Assume $0 < \varsigma(Z^n) < \infty$. Then,

$$D\left(\frac{Z^n}{\varsigma(Z^n)} \,\middle\|\, V^n\right) = -h(Z^n) + n \log \varsigma(Z^n) - \mathbb{E}\left[\log f_{V^n}\left(\varsigma^{-1}(Z^n) Z^n\right)\right] \tag{141}$$

$$= -h(Z^n) + n \log \varsigma(Z^n) + h(V^n) \tag{142}$$

in view of (138). Hence, the difference between right and left sides of (140) is equal to zero if and only if $Z^n = \lambda V^n$ for some $\lambda \neq 0$; otherwise, it is positive. □

7. Relative Information

45. For probability measures P and Q on the same measurable space $(\mathcal{A}, \mathscr{F})$, such that $P \ll Q$, the logarithm of their Radon–Nikodym derivative is the *relative information* denoted by

$$\imath_{P\|Q}(x) = \log \frac{\mathrm{d}P}{\mathrm{d}Q}(x). \tag{143}$$

46. As usual, we may employ the notation $\imath_{X\|Y}(x)$ to denote $\imath_{P_X\|P_Y}(x)$. The distributions of the random variables $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$ are referred to as *relative information spectra* (e.g., [21]). It can be shown that there is a one-to-one correspondence between the cumulative distributions of $\imath_{X\|Y}(X)$ and $\imath_{X\|Y}(Y)$. For example, if they are absolutely continuous random variables with respective probability density functions $f_{X\|Y}$ and $\bar{f}_{X\|Y}$, then

$$f_{X\|Y}(\alpha) = \exp(\alpha) \, \bar{f}_{X\|Y}(\alpha), \quad \alpha \in \mathbb{R}. \tag{144}$$

Obviously, the distributions of $\imath_{X\|Y}(X)$ and $\frac{\mathrm{d}P_X}{\mathrm{d}P_Y}(X)$ determine each other. One caveat is that relative information may take the value $-\infty$. It can be shown that

$$\mathbb{P}[\imath_{X\|Y}(X) = -\infty] = 0, \tag{145}$$

$$\mathbb{P}[\imath_{X\|Y}(Y) = -\infty] = 1 - \mathbb{E}\left[\exp(-\imath_{X\|Y}(X))\right]. \tag{146}$$

47. The information spectra determine all measures of the distance between the respective probability measures of interest (e.g., [22,23]), including f-divergences and Rényi divergences. For example, the *relative entropy* (or Kullback–Leibler divergence) of the dominated measure P with respect to the reference measure Q is the average of the relative information when the argument is distributed according to P, i.e., $D(X\|Y) = \mathbb{E}[\imath_{X\|Y}(X)]$. If $P \not\ll Q$, then $D(P\|Q) = \infty$.
48. The information spectra also determine the fundamental trade-off in hypothesis testing. Let $\alpha_\nu(P_1, P_0)$ denote the minimal probability of deciding P_0 when P_1 is true subject to the constraint that the probability of deciding P_1 when P_0 is true is no larger than ν. A consequence of the Neyman–Pearson lemma is

$$\alpha_\nu(P_1, P_0) = \min_{\gamma \in \mathbb{R}} \Big\{ \mathbb{P}\Big[\imath_{P_1\|P_0}(Y_1) \le \gamma\Big] - \exp(\gamma)\Big(\nu - \mathbb{P}\Big[\imath_{P_1\|P_0}(Y_0) > \gamma\Big]\Big)\Big\}, \tag{147}$$

where $Y_0 \sim P_0$ and $Y_1 \sim P_1$.
49. Cauchy distributions are absolutely continuous with respect to each other and, in view of (2),

$$\imath_{\lambda_1 V+\mu_1\|\lambda_0 V+\mu_0}(x) = \log\frac{|\lambda_1|}{|\lambda_0|} + \log\frac{(x-\mu_0)^2+\lambda_0^2}{(x-\mu_1)^2+\lambda_1^2}. \tag{148}$$

50. The following result, proved in Item 58, shows that the relative information spectrum corresponding to Cauchy distributions with respective scale/locations (λ_1, μ_1) and (λ_0, μ_0) depends on the four parameters through the single scalar

$$\zeta(\lambda_1, \mu_1, \lambda_0, \mu_0) = \frac{\lambda_1^2 + \lambda_0^2 + (\mu_1-\mu_0)^2}{2|\lambda_0\lambda_1|} \ge 1, \tag{149}$$

where equality holds if and only if $(\lambda_1, \mu_1) = (\lambda_0, \mu_0)$.

Theorem 7. *Suppose that $\lambda_1\lambda_0 \neq 0$, and V is standard Cauchy. Denote*

$$Z = \frac{\mathrm{d}P_{\lambda_1 V+\mu_1}}{\mathrm{d}P_{\lambda_0 V+\mu_0}}(\lambda_1 V + \mu_1). \tag{150}$$

Then,

(a)

$$\mathbb{E}[Z] = \zeta(\lambda_1, \mu_1, \lambda_0, \mu_0), \tag{151}$$

(b) Z has the same distribution as the random variable

$$\zeta + \sqrt{\zeta^2 - 1}\,\cos\Theta, \tag{152}$$

where Θ is uniformly distributed on $[-\pi, \pi]$ and $\zeta = \zeta(\lambda_1, \mu_1, \lambda_0, \mu_0)$. Therefore, the probability density function of Z is

$$f_Z(z) = \frac{1}{\pi}\frac{1}{\sqrt{\zeta^2 - 1 - (z-\zeta)^2}}, \tag{153}$$

on the interval $0 < \zeta - \sqrt{\zeta^2-1} < z < \zeta + \sqrt{\zeta^2-1}$.

51. The indefinite integral (e.g., see 2.261 in [24])

$$\int \frac{\mathrm{d}x}{\sqrt{2\zeta x - x^2 - 1}} = \arcsin\frac{x-\zeta}{\sqrt{\zeta^2-1}} \tag{154}$$

results, with $X_i = \lambda_i V + \mu_i$, $i = 0, 1$, in

$$\mathbb{P}[\imath_{X_1\|X_0}(X_1) \le \log t] = \begin{cases} 1, & \zeta + \sqrt{\zeta^2 - 1} \le t; \\ \frac{1}{2} + \frac{1}{\pi} \arcsin \frac{t-\zeta}{\sqrt{\zeta^2-1}}, & \zeta - \sqrt{\zeta^2 - 1} < t < \zeta + \sqrt{\zeta^2 - 1}; \\ 0, & 0 < t \le \zeta - \sqrt{\zeta^2 - 1}. \end{cases} \tag{155}$$

52. For future use, note that the endpoints of the support of (153) are their respective reciprocals. Furthermore,

$$f_Z\left(\frac{1}{z}\right) = z\, f_Z(z), \tag{156}$$

which implies

$$f_{\frac{1}{Z}}(z) = \frac{1}{z} f_Z(z). \tag{157}$$

8. Equivalent Pairs of Probability Measures

53. Suppose that P_1 and Q_1 are probability measures on $(\mathcal{A}_1, \mathscr{F}_1)$ such that $P_1 \ll Q_1$ and P_2 and Q_2 are probability measures on $(\mathcal{A}_2, \mathscr{F}_2)$ such that $P_2 \ll Q_2$. We say that (P_1, Q_1) and (P_2, Q_2) are *equivalent pairs*, and write $(P_1, Q_1) \equiv (P_2, Q_2)$, if the cumulative distribution functions of $\imath_{P_1\|Q_1}(X_1)$ and $\imath_{P_2\|Q_2}(X_2)$ are identical with $X_1 \sim P_1$ and $X_2 \sim P_2$. Naturally, $\equiv$ is an equivalence relationship. Because of the one-to-one correspondence indicated in Item 46, the definition of equivalent pairs does not change if we require equality of the information spectra under the dominated measure, i.e., that $\imath_{P_1\|Q_1}(Y_1)$ and $\imath_{P_2\|Q_2}(Y_2)$ be equally distributed $Y_1 \sim Q_1$ and $Y_2 \sim Q_2$. Obviously, the requirement that the information spectra coincide is the same as requiring that the distributions of $\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(Y_1)$ and $\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(Y_2)$ are equal. As in Item 46, we also employ the notation $(X_1, Y_1) \equiv (X_2, Y_2)$ to indicate $(P_1, Q_1) \equiv (P_2, Q_2)$ if $X_1 \sim P_1$, $X_2 \sim P_2$, $Y_1 \sim Q_1$, and $Y_2 \sim Q_2$.
54. Suppose that the output probability measures of a certain (random or deterministic) transformation are Q_0 and Q_1 when the input is distributed according to P_0 and P_1, respectively. If $(P_0, P_1) \equiv (Q_0, Q_1)$, then the transformation is a sufficient statistic for deciding between P_0 and P_1 (i.e., the case of a binary parameter).
55. If $(\mathcal{A}, \mathscr{F})$ is a measurable space on which the probability measures $P_{X_1} \ll P_{X_2}$ are defined, and $\phi\colon \mathcal{A} \to \mathcal{B}$ is a $(\mathscr{F}, \mathscr{G})$-measurable *injective* function, then $P_{\phi(X_1)} \ll P_{\phi(X_2)}$ are probability measures on $(\mathcal{B}, \mathscr{G})$ and

$$\imath_{X_1\|X_2}(x) = \imath_{\phi(X_1)\|\phi(X_2)}(\phi(x)). \tag{158}$$

Consequently, $(X_1, X_2) \equiv (\phi(X_1), \phi(X_2))$.
56. The most important special case of Item 55 is an affine transformation of an arbitrary real-valued random variable X, which enables the reduction of four-parameter problems into two-parameter problems: for all $(\lambda_2, \mu_1, \mu_2) \in \mathbb{R}^3$ and $\lambda_1 \neq 0$,

$$(\lambda_1 X + \mu_1, \lambda_2 X + \mu_2) \equiv (X, \lambda X + \mu), \tag{159}$$

with

$$\lambda = \frac{\lambda_2}{\lambda_1} \text{ and } \mu = \frac{\mu_2 - \mu_1}{\lambda_1}, \tag{160}$$

by choosing the affine function $\phi(x) = \frac{x - \mu_1}{\lambda_1}$.

57.

Theorem 8. *If $X^n \in \mathbb{R}^n$ is an even random vector, i.e., $P_{X^n} = P_{-X^n}$, then*

$$(X^n + \boldsymbol{\mu}_1, X^n + \boldsymbol{\mu}_2) \equiv (X^n + \boldsymbol{\mu}_3, X^n + \boldsymbol{\mu}_4), \tag{161}$$

whenever $|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| = |\boldsymbol{\mu}_3 - \boldsymbol{\mu}_4|$.

Proof.

(a) If $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 - \boldsymbol{\mu}_4$, then (161) holds even if X^n is not even because the function $\mathbf{x} - \boldsymbol{\mu}$ is injective, in particular, with $\boldsymbol{\mu} = \boldsymbol{\mu}_3 - \boldsymbol{\mu}_1 = \boldsymbol{\mu}_4 - \boldsymbol{\mu}_2$.

(b) If $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}_4 - \boldsymbol{\mu}_3$, then

$$\begin{aligned}
(X^n + \boldsymbol{\mu}_1, X^n + \boldsymbol{\mu}_2) &\equiv (X^n, X^n + \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) && (162)\\
&\equiv (X^n, X^n + \boldsymbol{\mu}_3 - \boldsymbol{\mu}_4) && (163)\\
&\equiv (-X^n + \boldsymbol{\mu}_3 - \boldsymbol{\mu}_4, -X^n) && (164)\\
&\equiv (X^n + \boldsymbol{\mu}_3 - \boldsymbol{\mu}_4, X^n) && (165)\\
&\equiv (X^n + \boldsymbol{\mu}_3, X^n + \boldsymbol{\mu}_4), && (166)
\end{aligned}$$

where (162) and (166) follow from Part (a), (164) follows because $-\mathbf{x} + \boldsymbol{\mu}_3 - \boldsymbol{\mu}_4$ is injective, and (165) holds because X^n is even.

□

58. We now proceed to prove Theorem 7.

Proof. Since λV and $-\lambda V$ have identical distributions, we may assume for convenience that $\lambda_1 > 0$ and $\lambda_0 > 0$. Furthermore, capitalizing on Item 56, we may assume $\lambda_1 = 1$, $\mu_1 = 0$, $\lambda_0 = \lambda$, and $\mu_0 = \mu$, and then recover the general result letting $\lambda = \frac{\lambda_0}{\lambda_1}$ and $\mu = \frac{\mu_0 - \mu_1}{\lambda_1}$. Invoking (A9) and (A10), we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathrm{d}P_V}{\mathrm{d}P_{\lambda V+\mu}}(V)\right] &= \frac{1}{\lambda}\mathbb{E}\left[\frac{(V-\mu)^2 + \lambda^2}{V^2+1}\right] && (167)\\
&= \frac{1}{\pi\,\lambda}\int_{-\infty}^{\infty} \frac{(t-\mu)^2 + \lambda^2}{(t^2+1)^2}\,\mathrm{d}t && (168)\\
&= \frac{\lambda^2 + \mu^2 + 1}{2\,\lambda}, && (169)
\end{aligned}$$

and we can verify that we recover (151) through the aforementioned substitution. Once we have obtained the expectation of $Z = \frac{\mathrm{d}P_V}{\mathrm{d}P_{\lambda V+\mu}}(V)$, we proceed to determine its distribution. Denoting the right side of (169) by ζ, we have

$$\begin{aligned}
Z - \mathbb{E}[Z] &= \frac{1}{\lambda}\frac{\lambda^2 + (V-\mu)^2}{1+V^2} - \zeta && (170)\\
&= \frac{1}{2\lambda}\frac{(1-\lambda^2-\mu^2)(V^2-1) - 4\mu V}{1+V^2} && (171)\\
&= \frac{1}{2\lambda}(1-\lambda^2-\mu^2)(\sin^2\Theta - \cos^2\Theta) - 4\mu\sin\Theta\cos\Theta && (172)\\
&= \frac{1}{2\lambda}\Big((\lambda^2+\mu^2-1)\cos 2\Theta - 2\mu\sin 2\Theta\Big) && (173)\\
&= \frac{1}{2\lambda}\sqrt{(\lambda^2+\mu^2-1)^2 + 4\mu^2}\,\cos\big(2\Theta + \phi_{\lambda,\mu}\big) && (174)\\
&= \sqrt{\zeta^2 - 1}\,\cos\big(2\Theta + \phi_{\lambda,\mu}\big), && (175)
\end{aligned}$$

where Θ is uniformly distributed on $[-\pi, \pi]$. We have substituted $V = \tan\Theta$ (see Item 4) in (172), and invoked elementary trigonometric identities in (173) and (174). Since the phase in (175) does not affect it, the distribution of Z is indeed as claimed in (152), and (153) follows because the probability density function of $\cos\Theta$ is

$$f_{\cos\Theta}(t) = \frac{1}{\pi}\frac{1}{\sqrt{1-t^2}}, \quad |t| < 1. \tag{176}$$

□

59. In general, it need not hold that $(X, Y) \equiv (Y, X)$—for example, if X and Y are zero-mean Gaussian with different variances. However, the class of scalar Cauchy distributions does satisfy this property since the result of Theorem 7 is invariant to swapping $\lambda_1 \leftrightarrow \lambda_0$ and $\mu_1 \leftrightarrow \mu_0$. More generally, Theorem 7 implies that, if $\lambda_1\lambda_0\gamma_1\gamma_0 \neq 0$, then

$$\begin{array}{c} (\lambda_1 V + \mu_1, \lambda_0 V + \mu_0) \equiv (\gamma_1 V + \nu_1, \gamma_0 V + \nu_0) \\ \Updownarrow \\ \dfrac{\lambda_1^2 + \lambda_0^2 + (\mu_1 - \mu_0)^2}{|\lambda_0\lambda_1|} = \dfrac{\gamma_1^2 + \gamma_0^2 + (\nu_1 - \nu_0)^2}{|\gamma_0\gamma_1|}. \end{array} \tag{177}$$

Curiously, (177) implies that $(V, V+1) \equiv (V, 2V+1)$.

60. For location–dilation families of random variables, we saw in Item 56 how to reduce a four-parameter problem into a two-parameter problem since $(\lambda_1 V + \mu_1, \lambda_0 V + \mu_0) \equiv (V, \lambda V + \mu)$ with the appropriate substitution. In the Cauchy case, Theorem 7 reveals that, in fact, we can go one step further and turn it into a one-parameter problem. We have two basic ways of doing this:

(a) $(\lambda_1 V + \mu_1, \lambda_0 V + \mu_0) \equiv (V, V + \mu)$ with $\mu^2 = 2\zeta - 2$.

(b) $(\lambda_1 V + \mu_1, \lambda_0 V + \mu_0) \equiv (V, \lambda V)$ with either

$$\lambda = \zeta - \sqrt{\zeta^2 - 1} < 1, \quad \text{or} \quad \lambda = \zeta + \sqrt{\zeta^2 - 1} > 1, \tag{178}$$

which are the solutions to $\zeta = \frac{\lambda^2+1}{2\lambda}$.

9. f-Divergences

This section studies the interplay of f-divergences and equivalent pairs of measures.

61. If $P \ll Q$ and $f\colon [0,\infty) \to \mathbb{R}$ is convex and right-continuous at 0, f-divergence is defined as

$$D_f(P \,\|\, Q) = \mathbb{E}\left[f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right)\right], \quad Y \sim Q. \tag{179}$$

62. The most important property of f-divergence is the data processing inequality

$$D_f(P_X \,\|\, Q_X) \geq D_f(P_Y \,\|\, Q_Y), \tag{180}$$

where P_Y and Q_Y are the responses of a (random or deterministic) transformation to P_X and Q_X, respectively. If f is strictly convex at 1 and $D_f(P_X \,\|\, Q_X) < \infty$, then $(P_X, Q_X) \equiv (P_Y, Q_Y)$ is necessary and sufficient for equality in (180).

63. If $(P, Q) \equiv (Q, P)$, then $D_f(P \,\|\, Q) = D_{f^\star}(P \,\|\, Q)$ with the transform $f^\star(t) = t\, f(\frac{1}{t})$, which satisfies $f^{\star\star} = f$.

64.

Theorem 9. *If $P_1 \ll Q_1$ and $P_2 \ll Q_2$, then*

$$(P_1, Q_1) \equiv (P_2, Q_2) \iff D_f(P_1 \,\|\, Q_1) = D_f(P_2 \,\|\, Q_2), \ \forall f, \tag{181}$$

where $\forall f$ *stands for all convex right-continuous* $f\colon [0,\infty) \to \mathbb{R}$.

Proof. As mentioned in Item 53, $(P_1, Q_1) \equiv (P_2, Q_2)$ is equivalent to $\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(Y_1)$ and $\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(Y_2)$ having identical distributions with $Y_1 \sim Q_1$ and $Y_2 \sim Q_2$.

$\Longrightarrow$ According to (179), $D_f(P \,\|\, Q)$ is determined by the distribution of the random variable $\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)$, $Y \sim Q$.

$\Longleftarrow$ For $t \in \mathbb{R}$, the function $f_t(x) = \mathrm{e}^{tx}$, $x \geq 0$, is convex and right-continuous at 0, and $D_{f_t}(P \,\|\, Q)$ is the moment generating function, evaluated at t, of the random variable $\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)$, $Y \sim Q$. Therefore, $D_{f_t}(P_1 \,\|\, Q_1) = D_{f_t}(P_2 \,\|\, Q_2)$ for all t implies that $(P_1, Q_1) \equiv (P_2, Q_2)$.

□

65. Since $P \ll Q$ is not necessary in order to define (finite) $D_f(P\|Q)$, it is possible to enlarge the scope of Theorem 9 by defining $(P_1, Q_1) \equiv (P_2, Q_2)$ dropping the restriction that $P_1 \ll Q_1$ and $P_2 \ll Q_2$. For that purpose, let μ_1 and μ_2 be σ-finite measures on $(\mathcal{A}_1, \mathscr{F}_1)$ and $(\mathcal{A}_2, \mathscr{F}_2)$, respectively, and denote $p_i = \frac{\mathrm{d}P_i}{\mathrm{d}\mu_i}$, $q_i = \frac{\mathrm{d}Q_i}{\mathrm{d}\mu_i}$, $i = 1, 2$. Then, we say $(P_1, Q_1) \equiv (P_2, Q_2)$ if

(a) when restricted to $[0, 1]$, the random variables $\frac{p_1(Y_1)}{q_1(Y_1)}$ and $\frac{p_2(Y_2)}{q_2(Y_2)}$ have identical distributions with $Y_1 \sim Q_1$ and $Y_2 \sim Q_2$;

(b) when restricted to $[0, 1]$, the random variables $\frac{q_1(X_1)}{p_1(X_1)}$ and $\frac{q_2(X_2)}{p_2(X_2)}$ have identical distributions with $X_1 \sim P_1$ and $X_2 \sim P_2$.

Note that those conditions imply that

(c) $Q_1(\{\omega \in \mathcal{A}_1 \colon p_1(\omega) = q_1(\omega)\}) = Q_2(\{\omega \in \mathcal{A}_2 \colon p_2(\omega) = q_2(\omega)\})$;

(d) $Q_1(\{\omega \in \mathcal{A}_1 \colon p_1(\omega) = 0\}) = Q_2(\{\omega \in \mathcal{A}_2 \colon p_2(\omega) = 0\})$;

(e) $P_1(\{\omega \in \mathcal{A}_1 \colon q_1(\omega) = 0\}) = P_2(\{\omega \in \mathcal{A}_2 \colon q_2(\omega) = 0\})$.

For example, if $P_1 \perp Q_1$ and $P_2 \perp Q_2$, then $(P_1, Q_1) \equiv (P_2, Q_2)$. To show the generalized version of Theorem 9, it is convenient to use the symmetrized form

$$D_f(P \,\|\, Q) = \int_{0 \leq p < q} q\, f\left(\frac{p}{q}\right) \mathrm{d}\mu + \int_{0 \leq q < p} p\, f^\star\left(\frac{q}{p}\right) \mathrm{d}\mu + f(1)\, Q[p = q]. \tag{182}$$

66. Suppose that there is a class $\mathcal{C}$ of probability measures on a given measurable space with the property that there exists a convex function $g\colon (0,\infty) \to \mathbb{R}$ (right-continuous at 0) such that, if $(P_1, Q_1) \in \mathcal{C}^2$ and $(P_2, Q_2) \in \mathcal{C}^2$, then

$$D_g(P_1 \,\|\, Q_1) = D_g(P_2 \,\|\, Q_2) \iff (P_1, Q_1) \equiv (P_2, Q_2). \tag{183}$$

In such case, Theorem 9 indicates that $\mathcal{C}^2$ can be partitioned into equivalence classes such that, within every equivalence class, the value of $D_f(P \,\|\, Q)$ is constant, though naturally dependent on f. Throughout $\mathcal{C}^2$, the value of $D_g(P \,\|\, Q)$ determines the value of $D_f(P \,\|\, Q)$, i.e., we can express $D_f(P \,\|\, Q) = \vartheta_{f,g}\big(D_g(P \,\|\, Q)\big)$, where $\vartheta_{f,g}$ is a non-decreasing function. Consider the following examples:

(a) Let $\mathcal{C}$ be the class of real-valued Gaussian probability measures with given variance $\sigma^2 > 0$. Then,

$$D\Big(\mathcal{N}\big(\mu_1, \sigma^2\big) \,\|\, \mathcal{N}\big(\mu_2, \sigma^2\big)\Big) = \frac{(\mu_1 - \mu_2)^2}{\sigma^2} \log \mathrm{e}. \tag{184}$$

Since Theorem 8 implies that $(\mathcal{N}(\mu_1, \sigma^2), \mathcal{N}(\mu_2, \sigma^2)) \equiv (\mathcal{N}(\mu_3, \sigma^2), \mathcal{N}(\mu_4, \sigma^2))$ as long as $(\mu_1 - \mu_2)^2 = (\mu_3 - \mu_4)^2$, (184) indicates that (183) is satisfied with $g(t)$ given by the right-continuous extension of $t \log t$. Therefore, we can con-

clude that, regardless of f, $D_f\big(\mathcal{N}(\mu_1,\sigma^2)\,\|\,\mathcal{N}(\mu_2,\sigma^2)\big)$ depends on (μ_1,μ_2,σ^2) only through $(\mu_1-\mu_2)^2/\sigma^2$.

(b) Let $\mathcal{C}$ be the collection of all Cauchy random variables. Theorem 7 reveals that (183) is also satisfied if $g(x)=x^2$ because, if $X\sim P$ and $Y\sim Q$, then

$$\mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}(X)\right]=\mathbb{E}\left[\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(Y)\right)^2\right]. \tag{185}$$

67. An immediate consequence of Theorems 7 and 9 is that, for any valid f, the f-divergence between Cauchy densities is symmetric,

$$D_f(\lambda_1 V+\mu_1\,\|\,\lambda_0 V+\mu_0)=D_f(\lambda_0 V+\mu_0\,\|\,\lambda_1 V+\mu_1). \tag{186}$$

This property does not generalize to the multivariate case. While, in view of Theorem 8,

$$(\boldsymbol{\Lambda}^{\frac{1}{2}}V^n+\boldsymbol{\mu}_1,\boldsymbol{\Lambda}^{\frac{1}{2}}V^n+\boldsymbol{\mu}_2)\equiv(\boldsymbol{\Lambda}^{\frac{1}{2}}V^n+\boldsymbol{\mu}_2,\boldsymbol{\Lambda}^{\frac{1}{2}}V^n+\boldsymbol{\mu}_1), \tag{187}$$

in general, $(\boldsymbol{\Lambda}^{\frac{1}{2}}V^n,V^n)\not\equiv(V^n,\boldsymbol{\Lambda}^{\frac{1}{2}}V^n)$ since the corresponding relative entropies do not coincide as shown in [8].

68. It follows from Item 66 and Theorem 7 that any f-divergence between Cauchy probability measures $D_f(\lambda_1 V+\mu_1\,\|\,\lambda_0 V+\mu_0)$ is a monotonically increasing function of $\zeta(\lambda_1,\mu_1,\lambda_0,\mu_0)$ given by (149). The following result shows how to obtain that function from f.

Theorem 10. *With f_Z given in* (153),

$$D_f(\lambda_1 V+\mu_1\,\|\,\lambda_0 V+\mu_0)=\int_{\zeta-\sqrt{\zeta^2-1}}^{\zeta+\sqrt{\zeta^2-1}} f\left(\frac{1}{z}\right)f_Z(z)\,\mathrm{d}z \tag{188}$$

$$=\mathbb{E}\left[f\left(\left(\zeta+\sqrt{\zeta^2-1}\,\cos\Theta\right)^{-1}\right)\right] \tag{189}$$

$$=\int_{\zeta-\sqrt{\zeta^2-1}}^{\zeta+\sqrt{\zeta^2-1}}\frac{1}{z}\,f(z)\,f_Z(z)\,\mathrm{d}z. \tag{190}$$

where Θ is uniformly distributed on $[0,\pi]$ in (189).

Proof. In view of (179) and the definition of Z in Theorem 7,

$$D_f(\lambda_1 V+\mu_1\,\|\,\lambda_0 V+\mu_0)=\mathbb{E}\left[f\left(\frac{1}{Z}\right)\right], \tag{191}$$

thereby justifying (188) and (189) since we saw in Theorem 7 that Z has the distribution of $\zeta+\sqrt{\zeta^2-1}\,\cos\Theta$ with Θ uniformly distributed on $[0,\pi]$. Item 52 results in (190). Alternatively, we can rely on Item 63 and substitute f by $f^\star$ on the right side of (188). □

69. Suppose now that we have two sequences of Cauchy measures with respective parameters $(\lambda_1^{(n)},\mu_1^{(n)})$ and $(\lambda_0^{(n)},\mu_0^{(n)})$ such that $\zeta(\lambda_1^{(n)},\mu_1^{(n)},\lambda_0^{(n)},\mu_0^{(n)})\to 1$. Then, Theorem 10 indicates that

$$D_f\left(\lambda_1^{(n)}V+\mu_1^{(n)}\,\|\,\lambda_0^{(n)}V+\mu_0^{(n)}\right)\to f(1). \tag{192}$$

The most common f-divergences are such that $f(1)=0$ since in that case $D_f(P\|Q)\geq 0$. In addition, adding the function $\alpha t-\alpha$ to $f(t)$ does not change the value of $D_f(P\|Q)$ and with appropriately chosen α, we can turn $f(t)$ into canonical form in which not

only $f(1) = 0$ but $f(t) \geq 0$. In the special case in which the second measure is fixed, Theorem 9 in [25] shows that, if ess sup $\frac{\mathrm{d}P_n}{\mathrm{d}Q}(Y) \to 1$ with $Y \sim Q$, then

$$\lim_{n\to\infty} \frac{D_f(P_n\|Q)}{D_g(P_n\|Q)} = \lim_{t\to 1} \frac{f(t)}{g(t)}, \tag{193}$$

provided the limit on the right side exists; otherwise, the left side lies between the left and right limits at 1. In the Cauchy case, we can allow the second probability to depend on n and sharpen that result by means of Theorem 10. In particular, it can be shown that

$$\lim_{n\to\infty} \frac{D_f\left(\lambda_1^{(n)} V + \mu_1^{(n)} \,\|\, \lambda_0^{(n)} V + \mu_0^{(n)}\right)}{D_g\left(\lambda_1^{(n)} V + \mu_1^{(n)} \,\|\, \lambda_0^{(n)} V + \mu_0^{(n)}\right)} = \frac{\dot{f}(0^-) + \dot{f}(0^+)}{\dot{g}(0^-) + \dot{g}(0^+)} \tag{194}$$

provided the right side is not $\frac{0}{0}$.

10. χ^2-Divergence

70. With either $f(x) = (x-1)^2$ or $f(x) = x^2 - 1$, f-divergence is the χ^2-*divergence*,

$$\chi^2(P\,\|\,Q) = \mathbb{E}\left[\frac{\mathrm{d}P}{\mathrm{d}Q}(X)\right] - 1, \;\; X \sim P. \tag{195}$$

71. If P and Q are Cauchy distributions, then (149), (151) and (195) result in

$$\chi^2(\lambda_1 V + \mu_1 \,\|\, \lambda_0 V + \mu_0) = \zeta(\lambda_1, \mu_1, \lambda_0, \mu_0) - 1 \tag{196}$$

$$= \frac{(|\lambda_0| - |\lambda_1|)^2 + (\mu_1 - \mu_0)^2}{2\,|\lambda_0\,\lambda_1|}, \tag{197}$$

a formula obtained in Appendix D of [26] using complex analysis and the Cauchy integral formula. In addition, invoking complex analysis and the maximal group invariant results in [27,28], ref. [26] shows that any f-divergence between Cauchy distributions can be expressed as a function of their χ^2 divergence, although [26] left open how to obtain that function, which is given by Theorem 10 substituting $\zeta = 1 + \chi^2$.

11. Relative Entropy

72. The relative entropy between Cauchy distributions is given by

$$D(\lambda_1 V + \mu_1 \,\|\, \lambda_0 V + \mu_0) = \log\left(\frac{(|\lambda_0| + |\lambda_1|)^2 + (\mu_1 - \mu_0)^2}{4\,|\lambda_0\,\lambda_1|}\right), \tag{198}$$

where $\lambda_1\lambda_0 \neq 0$. The special case $\lambda_1 = \lambda_0$ of (198) was found in Example 4 of [29]. The next four items give different simple justifications for (198). An alternative proof was recently given in Appendix C of [26] using complex analysis holomorphisms and the Cauchy integral formula. Yet another, much more involved, proof is reported in [30]. See also Remark 19 in [26] for another route invoking the Lévy–Khintchine formula and the Frullani integral.

73. Since for absolutely continuous random variables $D(X\,\|\,Y) = -h(X) - \mathbb{E}[\log f_Y(X)]$,

$$D(V\,\|\,\lambda\,V + \mu) = -h(V) + \log\frac{\pi}{|\lambda|} + \mathbb{E}\left[\log\left(\lambda^2 + (V-\mu)^2\right)\right] \tag{199}$$

$$= -\log(4|\lambda|) + \log\left((1+|\lambda|)^2 + \mu^2\right), \tag{200}$$

where (200) follows from (12) and (A4) with $\alpha^2 = \lambda^2 + \mu^2$ and $\cos\beta = \frac{\mu}{|\alpha|}$.

Now, substituting $\lambda = \frac{\lambda_0}{\lambda_1}$ and $\mu = \frac{\mu_0 - \mu_1}{\lambda_1}$, we obtain (198) since, according to Item 56, $(V, \lambda V + \mu) \equiv (\lambda_1 V + \mu_1, \lambda_0 V + \mu_0)$.

74. From the formula found in Example 4 of [29] and the fact that, according to (197), $\chi^2 = \frac{\mu^2}{2\lambda^2}$ when $\lambda_1 = \lambda_0 = \lambda$, we obtain

$$D(\lambda V + \mu \,\|\, \lambda V) = \log\left(1 + \frac{\mu^2}{4\lambda^2}\right) = \log\left(1 + \tfrac{1}{2}\chi^2\right). \tag{201}$$

Moreover, as argued in Item 60, (201) is also valid for the relative entropy between Cauchy distributions with $\lambda_1 \neq \lambda_0$ as long as χ^2 is given in (197). Indeed, we can verify that the right side of (201) becomes (198) with said substitution.

75. By the definition of relative entropy, and Theorem 7,

$$D(\lambda_1 V + \mu_1 \,\|\, \lambda_0 V + \mu_0) = \mathbb{E}[\log Z] \tag{202}$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \log\left(\zeta + \sqrt{\zeta^2 - 1}\, \cos\theta\right) \mathrm{d}\theta \tag{203}$$

$$= \log\left(\frac{1+\zeta}{2}\right), \tag{204}$$

where (204) follows from (A14). Then, (198) results by plugging into (204) the value of ζ in (149).

76. Evaluating (190) with $f(t) = t \log t$ results in (202).

77. If V is standard Cauchy, independent of Cauchy V_1 and V_0, then (198) results in

$$D(\lambda V + \epsilon V_1 \,\|\, \lambda V + \epsilon V_0) = \frac{\epsilon^2}{4\lambda^2}\Big((\lambda_1 - \lambda_0)^2 + (\mu_1 - \mu_0)^2\Big) \log \mathrm{e} + o(\epsilon^2), \tag{205}$$

where $V_1 = \lambda_1 V' + \lambda_1$ and $V_0 = \lambda_1 V' + \lambda_1$, and V' is an independent (or exact) copy of V. In contrast, the corresponding result in the Gaussian case in which X, X_1, X_0 are independent Gaussian with means μ, μ_1, μ_0 and variances $\sigma^2, \sigma_1^2, \sigma_0^2$, respectively, is

$$D(X + \epsilon X_1 \,\|\, X + \epsilon X_0) = \frac{\epsilon^2}{2\sigma^2}(\mu_1 - \mu_0)^2 \log \mathrm{e} + o(\epsilon^2). \tag{206}$$

In fact, it is shown in Lemma 1 of [31] that (206) holds even if X_1 and X_0 are not Gaussian but have finite variances. It is likely that (205) holds even if V_1 and V_0 are not Cauchy, but have finite strengths.

78. An important information theoretic result due to Csiszár [32] is that if $Q_1 \ll Q_2$ and P is such that

$$\mathbb{E}\Big[\imath_{Q_1 \| Q_2}(X)\Big] = D(Q_1 \,\|\, Q_2), \quad X \sim P, \tag{207}$$

then the following *Pythagorean identity* holds

$$D(P \,\|\, Q_2) = D(P \,\|\, Q_1) + D(Q_1 \,\|\, Q_2). \tag{208}$$

Among other applications, this result leads to elegant proofs of minimum relative entropy results. For example, the closest Gaussian to a given P with a finite second moment has the same first and second moments as P. If we let Q_1 and Q_2 be centered Cauchy with strengths λ_1 and λ_2, respectively, then the orthogonality condition (207) becomes, with the aid of (148) and (198),

$$\mathcal{V}_{X^2}\left(\lambda_2^{-1}\right) - \mathcal{V}_{X^2}\left(\lambda_1^{-1}\right) = 2 \log_{\mathrm{e}}\left(1 + \frac{\lambda_1}{\lambda_2}\right) - 2 \log_{\mathrm{e}} 2. \tag{209}$$

If, in addition, P is centered Cauchy, we can use (28) to verify that (209) holds only in the trivial cases in which either $\lambda_1 = \lambda_2$ or $P = Q_1$. For non-Cauchy P, (208) may indeed be satisfied with $\lambda_1 \neq \lambda_2$. For example, using (30), if $X = V_{2,2}$, then (209), and therefore (208), holds with $(\lambda_1, \lambda_2) = (2, 0.35459\ldots)$.

79. Mutually absolutely continuous random variables may be such that

$$D(X \,\|\, Z) < \infty = D(Z \,\|\, X). \tag{210}$$

An easy example is that of Gaussian X and Cauchy Z, or, if we let X be Cauchy, (210) holds with Z having the very heavy-tailed density function in (62).

80. While relative entropy is lower semi-continuous, it is not continuous. For example, using the Cauchy distribution, we can show that relative entropy is not stable against small contamination of a Gaussian random variable: if X is Gaussian independent of V, then no matter how small $\lambda \neq 0$,

$$D(\lambda|V| + X \,\|\, -\lambda|V| + X) = \infty. \tag{211}$$

12. Total Variation Distance

81. With $f(x) = |x-1|$, f-divergence becomes the *total variation distance* (with range [0,2]). Moreover, we have the following representation:

Theorem 11. *If $P \ll\gg Q$ and $(P, Q) \equiv (Q, P)$, then*

$$\tfrac{1}{2}|P - Q| = 2\,\mathbb{P}[Z > 1] - \mathbb{P}[Z \neq 1], \tag{212}$$

with $Z = \frac{\mathrm{d}P}{\mathrm{d}Q}(X)$, $X \sim P$.

Proof.

$$\begin{aligned}
\tfrac{1}{2}|P - Q| &= \max_{A \in \mathcal{F}} \{P(A) - Q(A)\} && (213)\\
&= P\left(\omega\colon \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) > 1\right) - Q\left(\omega\colon \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) > 1\right) && (214)\\
&= P\left(\omega\colon \frac{\mathrm{d}P}{\mathrm{d}Q}(\omega) > 1\right) - P\left(\omega\colon \frac{\mathrm{d}Q}{\mathrm{d}P}(\omega) > 1\right) && (215)\\
&= \mathbb{P}[Z > 1] - \mathbb{P}[Z < 1] && (216)
\end{aligned}$$

where (215) and (216) follow from $(P, Q) \equiv (Q, P)$ and $P \ll\gg Q$, respectively. □

82. Example 15 of [33] shows that the total variation distance between centered Cauchy distributions is

$$\begin{aligned}
\left|P_{\lambda_1 V} - P_{\lambda_0 V}\right| &= \frac{4}{\pi} \arctan\left(\frac{||\lambda_1| - |\lambda_0||}{2\sqrt{|\lambda_0 \lambda_1|}}\right) && (217)\\
&= \frac{4}{\pi} \arctan\left(\sqrt{\tfrac{1}{2}\,\chi^2(P_{\lambda_1 V} \| P_{\lambda_0 V})}\right) && (218)
\end{aligned}$$

in view of (197). Since any f-divergence between Cauchy distributions depends on the parameters only through the corresponding χ^2-divergence, (217)–(218) imply the general formula

$$\left|P_{\lambda_1 V + \mu_1} - P_{\lambda_0 V + \mu_0}\right| = \frac{4}{\pi} \arctan\left(\sqrt{\tfrac{1}{2}\,\chi^2(P_{\lambda_1 V + \mu_1} \| P_{\lambda_0 V + \mu_0})}\right). \tag{219}$$

Alternatively, applying Theorem 11 to the case of Cauchy random variables, note that, in this case, Z is an absolutely continuous random variable with density function (153). Therefore, $\mathbb{P}[Z \neq 1] = 1$, and

$$\mathbb{P}[Z > 1] = \frac{1}{\pi} \int_1^{\zeta + \sqrt{\zeta^2 - 1}} \frac{1}{\sqrt{2z\zeta - z^2 - 1}} \, \mathrm{d}z \tag{220}$$

$$= \frac{1}{2} + \frac{1}{\pi} \arctan \sqrt{\tfrac{1}{2}\chi^2}, \tag{221}$$

where (221) follows from (154) and the identity $\arcsin \frac{\sqrt{\delta}}{\sqrt{1+\delta}} = \arctan \sqrt{\delta}$ specialized to $\delta = \frac{1}{2}\chi^2 = \frac{1}{2}(\zeta - 1)$. Though more laborious (see [26]), (219) can also be verified by direct integration.

13. Hellinger Divergence

83. The Hellinger divergence, $\mathscr{H}_\alpha(P \,\|\, Q)$ of order $\alpha \in (0,1) \cup (1,\infty)$, is the f_α-divergence with

$$f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}. \tag{222}$$

Notable special cases are

$$\mathscr{H}_2(P \,\|\, Q) = \chi^2(P \,\|\, Q), \tag{223}$$

$$\lim_{\alpha \downarrow 1} \mathscr{H}_\alpha(P \,\|\, Q) = D(P \,\|\, Q), \tag{224}$$

$$\mathscr{H}_{\frac{1}{2}}(P \,\|\, Q) = 2\,\mathscr{H}^2(P \,\|\, Q), \tag{225}$$

where $\mathscr{H}^2(P \,\|\, Q)$ is known as the squared Hellinger distance.

84. For Cauchy random variables, Theorem 10 yields

$$\mathscr{H}_\alpha(\lambda_1 V + \mu_1 \,\|\, \lambda_0 V + \mu_0) = \frac{1}{\alpha - 1}\left(\mathbb{E}[Z^{-\alpha}] - 1\right) \tag{226}$$

$$= \frac{\mathrm{P}_{\alpha-1}(\zeta) - 1}{\alpha - 1}, \tag{227}$$

where ζ is as given in (149), and we have used (A15) and $\mathrm{P}_\alpha(\cdot)$ denotes the Legendre function of the first kind, which satisfies $\mathrm{P}_{-\alpha} = \mathrm{P}_{\alpha-1}$ (see 8.2.1. in [34]).

14. Rényi Divergence

85. For absolutely continuous probability measures P and Q, with corresponding probability density functions p and q, the *Rényi divergence of order* $\alpha \in [0,1) \cup (1,\infty)$ is [35]

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log\left(\int_{-\infty}^{\infty} p^\alpha(t)\, q^{1-\alpha}(t) \, \mathrm{d}t \right). \tag{228}$$

Note that, if $(P_1, Q_1) \equiv (P_2, Q_2)$, then $D_\alpha(P_1 \,\|\, Q_1) = D_\alpha(P_2 \,\|\, Q_2)$. Moreover, although Rényi divergence of order α is not an f-divergence, it is in one-to-one correspondence with the Hellinger divergence of order α:

$$D_\alpha(P \,\|\, Q) = \frac{1}{\alpha - 1} \log(1 + (\alpha - 1)\, \mathscr{H}_\alpha(P \,\|\, Q)). \tag{229}$$

86. An extensive table of order-α Rényi divergences for various continuous random variables can be found in [36]. An addition to that list for Cauchy random variables can be obtained plugging (227) into (229):

$$D_\alpha(\lambda_1 V+\mu_1 \,\|\, \lambda_0 V+\mu_0) = \frac{\log \mathrm{P}_{\alpha-1}(\zeta)}{\alpha-1} \tag{230}$$

$$= \frac{1}{\alpha-1}\log \mathrm{P}_{\alpha-1}\left(\frac{\lambda_1^2+\lambda_0^2+(\mu_1-\mu_0)^2}{2|\lambda_0\lambda_1|}\right), \tag{231}$$

for $\alpha \in (0,1)\cup(1,\infty)$.

87. Suppose that $\lambda \in (0,1)$. Then, (A16) yields

$$D_{\frac{1}{2}}(V \,\|\, \lambda V) = -2\log\left(\frac{2\sqrt{\lambda}}{\pi}\,\mathbf{K}\left(\sqrt{1-\lambda^2}\right)\right), \tag{232}$$

where $\mathbf{K}(\cdot)$ stands for the complete elliptical integral of the first kind in (A18). As indicated in Item 60, to obtain $D_{\frac{1}{2}}(\lambda_1 V+\mu_1 \,\|\, \lambda_0 V+\mu_0)$, we just need to substitute λ by $\zeta-\sqrt{\zeta^2-1}$ in (232), with ζ given by (149).

88. Notice that, specializing (86) to $(\alpha,\mu_0,\mu_1,\lambda_0,\lambda_1) = (\frac{1}{2},0,0,\lambda,1)$, (232) results in the identity

$$\mathrm{P}_{-\frac{1}{2}}\left(\frac{1}{2\lambda}+\frac{\lambda}{2}\right) = \frac{2\sqrt{\lambda}}{\pi}\,\mathbf{K}\left(\sqrt{1-\lambda^2}\right), \quad \lambda\in(0,1). \tag{233}$$

Writing the complete elliptical integral of the first kind and the Legendre function of the first kind as special cases of the Gauss hypergeometric function, González [37] noticed the simpler identity (see also 8.13.8 in [34])

$$\mathrm{P}_{-\frac{1}{2}}(\lambda) = \frac{2}{\pi}\,\mathbf{K}\left(\sqrt{\frac{1-\lambda}{2}}\right), \quad \lambda\in(0,1). \tag{234}$$

We can view (233) and (234) as complementary of each other since they constrain the argument of the Legendre function to belong to $(1,\infty)$ and $(0,1)$, respectively.

89. Since $\mathrm{P}_1(z)=z$, particularizing (230), we obtain

$$D_2(\lambda_1 V+\mu_1 \,\|\, \lambda_0 V+\mu_0) = \log\zeta = \log\left(\frac{\lambda_1^2+\lambda_0^2+(\mu_1-\mu_0)^2}{2|\lambda_0\lambda_1|}\right). \tag{235}$$

90. Since $\mathrm{P}_2(z)=\frac{1}{2}(3z^2-1)$, for Cauchy random variables, we obtain

$$D_3(P\,\|\,Q) = \frac{1}{2}\log\left(1+3\,\chi^2(P\,\|\,Q)+\frac{3}{2}\,\chi^4(P\,\|\,Q)\right). \tag{236}$$

91. For Cauchy random variables, the Rényi divergence for integer order 4 or higher can be obtained through (235), (236) and the recursion (dropping $(P\,\|\,Q)$ for typographical convenience)

$$(n+1)\exp((n+1)D_{n+2}) = (2n+1)\,\zeta\,\exp(nD_{n+1}) - n\exp((n-1)D_n), \tag{237}$$

which follows from (230) and the recursion of the Legendre polynomials

$$(n+1)\,\mathrm{P}_{n+1}(z) = (2n+1)\,z\,\mathrm{P}_n(z) - n\,\mathrm{P}_{n-1}(z), \tag{238}$$

which, in fact, also holds for non-integer n (see 8.5.3 in [34]).

92. The Chernoff information

$$C(P \parallel Q) = \sup_{\lambda \in (0,1)} (1-\lambda) D_\lambda(P \parallel Q) \tag{239}$$

satisfies $C(P \parallel Q) = C(Q \parallel P)$ regardless of (P, Q). If, as in the case of Cauchy measures, $(P, Q) \equiv (Q, P)$, then Chernoff information is equal to the Bhattacharyya distance:

$$C(P \parallel Q) = \tfrac{1}{2} D_{\frac{1}{2}}(P \parallel Q) = \log \frac{1}{\int_{-\infty}^{\infty} \sqrt{p(t)\, q(t)}\, \mathrm{d}t} = -\log\Big(1 - \mathscr{H}^2(P \| Q)\Big), \tag{240}$$

where $\mathscr{H}^2(P\|Q)$ is the squared Hellinger distance, which is the f-divergence with $f(t) = \frac{1}{2}(1-\sqrt{t})^2$. Together with Item 87, (240) gives the Chernoff information for Cauchy distributions. While it involves the complete elliptical integral function, its simplicity should be contrasted with the formidable expression for Gaussian distributions, recently derived in [38]. The reason (240) holds is that the supremum in (239) is achieved at $\lambda^* = \frac{1}{2}$. To see this, note that

$$f(\lambda) = (1-\lambda) D_\lambda(P \parallel Q) = \lambda\, D_{1-\lambda}(Q \parallel P) \tag{241}$$
$$= \lambda\, D_{1-\lambda}(P \parallel Q) \tag{242}$$
$$= f(1-\lambda), \tag{243}$$

where (241) reflects the skew-symmetry of Rényi divergence, and (242) holds because $(P, Q) \equiv (Q, P)$. Since $f(\lambda)\colon \lambda \in [0,1]$ is concave and its own mirror image, it is maximized at $\lambda^* = \frac{1}{2}$.

15. Fisher's Information

93. The score function of the standard Cauchy density (1) is

$$\rho_V(x) = \nabla \log_{\mathrm{e}} f_V(x) = -\nabla \log_{\mathrm{e}}(1+x^2) = -\frac{2\,x}{1+x^2}. \tag{244}$$

Then, $\rho_V(V)$ is a zero-mean random variable with second moment equal to Fisher's information

$$J(V) = \mathbb{E}\Big[\rho_V^2(V)\Big] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{4\,t^2}{(1+t^2)^3}\, \mathrm{d}t = \frac{1}{2}, \tag{245}$$

where we have used (A11). Since Fisher's information is invariant to location and scales as $J(X) = \alpha^2\, J(\alpha X)$, we obtain

$$J(\lambda V + \mu) = \frac{1}{2\,\lambda^2}. \tag{246}$$

Together with (117), the product of entropy power and Fisher information is $\frac{4\pi}{\mathrm{e}}$, thereby abiding by Stam's inequality [4], $1 \le N(X)\, J(X)$.

94. Introduced in [39], Fisher's information of a density function (245) quantifies its similarity with a slightly shifted version of itself. A more general notion is the Fisher information matrix of a random transformation $P_{Y|X}\colon \mathbb{R}^k \to \mathcal{Y}$ satisfying the regularity condition

$$D(P_{Y|X=\boldsymbol{\alpha}} \parallel P_{Y|X=\boldsymbol{\theta}}) = o(\|\boldsymbol{\alpha} - \boldsymbol{\theta}\|). \tag{247}$$

Then, the Fisher information matrix of $P_{Y|X}$ at $\boldsymbol{\theta}$ has coefficients

$$J_{ij}(\boldsymbol{\theta}, P_{Y|X}) = \mathbb{E}\left[\frac{\partial}{\partial \alpha_i} \imath_{P_{Y|X=\boldsymbol{\alpha}} \| P_{Y|X=\boldsymbol{\theta}}}(Y_{\boldsymbol{\theta}}) \frac{\partial}{\partial \alpha_j} \imath_{P_{Y|X=\boldsymbol{\alpha}} \| P_{Y|X=\boldsymbol{\theta}}}(Y_{\boldsymbol{\theta}})\right]|_{\boldsymbol{\alpha} \leftarrow \boldsymbol{\theta}}, \tag{248}$$

and satisfies (with relative entropy in nats)

$$D(P_{Y|X=\boldsymbol{\alpha}} \,\|\, P_{Y|X=\boldsymbol{\theta}}) = \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\theta})^\top \mathbf{J}(\boldsymbol{\theta}, P_{Y|X})(\boldsymbol{\alpha} - \boldsymbol{\theta}) + o(\|\boldsymbol{\alpha} - \boldsymbol{\theta}\|^2). \tag{249}$$

For the Cauchy family, the parametrization vector has two components, location and strength, namely, $\boldsymbol{\theta}^\top = (\mu, \lambda)$. The regularity condition (247) is satisfied in view of (205), and we can use the closed-form expression in (205) to obtain

$$J_{11}(\boldsymbol{\theta}, P_{Y|X}) = J_{22}(\boldsymbol{\theta}, P_{Y|X}) = \frac{1}{2\lambda^2}, \tag{250}$$

$$J_{12}(\boldsymbol{\theta}, P_{Y|X}) = J_{21}(\boldsymbol{\theta}, P_{Y|X}) = 0. \tag{251}$$

95. The relative Fisher information is defined as

$$J(P \,\|\, Q) = \mathbb{E}\left[\left(\nabla \imath_{P\|Q}(X)\right)^2\right], \quad X \sim P. \tag{252}$$

Although the purpose of this definition is to avoid some of the pitfalls of the classical definition of Fisher's information, not only do equivalent pairs fail to have the same relative Fisher information but, unlike relative entropy or f-divergence, relative Fisher information is not transparent to injective transformations. For example, $J(X\|Y) = \lambda^2 J(\lambda X\| \lambda Y)$. Centered Cauchy random variables illustrate this fact since

$$J(V \,\|\, \lambda V) = \frac{(4+\lambda)(\lambda-1)^2}{2\lambda(1+\lambda)^2} \quad \text{and} \quad J(\lambda V \,\|\, V) = \frac{(4\lambda+1)(\lambda-1)^2}{2\lambda^2(1+\lambda)^2}. \tag{253}$$

96. de Bruijn's identity [4] states that, if $N \sim \mathcal{N}(0,1)$ is independent of X, then, in nats,

$$\frac{\mathrm{d}}{\mathrm{d}t} h\left(X + \sqrt{t}\, N\right) = \frac{1}{2} J\left(X + \sqrt{t}\, N\right), \quad t > 0. \tag{254}$$

As well as serving as the key component in the original proofs of the entropy power inequality, the differential equation in (254) provides a concrete link between Shannon theory and its prehistory. As we show in Theorem 12, it turns out that there is a Cauchy counterpart of de Bruijn's identity (254). Before stating the result, we introduce the following notation for a parametrized random variable Y_t (to be specified later):

$$\nabla \log_{\mathrm{e}} f_{Y_t}(y) = \frac{\partial}{\partial y} \log_{\mathrm{e}} f_{Y_t}(y) = f_{Y_t}^{-1}(y) \frac{\partial}{\partial y} f_{Y_t}(y), \tag{255}$$

$$\nabla_2 \log_{\mathrm{e}} f_{Y_t}(y) = \frac{\partial}{\partial t} \log_{\mathrm{e}} f_{Y_t}(y) = f_{Y_t}^{-1}(y) \frac{\partial}{\partial t} f_{Y_t}(y), \tag{256}$$

$$J(Y_t) = \mathbb{E}\left[\left(\nabla \log_{\mathrm{e}} f_{Y_t}(Y_t)\right)^2\right], \tag{257}$$

$$K(Y_t) = \mathbb{E}\left[\left(\nabla_2 \log_{\mathrm{e}} f_{Y_t}(Y_t)\right)^2\right], \tag{258}$$

i.e., $J(Y_t)$ and $K(Y_t)$ are the Fisher information with respect to location and with respect to dilation, respectively (corresponding to the coefficients J_{11} and J_{22} of the

Fisher information matrix when $\boldsymbol{\theta}^\top = (\mu, \lambda)$ as in Item 94. The key to (254) is that $Y_t = X + \sqrt{t}\,N$, $N \sim \mathcal{N}(0,1)$ satisfies the partial differential equation

$$\frac{\partial^2}{\partial y^2} f_{Y_t}(y) = \frac{\partial}{\partial t} f_{Y_t}(y). \tag{259}$$

Theorem 12. *Suppose that X is independent of standard Cauchy V. Then, in nats,*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} h(X + t\,V) = -J(X + t\,V) - K(X + t\,V), \quad t > 0. \tag{260}$$

Proof. Equation (259) does not hold in the current case in which $Y_t = X + t\,V$, and

$$f_{Y_t}(y) = \frac{t}{\pi}\,\mathbb{E}\left[\frac{1}{t^2 + (X-y)^2}\right]. \tag{261}$$

However, some algebra (the differentiation/integration swaps can be justified invoking the bounded convergence theorem) indicates that the convolution with the Cauchy density satisfies the Laplace partial differential equation

$$\frac{\partial^2}{\partial y^2} f_{Y_t}(y) = -\frac{\partial^2}{\partial t^2} f_{Y_t}(y) = \frac{2t}{\pi}\,\mathbb{E}\left[\frac{3(X-y)^2 - t^2}{(t^2 + (X-y)^2)^3}\right]. \tag{262}$$

The derivative of the differential entropy of Y_t is, in nats,

$$\frac{\mathrm{d}}{\mathrm{d}t} h(Y_t) = -\int_{-\infty}^{\infty} \frac{\partial}{\partial t} f_{Y_t}(y)\,\mathrm{d}y - \int_{-\infty}^{\infty} \log_{\mathrm{e}} f_{Y_t}(y)\,\frac{\partial}{\partial t} f_{Y_t}(y)\,\mathrm{d}y \tag{263}$$

$$= -\frac{\partial}{\partial t}\int_{-\infty}^{\infty} f_{Y_t}(y)\,\mathrm{d}y - \int_{-\infty}^{\infty} \log_{\mathrm{e}} f_{Y_t}(y)\,\frac{\partial}{\partial t} f_{Y_t}(y)\,\mathrm{d}y. \tag{264}$$

Taking another derivative, the left side of (260) becomes

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} h(Y_t) = -\int_{-\infty}^{\infty} \frac{\partial^2}{\partial t^2} f_{Y_t}(y)\,\log_{\mathrm{e}} f_{Y_t}(y)\,\mathrm{d}y - \int_{-\infty}^{\infty} \frac{\partial}{\partial t} f_{Y_t}(y)\,\frac{\partial}{\partial t}\log_{\mathrm{e}} f_{Y_t}(y)\,\mathrm{d}y \tag{265}$$

$$= \int_{-\infty}^{\infty} \frac{\partial^2}{\partial y^2} f_{Y_t}(y)\,\log_{\mathrm{e}} f_{Y_t}(y)\,\mathrm{d}y - \int_{-\infty}^{\infty} f_{Y_t}^{-1}(y)\left(\frac{\partial}{\partial t} f_{Y_t}(y)\right)^2 \mathrm{d}y \tag{266}$$

$$= \int_{-\infty}^{\infty} \frac{\partial^2}{\partial y^2} f_{Y_t}(y)\,\log_{\mathrm{e}} f_{Y_t}(y)\,\mathrm{d}y - K(Y_t) \tag{267}$$

$$= -J(Y_t) - K(Y_t), \tag{268}$$

where

- (265) $\Longleftarrow$ the first term on the right side of (264) is zero;
- (266) $\Longleftarrow$ (262);
- (267) $\Longleftarrow$ (258);
- (268) $\Longleftarrow$ integration by parts, exactly as in [4] (or p. 673 of [19]).

□

97. Theorem 12 reveals that the increasing function $f_X(t) = h(X + t\,V)$ is concave (which does not follow from the concavity of differential entropy functional of the density). In contrast, it was shown by Costa [40] that the entropy power $N\left(X + \sqrt{t}\,N\right)$, with $N \sim \mathcal{N}(0,1)$ is concave in t.

16. Mutual Information

98. Most of this section is devoted to an additive noise model. We begin with the simplest case in which X_C is centered Cauchy independent of W_C, also centered Cauchy with $\varsigma(W_C) > 0$. Then, (11) yields

$$I(X_C; X_C + W_C) = h(X_C + W_C) - h(W_C) \tag{269}$$
$$= \log(4\pi(\varsigma(X_C) + \varsigma(W_C))) - \log(4\pi\varsigma(W_C)) \tag{270}$$
$$= \log\left(1 + \frac{\varsigma(X_C)}{\varsigma(W_C)}\right), \tag{271}$$

thereby establishing a pleasing parallelism with Shannon's formula [1] for the mutual information between a Gaussian random variable and its sum with an independent Gaussian random variable. Aside from a factor of $\frac{1}{2}$, in the Cauchy case, the role of the variance is taken by the strength. Incidentally, as shown in [2], if N is standard exponential on $(0, \infty)$, an independent X on $[0, \infty)$ can be found so that $X + N$ is exponential, in which case the formula (271) also applies because the ratio of strengths of exponentials is equal to the ratio of their means. More generally, if input and noise are independent non-centered Cauchy, their locations do not affect the mutual information, but they do affect their strengths, so, in that case, (271) holds provided that the strengths are evaluated for the centered versions of the Cauchy random variables.

99. It is instructive, as well as useful in the sequel, to obtain (271) through a more circuitous route. Since $Y_C = X_C + W_C$ is centered Cauchy with strength $\varsigma(Y_C) = \varsigma(X_C) + \varsigma(W_C)$, the *information density* (e.g., [41]) is defined as

$$\imath_{X_C;Y_C}(x; y) = \log \frac{\mathrm{d}P_{X_C Y_C}}{\mathrm{d}(P_{X_C} \times P_{Y_C})}(x, y) \tag{272}$$
$$= \log \frac{f_{Y_C|X_C}(y|x)}{f_{Y_C}(y)} \tag{273}$$
$$= \log \frac{\varsigma(Y_C)}{\varsigma(W_C)} + \log\left(1 + \frac{y^2}{\varsigma^2(Y_C)}\right) - \log\left(1 + \frac{(y-x)^2}{\varsigma^2(W_C)}\right). \tag{274}$$

Averaging with respect to $(X_C, Y_C) = (X_C, X_C + W_C)$, we obtain

$$I(X_C; Y_C) = \mathbb{E}[\imath_{X_C;Y_C}(X_C; Y_C)] \tag{275}$$
$$= \log \frac{\varsigma(Y_C)}{\varsigma(W_C)} + \log 4 - \log 4 = \log\left(1 + \frac{\varsigma(X_C)}{\varsigma(W_C)}\right). \tag{276}$$

100. If the strengths of output $Y = X + N$ and independent noise N are finite and their differential entropies are not $-\infty$, we can obtain a general representation of the mutual information without requiring that either input or noise be Cauchy. Invoking (56) and $I(X; X + N) = h(X + N) - h(N)$, we have

$$I(X; Y) = \log \frac{N_C(Y)}{N_C(N)} \tag{277}$$
$$= \log \frac{\varsigma(Y)}{\varsigma(N)} + D(N \,\|\, \varsigma(N)V) - D(Y \,\|\, \varsigma(Y)V), \tag{278}$$

since, as we saw in (49), the finiteness of the strengths guarantees the finiteness of the relative entropies in (278). We can readily verify the alternative representation in which strength is replaced by standard deviation, and the standard Cauchy V is replaced by standard normal W:

$$I(X;Y) = \frac{1}{2}\log\frac{N(Y)}{N(N)} \tag{279}$$

$$= \log\frac{\sigma(Y)}{\sigma(N)} + D(N\,\|\,\sigma(N)W) - D(Y\,\|\,\sigma(Y)W). \tag{280}$$

A byproduct of (278) is the upper bound

$$I(X;Y) \le \log\frac{\varsigma(Y)}{N_{\mathsf{C}}(N)} \tag{281}$$

$$= \log\frac{\varsigma(Y)}{\varsigma(N)} + D(N\,\|\,\varsigma(N)V), \tag{282}$$

where (281) follows from $N_{\mathsf{C}}(Y) \le \varsigma(Y)$, and (282) follows by dropping the last term on the right side of (278). Note that (281) is the counterpart of the upper bound given by Shannon [1] in which the standard deviation of Y takes the place of the strength in the numerator, and the square root of the noise entropy power takes the place of the entropy strength in the denominator. Shannon gave his bound three years before Kullback and Leibler introduced relative entropy in [42]. The counterpart of (282) with analogous substitutions of strengths by standard deviations was given by Pinsker [43], and by Ihara [44] for continuous-time processes.

101. We proceed to investigate the maximal mutual information between the (possibly non-Cauchy) input and its additive Cauchy-noise contaminated version.

Theorem 13. *Maximal mutual information: output strength constraint. For any* $\eta \ge \varsigma(W_{\mathsf{C}}) > 0$,

$$\max_{X:\,\varsigma(X+W_{\mathsf{C}})\le\eta} I(X;X+W_{\mathsf{C}}) = \log\frac{\eta}{\varsigma(W_{\mathsf{C}})}, \tag{283}$$

where W_{C} *is centered Cauchy independent of* X*. The maximum in* (283) *is attained uniquely by the centered Cauchy distribution with strength* $\eta - \varsigma(W_{\mathsf{C}})$*.*

Proof. For centered Cauchy noise, the upper bound in (282) simplifies to

$$I(X;X+W_{\mathsf{C}}) \le \log\frac{\varsigma(X+W_{\mathsf{C}})}{\varsigma(W_{\mathsf{C}})}, \tag{284}$$

which shows $\le$ in (283). If the input is centered Cauchy X_{C} with strength $\eta - \varsigma(W_{\mathsf{C}})$, then $\varsigma(X_{\mathsf{C}}+W_{\mathsf{C}}) = \eta$, and $I(X_{\mathsf{C}};X_{\mathsf{C}}+W_{\mathsf{C}})$ is equal to the right side in view of (271). □

102. In the information theory literature, the maximization of mutual information over the input distribution is usually carried out under a constraint on the average cost $\mathbb{E}[\mathsf{b}(X)]$ for some real-valued function b. Before we investigate whether the optimization in (283) can be cast into that conventional paradigm, it is instructive to realize that the maximization of mutual information in the case of input-independent additive Gaussian noise can be viewed as one in which we allow any input such that the *output variance* is constrained, and because the output variance is the sum of input and noise variances that the familiar optimization over variance constrained inputs obtains. Likewise, in the case of additive exponential noise and random variables taking nonnegative values, if we constrain the *output mean*, automatically we are constraining the input mean. In contrast, the output strength is not equal to the sum of Cauchy noise strength and the input strength, unless the input is Cauchy. Indeed, as we saw

in Theorem 1-(d), the output strength depends not only on the input strength but on the shape of its probability density function. Since the noise is Cauchy, (45) yields

$$\varsigma(X+W_{\mathsf{C}}) \le \eta \iff \varsigma_{2,\theta}(X) \le \varsigma(W_{\mathsf{C}}) + \eta, \text{ with } \theta = 2\log\frac{2\eta}{\eta+\varsigma(W_{\mathsf{C}})} \tag{285}$$

$$\iff \mathbb{E}\Big[\log\Big(\varsigma(W_{\mathsf{C}}) + \eta + X^2\Big)\Big] \le 2\,\log(2\eta), \tag{286}$$

which is the same input constraint found in [45] (see also Lemma 6 in [46] and Section V in [47]) in which η affects not only the allowed expected cost but the definition of the cost function itself. If X is centered Cauchy with strength $\eta - \varsigma(W_{\mathsf{C}})$, then (286) is satisfied with equality, in keeping with the fact that that input achieves the maximum in (283). Any alternative input with the same strength that produces output strength lower than or equal to η can only result in lower mutual information. However, as we saw in Item 29, we can indeed find input distributions with strength $\eta - \varsigma(W_{\mathsf{C}})$ that can produce output strength higher than η. Can any of those input distributions achieve $I(X;Y) > \log\frac{\eta}{\varsigma(W_{\mathsf{C}})}$? The answer is affirmative. If we let $X = V_{\beta,2}$, defined in (9), we can verify numerically that, for $\beta \in [0.8, 1)$,

$$I(X; X+V) > \log(\varsigma(X)+1). \tag{287}$$

We conclude that, at least for $\frac{\theta}{\varsigma(W_{\mathsf{C}})} \in (1, \varsigma(V_{0.8,2})) = (1, 3.126\ldots)$, the capacity–input–strength function satisfies

$$C(\theta) = \max_{X:\,\varsigma(X)\le\theta} I(X; X+W_{\mathsf{C}}) > \log\left(1 + \frac{\theta}{\varsigma(W_{\mathsf{C}})}\right). \tag{288}$$

103. Although not always acknowledged, the key step in the maximization of mutual information over the input distribution for a given random transformation is to identify the optimal *output* distribution. The results in Items 101 and 102 point out that it is mathematically more natural to impose constraints on the attributes of the observed noisy signal than on the transmitted noiseless signal. In the usual framework of power constraints, both formulations are equivalent as an increase in the gain of the receiver antenna (or a decrease in the front-end amplifier thermal noise) of κ dB has the same effect as an increase of κ dB in the gain of the transmitter antenna (or increase in the output power of the transmitted amplifier). When, as in the case of strength, both formulations lead to different solutions, it is worthwhile to recognize that what we usually view as transmitter/encoder constraints also involve receiver features.
104. Consider a multiaccess channel $Y_i = X_{1i} + X_{2i} + W_i$, where W_i is a sequence of strength $\varsigma(W)$ independent centered Cauchy random variables. While the capacity region is unknown if we place individual cost or strength constraints on the transmitters, it is easily solvable if we impose an output strength constraint. In that case, the capacity region is the triangle

$$C_\eta = \left\{(R_1, R_2) \in [0,\infty)^2 \colon R_1 + R_2 \le \log\frac{\eta}{\varsigma(W)}\right\}, \tag{289}$$

where $\eta > \varsigma(W)$ is the output strength constraint. To see this, note (a) the corner points are achievable thanks to Theorem 13; (b) if the transmitters are synchronous, a time-sharing strategy with Cauchy distributed inputs satisfies the output strength constraint in view of (107); (c) replacing the independent encoders by a single encoder which encodes both messages would not be able to achieve higher rate sum. It is also possible to achieve (289) using the successive decoding strategy invented by Cover [48] and Wyner [49] for the Gaussian multiple-access channel: fix $\alpha \in (0,1)$; to achieve

$R_1 = \alpha \log \frac{\eta}{\varsigma(W)}$ and $R_2 = (1-\alpha) \log \frac{\eta}{\varsigma(W)}$, we let the transmitters use random coding with sequences of independent Cauchy random variables with respective strengths

$$\varsigma_1 = \eta - \varsigma^{\alpha}(W)\eta^{1-\alpha} > 0, \tag{290}$$

$$\varsigma_2 = \varsigma^{\alpha}(W)\eta^{1-\alpha} - \varsigma(W) > 0, \tag{291}$$

which abide by the output strength constraint since $\varsigma_1 + \varsigma_2 + \varsigma(W) = \eta$, and

$$R_1 = \log\left(1 + \frac{\varsigma_1}{\varsigma_2 + \varsigma(W)}\right), \tag{292}$$

$$R_2 = \log\left(1 + \frac{\varsigma_2}{\varsigma(W)}\right), \tag{293}$$

a rate-pair which is achievable by successive decoding by using a single-user decoder for user 1, which treats the codeword transmitted by user 2 as noise; upon decoding the message of user 1, it is re-encoded and subtracted from the received signal, thereby presenting a single-user decoder for user 2 with a signal devoid of any trace of user 1 (with high probability).

105. The capacity per unit energy of the additive Cauchy-noise channel $Y_i = X_i + \lambda V_i$, where $\{V_i\}$ is an independent sequence of standard Cauchy random variables, was shown in [29] to be equal to $(4\lambda^2)^{-1} \log \mathrm{e}$, even though the capacity-cost function of such a channel is unknown. A corollary to Theorem 13 is that the capacity per unit output strength of the same channel is

$$\mathbf{C}_O = \frac{1}{\lambda} \max_{\eta \geq \lambda} \frac{\lambda}{\eta} \log \frac{\eta}{\lambda} = \frac{\log \mathrm{e}}{\lambda \, \mathrm{e}}. \tag{294}$$

By only considering Cauchy distributed inputs, the capacity per unit input strength is lower bounded by

$$\mathbf{C}_I \geq \max_{\gamma > 0} \frac{1}{\gamma} \log\left(1 + \frac{\gamma}{\lambda}\right) = \frac{\log \mathrm{e}}{\lambda} \tag{295}$$

but is otherwise unknown as it is not encompassed by the formula in [29].

106. We turn to the scenario, dual to that in Theorem 13, in which the input is Cauchy but the noise need not be. As Shannon showed in [1], if the input is Gaussian, among all noise distributions with given second moment, independent Gaussian noise is the least favorable. Shannon showed that fact applying the entropy power inequality to the numerator on the right side of (279), and then further weakened the resulting lower bound by replacing the noise entropy power in the denominator by its variance. Taking a cue from this simple approach, we apply the entropy strength inequality (124) to (277) to obtain

$$I(X_{\mathsf{C}}; X_{\mathsf{C}} + W) = \frac{1}{2} \log \frac{N_{\mathsf{C}}^2(Y)}{N_{\mathsf{C}}^2(W)} \tag{296}$$

$$\geq \frac{1}{2} \log \frac{N_{\mathsf{C}}^2(X_{\mathsf{C}}) + N_{\mathsf{C}}^2(W)}{N_{\mathsf{C}}^2(W)} \tag{297}$$

$$= \frac{1}{2} \log\left(1 + \frac{\varsigma^2(X_{\mathsf{C}})}{N_{\mathsf{C}}^2(W)}\right) \tag{298}$$

$$\geq \frac{1}{2} \log\left(1 + \frac{\varsigma^2(X_{\mathsf{C}})}{\varsigma_{\mathsf{C}}^2(W)}\right), \tag{299}$$

where (299) follows from $N_{\mathsf{C}}^2(W) \leq \varsigma_{\mathsf{C}}^2(W)$. Unfortunately, unlike the case of Gaussian input, this route falls short of showing that Cauchy noise of a given strength

is least favorable because the right side of (299) is strictly smaller than the Cauchy-input Cauchy-noise mutual information in (271). Evidently, while the entropy power inequality is tight for Gaussian random variables, it is not for Cauchy random variables as we observed in Item 39. For this approach to succeed showing that, under a strength constraint, the least favorable noise is centered Cauchy we would need that, if W is independent of standard Cauchy V, then $N_{\mathsf{C}}(V+W) - N_{\mathsf{C}}(W) \geq 1$. (See Item 119-(a).)

107. As in Item 102, the counterpart in the Cauchy-input case is more challenging due to the fact that, unlike variance, the output strength need not be equal to the sum of input and noise strength. The next two results give lower bounds which, although achieved by Cauchy noise, do not just depend on the noise distribution through its strength.

Theorem 14. *If X_{C} is centered Cauchy, independent of W with $0 < \varsigma(W) < \infty$, denote $Y = X_{\mathsf{C}} + W$. Then,*

$$I(X_{\mathsf{C}}; X_{\mathsf{C}} + W) \geq \log \frac{\varsigma(Y)}{\varsigma(W)} - \left| \log \frac{\varsigma(W)}{\varsigma(Y) - \varsigma(X_{\mathsf{C}})} \right|, \tag{300}$$

with equality if W is centered Cauchy.

Proof. Let us abbreviate $\varsigma = \varsigma(Y) - \varsigma(X_{\mathsf{C}})$. Consider the following chain:

$$\begin{aligned} D(Y \,\|\, \varsigma(Y)V) - D(W \,\|\, \varsigma(W)V) &= D(X_{\mathsf{C}} + W \,\|\, X_{\mathsf{C}} + \varsigma\, V) - D(W \,\|\, \varsigma(W)V) && (301)\\ &\leq D(W \,\|\, \varsigma\, V) - D(W \,\|\, \varsigma(W)V) && (302)\\ &= \log \frac{\varsigma(W)}{\varsigma} + \mathbb{E}\left[\log \frac{\varsigma^2 + W^2}{\varsigma^2(W) + W^2}\right] && (303)\\ &\leq \left| \log \frac{\varsigma(W)}{\varsigma} \right|, && (304) \end{aligned}$$

where

- (301) $\Longleftarrow$ X_{C} is centered Cauchy;
- (302) $\Longleftarrow$ relative entropy data processing theorem applied to a random transformation that consists of the addition of independent "noise" X_{C};
- (303) $\Longleftarrow$ both relative entropies are finite since $\varsigma(W) < \infty$;
- (304) $\Longleftarrow$ the elementary observation

$$\log \frac{\varsigma^2 + t^2}{\varsigma^2(W) + t^2} \leq \begin{cases} 0, & \varsigma < \varsigma(W); \\ 2 \log \frac{\varsigma}{\varsigma(W)}, & \varsigma \geq \varsigma(W). \end{cases} \tag{305}$$

The desired bound (300) now follows in view of (278). It holds with equality in W being centered Cauchy as, in that case, $\varsigma(Y) = \varsigma(X_{\mathsf{C}}) + \varsigma(W_{\mathsf{C}})$. □

Although the lower bound in Theorem 14 is achieved by a centered Cauchy, it does not rule out the existence of W such that $\varsigma(W) = \varsigma(W_{\mathsf{C}})$ and $I(X_{\mathsf{C}}; X_{\mathsf{C}} + W) < I(X_{\mathsf{C}}; X_{\mathsf{C}} + W_{\mathsf{C}})$.

108. For the following lower bound, it is advisable to assume for notational simplicity and without loss of generality that $\varsigma(X_{\mathsf{C}}) = 1$. To remove that restriction, we may simply replace W by $\varsigma(X_{\mathsf{C}})W$.

Theorem 15. *Let V be standard Cauchy independent of W. Then,*

$$I(V; V + W) \geq \log\left(1 + \frac{1}{\lambda(W)}\right), \tag{306}$$

where $\lambda(W)$ is the solution to

$$\mathbb{E}\left[\log \frac{(2+\lambda)^2+W^2}{\lambda^2+W^2}\right] = 2\log\left(1+\frac{1}{\lambda}\right). \tag{307}$$

Equality holds in (306) if W is a centered Cauchy random variable, in which case, $\lambda(W) = \varsigma(W)$.

Proof. It can be shown that, if $P_{XY} = P_X P_{Y|X} = P_Y P_{X|Y}$ and $Q_{Y|X}$ is an auxiliary random transformation such that $P_X Q_{Y|X} = Q_Y Q_{X|Y}$ where Q_Y is the response of $Q_{Y|X}$ to P_X, then

$$I(X;Y) = D(P_{X|Y} \| Q_{X|Y} \,|\, P_Y) + \mathbb{E}\left[\imath_{X;\overline{Y}}(X;Y)\right], \tag{308}$$

where $(X,Y) \sim P_X P_{Y|X}$ and the information density $\imath_{X;\overline{Y}}$ corresponds to the joint probability measure $P_X Q_{Y|X}$. We can particularize this decomposition of mutual information to the case where $P_X = P_V$, $P_{Y|X=x} = P_{W+x}$, $Q_{Y|X=x} = P_{W_{\mathsf{C}}+x}$ where W_{C} is centered Cauchy with strength $\lambda > 0$. Then, $P_X Q_{Y|X}$ is the joint distribution of V and $V + W_{\mathsf{C}}$, and

$$\imath_{X;\overline{Y}}(x;y) = \log\frac{\lambda}{1+\lambda} - \log\left(\lambda^2+(y-x)^2\right) + \log\left((1+\lambda)^2+y^2\right). \tag{309}$$

Taking expectation with respect to $(x,y) = (V, V+t)$, and invoking (52), we obtain

$$\mathbb{E}\left[\imath_{X;\overline{Y}}(V;V+t)\right] = \log\frac{\lambda}{1+\lambda} + \mathbb{E}\left[\log\frac{(1+\lambda)^2+(V+t)^2}{\lambda^2+t^2}\right] \tag{310}$$

$$= \log\frac{\lambda}{1+\lambda} + \log\frac{(2+\lambda)^2+t^2}{\lambda^2+t^2}. \tag{311}$$

Finally, taking expectation with respect to $t = W$, we obtain

$$\mathbb{E}\left[\imath_{X;\overline{Y}}(V;V+W)\right] = \mathbb{E}\left[\log\frac{(2+\lambda)^2+W^2}{\lambda^2+W^2}\right] - \log\left(1+\frac{1}{\lambda}\right). \tag{312}$$

If $\lambda = \lambda(W)$, namely, the solution to (307), then (306) follows as a result of (308). If $W = \varsigma(W)V$, then the solution to (307) is $\lambda(W) = \varsigma(W)$, and the equality in (306) can be seen by specializing (271) to $(\varsigma(X_{\mathsf{C}}), \varsigma(W_{\mathsf{C}})) = (1, \varsigma(W))$. □

109. As we just saw, if W is centered Cauchy, then the solution to (307) satisfies $\lambda(W) = \varsigma(W)$. On the other hand, we have

$$0.302.. = \varsigma(V_{2,2}) < \lambda(V_{2,2}) = 0.349\ldots \tag{313}$$

$$4.961... = \lambda(W) < \varsigma(W) = 5.845\ldots \tag{314}$$

if W has the probability density function in (100).

110. As the proof indicates, at the expense of additional computation, we may sharpen the lower bound in Theorem 15 to show

$$I(V;V+W) \geq \max_{\lambda>0}\left\{\mathbb{E}\left[\log\frac{(2+\lambda)^2+W^2}{\lambda^2+W^2}\right] - \log\left(1+\frac{1}{\lambda}\right)\right\}, \tag{315}$$

which is attained at the solution to

$$\frac{\lambda}{2+\lambda}\eta_{W^2}\left(\frac{1}{(2+\lambda)^2}\right) - \eta_{W^2}\left(\frac{1}{\lambda^2}\right) + \frac{1}{2\lambda+2} = 0. \tag{316}$$

111.

Theorem 16. *The rate–distortion function of a memoryless source whose distribution is centered Cauchy with strength $\varsigma(X)$ such that the time-average of the distortion strength is upper bounded by D is given by*

$$R(D) = \begin{cases} \log \frac{\varsigma(X)}{D}, & 0 < D < \varsigma(X); \\ 0, & D \geq \varsigma(X). \end{cases} \tag{317}$$

Proof. If $D \geq \varsigma(X)$, reproducing the source by $(0, \ldots, 0)$ results in time-average of the distortion strength equal to $\frac{1}{n}\sum_{i=1}^{n} \varsigma(X_i) = \varsigma(X)$. Therefore, $R(D) = 0$. If $0 < D < \varsigma(X)$, we proceed to determine the minimal $I(X;\hat{X})$ among all $P_{\hat{X}|X}$ such that $\varsigma(X - \hat{X}) \leq D$. For any such random transformation,

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) && (318)\\ &= h(X) - h(X - \hat{X}|\hat{X}) && (319)\\ &\geq h(X) - h(X - \hat{X}) && (320)\\ &= \log(4\pi\varsigma(X)) - h(X - \hat{X}) && (321)\\ &\geq \log(4\pi\varsigma(X)) - \log\big(4\pi\varsigma(X - \hat{X})\big) && (322)\\ &\geq \log \frac{\varsigma(X)}{D}, && (323) \end{aligned}$$

where (320) holds because conditioning cannot increase differential entropy, and (322) follows from Theorem 3 applied to $Z = X - \hat{X}$. The fact that there is an allowable $P_{\hat{X}|X}$ that achieves the lower bound with equality is best seen by letting $X = \hat{X} + Z$, where Z and $\hat{X}$ are independent centered Cauchy random variables with $\varsigma(Z) = D$ and $\varsigma(\hat{X}) = \varsigma(X) - D$. Then, $P_{\hat{X}|X}P_X = P_{X|\hat{X}}P_{\hat{X}}$ is such that the X marginal is indeed centered Cauchy with strength $\varsigma(X)$, and $\varsigma(X - \hat{X}) = D$. Recalling (271),

$$I(\hat{X};X) = \log\left(1 + \frac{\varsigma(X) - D}{\varsigma(Z)}\right) = \log \frac{\varsigma(X)}{D}, \tag{324}$$

and the lower bound in (323) can indeed be satisfied with equality. We are not finished yet since we need to justify that the rate–distortion function is indeed

$$R(D) = \min_{P_{\hat{X}|X}:\, \varsigma(X-\hat{X}) \leq D} I(X;\hat{X}), \tag{325}$$

which does not follow from the conventional memoryless lossy compression theorem with average distortion because, although the distortion measure is separable, it is not the average of a function with respect to the joint probability measure $P_{X\hat{X}}$. This departure from the conventional setting does not impact the direct part of the theorem (i.e., $\leq$ in (325)), but it does affect the converse and in particular the proof of the fact that the n-version of the right side of (325) single-letterizes. To that end, it is sufficient to show that the function of D on the right side of (325) is convex (e.g., see pp. 316–317 in [19]). In the conventional setting, this follows from the convexity of the mutual information in the random transformation since, with a distortion function $d(\cdot,\cdot)$, we have

$$\mathbb{E}[d(X,\hat{X}_\alpha)] = \alpha\, \mathbb{E}[d(X,\hat{X}_1)] + (1-\alpha)\, \mathbb{E}[d(X,\hat{X}_0)], \tag{326}$$

where $(X,\hat{X}_1) \sim P_X P^1_{\hat{X}|X}$, $(X,\hat{X}_0) \sim P_X P^0_{\hat{X}|X}$, and $(X,\hat{X}_\alpha) \sim \alpha P_X P^1_{\hat{X}|X} + (1-\alpha) P_X P^0_{\hat{X}|X}$. Unfortunately, as we saw in Item 35, strength is not convex on the probability measure so, in general, we cannot claim that

$$\varsigma(X - \hat{X}_\alpha) \leq \alpha\, \varsigma(X - \hat{X}_1) + (1-\alpha)\varsigma(X - \hat{X}_0). \tag{327}$$

The way out of this quandary is to realize that (327) is only needed for those $P^0_{\hat{X}|X}$ and $P^1_{\hat{X}|X}$ that attain the minimum on the right side of (325) for different distortion bounds D_0 and D_1. As we saw earlier in this proof, those optimal random transformations are such that $X - \hat{X}_0$ and $X - \hat{X}_1$ are centered Cauchy. Fortuitously, as we noted in (107), (327) does indeed hold when we restrict attention to mixtures of centered Cauchy distributions. □

Theorem 16 gives another example in which the Shannon lower bound to the rate–distortion function is tight. In addition to Gaussian sources with mean–square distortion, other examples can be found in [50]. Another interesting aspect of the lossy compression of memoryless Cauchy sources under strength distortion measure is that it is optimally successively refinable in the sense of [51,52]. As in the Gaussian case, this is a simple consequence of the stability of the Cauchy distribution and the fact that the strength of the sum of independent Cauchy random variables is equal to the sum of their respective strengths (Item 27).

112. The continuity of mutual information can be shown under the following sufficient conditions

Theorem 17. *Suppose that X_n is a sequence of real-valued random variables that vanishes in strength, Z is independent of X_n, $h(Z) > -\infty$ and $0 < \varsigma(Z) < \infty$. Then,*

$$\lim_{n\to\infty} I(X_n; X_n + Z) = 0. \tag{328}$$

Proof. Under the assumptions, $h(Z) \in \mathbb{R}$. Therefore, $I(X_n; X_n + Z) = h(X_n + Z) - h(Z)$, and (328) follows from Theorem 1-(m). □

113. The assumption $h(Z) > -\infty$ is not superfluous for the validity of Theorem 17 even though it was not needed in Theorem 1-(m). Suppose that Z is integer valued, and $X_n = (nL)^{-1} \in (0, \frac{1}{2})$ where $L \in \{2, 3, \ldots\}$ has probability mass function

$$P_L(k) = \frac{0.986551\ldots}{k \log_2^2 k}, \quad k = 2, 3, \ldots \tag{329}$$

Then, $I(X_n; X_n + Z) = H(X_n) = H(L) = \infty$, while $\mathbb{E}[|X_n|] = \frac{0.328289\ldots}{n}$, and therefore, $\varsigma(X_n) \to 0$.

114. In the case in which V^n and W^n are standard spherical multivariate Cauchy random variables with densities in (6), it follows from (7) that $\lambda_X V^n + \lambda_W W^n$ has the same distribution as $(|\lambda_X| + |\lambda_W|) V^n$. Therefore,

$$I(V^n; \lambda_X V^n + \lambda_W W^n) = h(\lambda_X V^n + \lambda_W W^n) - h(\lambda_W W^n) \tag{330}$$

$$= n \log\left(1 + \frac{|\lambda_X|}{|\lambda_W|}\right), \tag{331}$$

where we have used the scaling law $h(\alpha X^n) = n \log |\alpha| + h(X^n)$. There is no possibility of a Cauchy-counterpart of the celebrated log-determinant formula for additive Gaussian vectors (e.g., Theorem 9.2.1 in [41]) because, as pointed out in Item 7, $\mathbf{\Lambda}^{\frac{1}{2}} V^n + \bar{\mathbf{\Lambda}}^{\frac{1}{2}} W^n$ is not distributed according to the ellipsoidal density in (8) unless $\mathbf{\Lambda}$ and $\bar{\mathbf{\Lambda}}$ are proportional, in which case the setup reverts to that in (330).

115. To conclude this section, we leave aside additive noise models and consider the mutual information between a partition of the components of the standard spherical multivariate Cauchy density (6). If $\mathcal{I} \cap \mathcal{J} = \varnothing$, then (17) yields

$$I\big(\{V_i, i \in \mathcal{I}\}; \{V_j, j \in \mathcal{J}\}\big) = h_{|\mathcal{I}|} + h_{|\mathcal{J}|} - h_{|\mathcal{I}|+|\mathcal{J}|}, \tag{332}$$

where h_n stands for the right side of (17). For example, if $i \neq j$, then, in nats,

$$I(V_i;V_j) = 2\,h(V_1) - h(V_1,V_2) \tag{333}$$
$$= 2\log_{\mathrm{e}}(4\pi) - \tfrac{3}{2}\big(\log_{\mathrm{e}}(4\pi) + \gamma + \psi(\tfrac{3}{2})\big) - \log_{\mathrm{e}}\Gamma(\tfrac{3}{2}) \tag{334}$$
$$= \log_{\mathrm{e}}(8\pi) - 3 = 0.22417\ldots \tag{335}$$

More generally, the dependence index among the n random variables in the standard spherical multivariate Cauchy density is (see also [9,53]), in nats,

$$D(P_{V^n} \,\|\, P_{V_1} \times \cdots \times P_{V_n}) = n\,h(V_1) - h(V^n) \tag{336}$$
$$= \frac{n-1}{2}\log_{\mathrm{e}}(4\pi) + \log_{\mathrm{e}}\Gamma\left(\frac{n+1}{2}\right) - \frac{n+1}{2}\left(\gamma + \psi\left(\frac{n+1}{2}\right)\right) \tag{337}$$
$$= \begin{cases} \frac{n}{2}\log_{\mathrm{e}}(8\pi) + \sum_{k=1}^{\frac{n}{2}}\left(\log_{\mathrm{e}}(2k-1) - \frac{n+1}{2k-1}\right), & n \text{ even;} \\ \frac{n-1}{2}\log_{\mathrm{e}}(4\pi) + \sum_{k=1}^{\frac{n-1}{2}}\left(\log_{\mathrm{e}} k - \frac{n+1}{2k}\right), & n \text{ odd.} \end{cases} \tag{338}$$

116. The *shared information* of n random variables is a generalization of mutual information introduced in [54] for deriving the fundamental limit of interactive data exchange among agents who have access to the individual components and establish a dialog to ensure that all of them find out the value of the random vector. The shared information of X^n is defined as

$$S(X^n) = \min_{\Pi} \frac{1}{|\Pi|-1} D\left(P_{X^n} \,\middle\|\, \prod_{\ell=1}^{|\Pi|} P_{X(\mathcal{I}_\ell)}\right), \tag{339}$$

where $X(\mathcal{J}) = \{X_i, i \in \mathcal{J}\}$, with $\mathcal{J} \subset \mathcal{I} = \{1,\ldots,n\}$, and the minimum is over all partitions of $\mathcal{I}$:

$$\Pi = \{\mathcal{I}_\ell \neq \varnothing, \ell = 1,\ldots,|\Pi|\}, \text{ with } \cup_{\ell=1}^{|\Pi|}\mathcal{I}_\ell = \mathcal{I},\ \mathcal{I}_\ell \cap \mathcal{I}_j = \varnothing, \ell \neq j,$$

such that $|\Pi| > 1$. If we divide (338) by $n-1$, we obtain the *shared information* of n random variables distributed according to the standard spherical multivariate Cauchy model. This is a consequence of the following result, which is of independent interest.

Theorem 18. *If X^n are exchangeable random variables, any subset of which have finite differential entropy, then for any partition Π of $\{1,\ldots,n\}$,*

$$\frac{1}{|\Pi|-1} D\left(P_{X^n} \,\middle\|\, \prod_{\ell=1}^{|\Pi|} P_{X(\mathcal{I}_\ell)}\right) \geq \frac{1}{n-1} D(P_{X^n} \,\|\, P_{X_1} \times \cdots \times P_{X_n}). \tag{340}$$

Proof. Fix any partition Π with $|\Pi| = L \in \{2,\ldots,n-1\}$ chunks. Denote by n_ℓ the number of chunks in Π with cardinality $\ell \in \{1,\ldots,n-1\}$. Therefore,

$$\sum_{\ell=1}^{n-1} n_\ell = L, \text{ and } \sum_{\ell=1}^{n-1} \ell\, n_\ell = n. \tag{341}$$

By exchangeability, any chunk of cardinality k has the same differential entropy, which we denote by h_k. Then,

$$D\left(P_{X^n} \,\middle\|\, \prod_{\ell=1}^{|\Pi|} P_{X(\mathcal{I}_\ell)}\right) = -h_n + \sum_{\ell=1}^{n-1} n_\ell\, h_\ell, \tag{342}$$

and the difference of the left minus the right sides of (340) multiplied by $(n-1)(L-1)$ is readily seen to equal

$$
\begin{aligned}
&-(n-1)\,h_n + (n-1)\sum_{\ell=1}^{n-1} n_\ell\, h_\ell + (L-1)\,h_n - (L-1)n\,h_1 \\
&= ((n-1)n_1 - n(L-1))\,h_1 + (L-n)\,h_n + (n-1)\sum_{\ell=2}^{n-1} n_\ell\, h_\ell && (343)\\
&\geq \left((n-1)n_1 - n(L-1) + \sum_{\ell=2}^{n-1}(n-\ell)n_\ell\right) h_1 + \left(L-n+\sum_{\ell=2}^{n-1}(\ell-1)n_\ell\right) h_n && (344)\\
&= 0 && (345)
\end{aligned}
$$

where

- (344) $\Longleftarrow$ for all $\ell \in \{2,\ldots,n-1\}$,

$$h_\ell \geq \frac{\ell-1}{n-1}\,h_n + \frac{n-\ell}{n-1}\,h_1, \tag{346}$$

  since $h_1,\ldots,h_n$ is a concave sequence, i.e., $2\,h_k \geq h_{k-1} + h_{k+1}$ as a result of the sub-modularity of differential entropy.
- (345) $\Longleftarrow$ (341).

□

Naturally, the same proof applies to n discrete exchangeable random variables with finite joint entropy.

17. Outlook

117. We have seen that a number of key information theoretic properties pertaining to the Gaussian law are also satisfied in the Cauchy case. Conceptually, those extensions shed light on the underlying reason the conventional Gaussian results hold. Naturally, we would like to explore how far beyond the Cauchy law those results can be expanded. As far as the maximization of differential entropy is concerned, the essential step is to redefine strength tailoring it to the desired law: Fix a reference random variable W with probability density function f_W and finite differential entropy $h(W) \in \mathbb{R}$, and define the *W-strength* of a real valued random variable Z as

$$\varsigma_W(Z) = \inf\left\{\varsigma > 0\colon\ -\mathbb{E}\left[\log f_W\left(\frac{Z}{\varsigma}\right)\right] \leq h(W)\right\}. \tag{347}$$

For example,

(a) For $\alpha > 0$, $\varsigma_W(\alpha W) = \alpha$;

(b) if W is standard normal, then $\varsigma_W^2(Z) = \mathbb{E}[Z^2]$;

(c) if V is standard Cauchy, then $\varsigma_V(Z) = \varsigma(Z)$;

(d) if W is standard exponential, then $\varsigma_W(Z) = \mathbb{E}[Z]$ if $Z \geq 0$ a.s., otherwise, $\varsigma_W(Z) = \infty$;

(e) if W is standard ($\mu = 1$) Subbotin (108) with $p > 0$, then, $\varsigma_W^p(Z) = \mathbb{E}[|Z|^p]$;

(f) if W has the Rider distribution in (9), then $\varsigma_W(Z) = \varsigma_{\rho,\theta}(Z)$ defined in (126) for θ chosen as in (110);

(g) if W is uniformly distributed on $[-1,1]$, $\varsigma_W(Z) = \operatorname{ess\,sup}|Z|$;

(h) if W is standard Rayleigh, then $\varsigma_W(Z) = \inf\left\{\varsigma > 0\colon \mathbb{E}\left[\frac{Z^2}{\varsigma^2} - \log_{\mathrm{e}}\frac{Z^2}{2\varsigma^2}\right] \leq 2+\gamma\right\}$ if $Z \geq 0$ a.s., otherwise, $\varsigma_W(Z) = \infty$.

The pivotal Theorems 3 and 4 admit the following generalization.

Theorem 19. *Suppose $h(W) \in \mathbb{R}$ and $\varsigma > 0$. Then,*

$$\max_{Z\colon \varsigma_W(Z) \le \varsigma} h(Z) = h(W) + \log \varsigma. \tag{348}$$

Proof. Fix any Z in the feasible set. For any $\sigma \ge \varsigma_W(Z)$ such that $-\mathbb{E}\Big[\log f_W\Big(\frac{Z}{\sigma}\Big)\Big] \le h(W)$, we have

$$0 \le D(\sigma^{-1}Z \,\|\, W) = -h(Z) + \log \sigma - \mathbb{E}\left[\log f_W\left(\frac{Z}{\sigma}\right)\right] \tag{349}$$

$$\le -h(Z) + \log \sigma + h(W). \tag{350}$$

Therefore, $h(Z) \le h(W) + \log \varsigma_W(Z)$, by definition of $\varsigma_W(Z)$, thereby establishing $\le$ in (348). Equality holds since $\varsigma_W(\varsigma W) = \varsigma$. □

A corollary to Theorem 19 is a very general form of the Shannon lower bound for the rate–distortion function of a memoryless source Z such that the distortion is constrained to have W-strength not higher than D, namely,

$$R(D) \ge h(Z) - h(W) - \log D. \tag{351}$$

Theorem 19 finds an immediate extension to the multivariate case

$$\max_{Z^n\colon \varsigma_{W^n}(Z^n) \le \varsigma} h(Z^n) = h(W^n) + n \log \varsigma, \tag{352}$$

where, for W^n with $h(W^n) \in \mathbb{R}$, we have defined

$$\varsigma_{W^n}(Z^n) = \inf\Big\{\varsigma > 0\colon -\mathbb{E}\Big[\log f_{W^n}\Big(\varsigma^{-1}Z^n\Big)\Big] \le h(W^n)\Big\}. \tag{353}$$

For example, if W^n is zero-mean multivariate Gaussian with positive definite covariance $\mathbf{\Sigma}$, then $\varsigma^2_{W^n}(Z^n) = \frac{1}{n}\mathbb{E}\big[Z^{n\top}\mathbf{\Sigma}^{-1}Z^n\big]$.

118. One aspect in which we have shown that Cauchy distributions lend themselves to simplification unavailable in the Gaussian case is the single-parametrization of their likelihood ratio, which paves the way for a slew of closed-form expressions for f-divergences and Rényi divergences. It would be interesting to identify other multiparameter (even just scale/location) families of distributions that enjoy the same property. To that end, it is natural, though by no means hopeful, to study various generalizations of the Cauchy distribution such as the Student-t random variable, or more generally, the Rider distribution in (9). The information theoretic study of general stable distributions is hampered by the fact that they are characterized by their characteristic functions (e.g., p. 164 in [55]), which so far, have not lent themselves to the determination of relative entropy or even differential entropy.
119. Although we cannot expect that the cornucopia of information theoretic results in the Gaussian case can be extended to other domains, we have been able to show that a number of those results do find counterparts in the Cauchy case. Nevertheless, much remains to be explored. To name a few,

    (a) The concavity of the entropy-strength $N_{\mathsf{C}}(X + tV)$—a counterpart of Costa's entropy power inequality [40] would guarantee the least favorability of Cauchy noise among all strength-constrained noises as well as the entropy strength inequality

    $$N_{\mathsf{C}}(X + t\,V) \ge N_{\mathsf{C}}(t\,V) + N_{\mathsf{C}}(X). \tag{354}$$

    (b) Information theoretic analyses quantifying the approach to normality in the central limit theorem are well-known (e.g., [56–58]). It would be interesting to explore the decrease in the relative entropy (relative to the Cauchy law) of

independent sums distributed according to a law in the domain of attraction of the Cauchy distribution [55].

(c) Since de Bruijn's identity is one of the ancestors of the I-MMSE formula of [59], and we now have a counterpart of de Bruijn's identity for convolutions with scaled Cauchy, it is natural to wonder if there may be some sort of integral representation of the mutual information between a random variable and its noisy version contaminated by additive Cauchy noise. In this respect, note that counterparts for the I-MMSE formula for models other than additive Gaussian noise have been found in [60–62].

(d) Mutual information is robust against the addition of small non-Gaussian contamination in the sense that its effects are the same as if it were Gaussian [63]. The proof methods rely on Taylor series expansions that require the existence of moments. Any Cauchy counterparts (recall Item 77) would require substantially different methods.

(e) Pinsker [41] showed that Gaussian processes are information stable imposing only very mild assumptions. The key is that, modulo a factor, the variance of the information density is upper bounded by its mean, the mutual information. Does the spherical multivariate Cauchy distribution enjoy similar properties?

120. Although not surveyed here, there are indeed a number of results in the engineering literature advocating Cauchy models in certain heavy-tailed infinite-variance scenarios (see, e.g., [45] and the references therein.) At the end, either we abide by the information theoretic maxim that "there is nothing more practical than a beautiful formula", or we pay heed to Poisson, who after pointing out in [64] that Laplace's proof of the central limit theorem broke down for what we now refer to as the Cauchy law, remarked that "Mais nous ne tiendrons pas compte de ce cas particulier, quil nous suffira d'avoir remarqué à cause de sa singularité, et qui ne se recontre sans doute pas dans la pratique".

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Definite Integrals

$$\int_0^x \frac{1}{1+t^2}\,\mathrm{d}t = \arctan(x), \tag{A1}$$

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \log\cos(\pi t)\,\mathrm{d}t = \log\tfrac{1}{2}, \tag{A2}$$

$$\int_{-\infty}^{\infty} \frac{\log(1+t^2)}{1+t^2}\,\mathrm{d}t = \pi\log 4, \tag{A3}$$

$$\int_{-\infty}^{\infty} \frac{\log(\alpha^2 - 2\alpha t\cos\beta + t^2)}{1+t^2}\,\mathrm{d}t = \pi\log(1+\alpha^2+2\alpha|\sin\beta|), \tag{A4}$$

$$\int_{-\infty}^{\infty} \frac{\log(1+t^2)}{1+(\xi t-\kappa)^2}\,\mathrm{d}t = \frac{\pi}{\xi}\Big(\log\Big(\kappa^2+(\xi+1)^2\Big) - 2\log\xi\Big), \quad \xi>0, \tag{A5}$$

$$\kappa_{\beta,\rho}\int_{-\infty}^{\infty} \frac{\log_{\mathrm{e}}(1+|t|^\rho)}{(1+|t|^\rho)^\beta}\,\mathrm{d}t = \psi(\beta) - \psi\Big(\beta-\tfrac{1}{\rho}\Big), \quad \beta\rho>1, \tag{A6}$$

$$\int_{-\infty}^{\infty} \frac{\log_{\mathrm{e}}(1+\theta^2t^2)}{(1+t^2)^2} = \pi\Big(\log_{\mathrm{e}}(1+|\theta|) - \frac{|\theta|}{1+|\theta|}\Big), \tag{A7}$$

$$\int_{-\alpha}^{\alpha} \log_{\mathrm{e}}(t^2+\varsigma^2)\,\mathrm{d}t = 4\varsigma \arctan\left(\frac{\alpha}{\varsigma}\right) - 4\alpha + 2\alpha\,\log_{\mathrm{e}}\left(\alpha^2+\varsigma^2\right), \tag{A8}$$

$$\int_{-\infty}^{\infty} \frac{t^2}{(1+t^2)^2}\,\mathrm{d}t = \frac{\pi}{2}, \tag{A9}$$

$$\int_{-\infty}^{\infty} \frac{1}{(1+t^2)^2}\,\mathrm{d}t = \frac{\pi}{2}, \tag{A10}$$

$$\int_{-\infty}^{\infty} \frac{t^2}{(1+t^2)^3}\,\mathrm{d}t = \frac{\pi}{8}, \tag{A11}$$

$$\int_{-\infty}^{\infty} \frac{1}{(\beta^2+t^2)^\nu}\,\mathrm{d}t = \sqrt{\pi}\,\beta^{1-2\nu}\frac{\Gamma\left(\nu-\frac{1}{2}\right)}{\Gamma(\nu)}, \quad \nu > \frac{1}{2}, \tag{A12}$$

$$\int_{0}^{\infty} \frac{1}{(1+t^\rho)^\nu}\,\mathrm{d}t = \frac{\Gamma\left(\nu-\frac{1}{\rho}\right)\Gamma\left(1+\frac{1}{\rho}\right)}{\Gamma(\nu)}, \quad \nu > \frac{1}{\rho} > 0, \tag{A13}$$

$$\int_{0}^{\pi} \log(\alpha+\beta\,\cos\theta)\,\mathrm{d}\theta = \pi\,\log\left(\frac{\alpha}{2}+\frac{1}{2}\sqrt{\alpha^2-\beta^2}\right), \quad \alpha \geq |\beta| > 0, \tag{A14}$$

$$\int_{0}^{\pi} \log\left(\beta+\sqrt{\beta^2-1}\,\cos\theta\right)^\alpha \mathrm{d}\theta = \pi\,\mathrm{P}_\alpha(\beta), \quad \beta > 0, \tag{A15}$$

$$\int_{0}^{\infty} \frac{\mathrm{d}t}{\sqrt{1+t^2}\,\sqrt{\beta^2+t^2}} = \mathbf{K}\left(\sqrt{1-\beta^2}\right), \quad \beta \in (0,1), \tag{A16}$$

where

- (A2) is a special case of 4.384.21 in [24];
- (A3) is a special case of (A4);
- (A4) is 4.296.2 in [24];
- (A5) follows from (A4) by change of variable;
- (A6), with $\kappa_{\beta,\rho}$ defined in (10) and $\psi(\cdot)$ denoting the digamma function, follows from 4.256 in [24] by change of variable $x = (1+t^p)^{-\frac{1}{2n}}$ and $n = m\,p$;
- (A7) is a special case of 4.295.25 in [24];
- (A8) follows from 2.733.1 in [24];
- (A9)–(A10) follow from 3.252.6 in [24];
- (A11) can be obtained by integration by parts and (A10);
- (A12), with $\Gamma(\cdot)$ denoting the gamma function, is a special case of 3.251.11 in [24];
- (A13) can be obtained from 3.251.11 in [24] by change of variable;
- (A14) is 4.224.9 in [24];
- (A15) is 8.822.1 in [24] with $\mathrm{P}_\alpha(x)$ the Legendre function of the first kind, which is a solution to

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\left(1-x^2\right)\frac{\mathrm{d}u(x)}{\mathrm{d}x}\right) + \alpha(\alpha+1)\,u(x) = 0; \tag{A17}$$

- (A16) is a special case of 3.152.1 in [24] with the complete elliptic integral of the first kind defined as 8.112.1 in [24], namely,

$$\mathbf{K}(k) = \int_{0}^{\frac{\pi}{2}} \frac{\mathrm{d}\alpha}{\sqrt{1-k^2\sin^2\alpha}}, \quad |k| < 1. \tag{A18}$$

Note that MATHEMATICA defines the complete elliptic integral function `EllipticK` such that

$$\mathbf{K}(k) = \frac{\texttt{EllipticK}\left(\frac{-k^2}{1-k^2}\right)}{\sqrt{1-k^2}}, \quad |k| < 1. \tag{A19}$$

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. 623–656. [CrossRef]
2. Verdú, S. The exponential distribution in information theory. *Probl. Inf. Transm.* **1996**, *32*, 86–95.
3. Anantharam, V.; Verdú, S. Bits through queues. *IEEE Trans. Inf. Theory* **1996**, *42*, 4–18. [CrossRef]
4. Stam, A. Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Inf. Control.* **1959**, *2*, 101–112. [CrossRef]
5. Ferguson, T.S. A representation of the symmetric bivariate Cauchy distribution. *Ann. Math. Stat.* **1962**, *33*, 1256–1266. [CrossRef]
6. Fang, K.T.; Kotz, S.; Ng, K.W. *Symmetric Multivariate and Related Distributions*; CRC Press: Boca Raton, FL, USA, 2018.
7. Rider, P.R. Generalized Cauchy distributions. *Ann. Inst. Stat. Math.* **1958**, *9*, 215–223. [CrossRef]
8. Bouhlel, N.; Rousseau, D. A generic formula and some special cases for the Kullback–Leibler divergence between central multivariate Cauchy distributions. *Entropy* **2022**, *24*, 838. [CrossRef]
9. Abe, S.; Rajagopal, A.K. Information theoretic approach to statistical properties of multivariate Cauchy-Lorentz distributions. *J. Phys. A Math. Gen.* **2001**, *34*, 8727–8731. [CrossRef]
10. Tulino, A.M.; Verdú, S. Random matrix theory and wireless communications. *Found. Trends Commun. Inf. Theory* **2004**, *1*, 1–182. [CrossRef]
11. Widder, D.V. The Stieltjes transform. *Trans. Am. Math. Soc.* **1938**, *43*, 7–60. [CrossRef]
12. Kullback, S. *Information Theory and Statistics*; Dover: New York, NY, USA, 1968; Originally published in 1959 by John Wiley.
13. Wu, Y.; Verdú, S. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans. Inf. Theory* **2010**, *56*, 3721–3747. [CrossRef]
14. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain Markov process expectations for large time, I. *Commun. Pure Appl. Math.* **1975**, *28*, 1–47. [CrossRef]
15. Donsker, M.D.; Varadhan, S.R.S. Asymptotic evaluation of certain Markov process expectations for large time, III. *Commun. Pure Appl. Math.* **1977**, *29*, 369–461. [CrossRef]
16. Lapidoth, A.; Moser, S.M. Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Trans. Inf. Theory* **2003**, *49*, 2426–2467. [CrossRef]
17. Subbotin, M.T. On the law of frequency of error. *Mat. Sb.* **1923**, *31*, 296–301.
18. Kapur, J.N. *Maximum-Entropy Models in Science and Engineering*; Wiley-Eastern: New Delhi, India, 1989.
19. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: New York, NY, USA, 2006.
20. Dembo, A.; Cover, T.M.; Thomas, J.A. Information theoretic inequalities. *IEEE Trans. Inf. Theory* **1991**, *37*, 1501–1518. [CrossRef]
21. Han, T.S. *Information Spectrum Methods in Information Theory*; Springer: Heidelberg, Germany, 2003.
22. Vajda, I. *Theory of Statistical Inference and Information*; Kluwer: Dordrecht, The Netherlands, 1989.
23. Deza, E.; Deza, M.M. *Dictionary of Distances*; Elsevier: Amsterdam, The Netherlands, 2006.
24. Gradshteyn, I.S.; Ryzhik, I.M. *Table of Integrals, Series, and Products*, 7th ed.; Academic Press: Burlington, MA, USA, 2007.
25. Sason, I.; Verdú, S. f-divergence inequalities. *IEEE Trans. Inf. Theory* **2016**, *62*, 5973–6006. [CrossRef]
26. Nielsen, F.; Okamura, K. On f-divergences between Cauchy distributions. In Proceedings of the International Conference on Geometric Science of Information, Paris, France, 21–23 July 2021; pp. 799–807.
27. Eaton, M.L. Group Invariance Applications in Statistics. In *Proceedings of the Regional Conference Series in Probability and Statistics*; Institute of Mathematical Statistics: Hayward, CA, USA, 1989; Volume 1.
28. McCullagh, P. On the distribution of the Cauchy maximum-likelihood estimator. *Proc. R. Soc. London. Ser. A Math. Phys. Sci.* **1993**, *440*, 475–479.
29. Verdú, S. On channel capacity per unit cost. *IEEE Trans. Inf. Theory* **1990**, *36*, 1019–1030. [CrossRef]
30. Chyzak, F.; Nielsen, F. A closed-form formula for the Kullback–Leibler divergence between Cauchy distributions. *arXiv* **2019**, arXiv:1905.10965.
31. Verdú, S. Mismatched estimation and relative entropy. *IEEE Trans. Inf. Theory* **2010**, *56*, 3712–3720. [CrossRef]
32. Csiszár, I. *I*-Divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **1975**, *3*, 146–158. [CrossRef]
33. Sason, I.; Verdú, S. Bounds among f-divergences. *arXiv* **2015**, arXiv:1508.00335.
34. Abramowitz, M.; Stegun, I.A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*; US Government Printing Office: Washington, DC, USA, 1964; Volume 55.
35. Rényi, A. On measures of information and entropy. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*; Neyman, J., Ed.; University of California Press: Berkeley, CA, USA, 1961; pp. 547–561.

36. Gil, M.; Alajaji, F.; Linder, T. Rényi divergence measures for commonly used univariate continuous distributions. *Inf. Sci.* **2013**, *249*, 124–131. [CrossRef]
37. González, M. Elliptic integrals in terms of Legendre polynomials. *Glasg. Math. J.* **1954**, *2*, 97–99. [CrossRef]
38. Nielsen, F. Revisiting Chernoff information with likelihood ratio exponential families. *Entropy* **2022**, *24*, 1400. [CrossRef]
39. Fisher, R.A. Theory of statistical estimation. *Math. Proc. Camb. Math. Soc.* **1925**, *22*, 700–725. [CrossRef]
40. Costa, M.H.M. A new entropy power inequality. *IEEE Trans. Inf. Theory* **1985**, *31*, 751–760. [CrossRef]
41. Pinsker, M.S. *Information and Information Stability of Random Variables and Processes*; Holden-Day: San Francisco, CA, USA, 1964; Originally published in Russian in 1960.
42. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
43. Pinsker, M.S. Calculation of the rate of message generation by a stationary random process and the capacity of a stationary channel. *Dokl. Akad. Nauk* **1956**, *111*, 753–766.
44. Ihara, S. On the capacity of channels with additive non-Gaussian noise. *Inf. Control.* **1978**, *37*, 34–39. [CrossRef]
45. Fahs, J.; Abou-Faycal, I.C. A Cauchy input achieves the capacity of a Cauchy channel under a logarithmic constraint. In Proceedings of the 2014 IEEE International Symposium on Information Theory, Honolulu, HI, USA, 29 June–4 July 2014; pp. 3077–3081.
46. Rioul, O.; Magossi, J.C. On Shannon's formula and Hartley's rule: Beyond the mathematical coincidence. *Entropy* **2014**, *16*, 4892–4910. [CrossRef]
47. Dytso, A.; Egan, M.; Perlaza, S.; Poor, H.; Shamai, S. Optimal inputs for some classes of degraded wiretap channels. In Proceedings of the 2018 IEEE Information Theory Workshop, Guangzhou, China, 25–29 November 2018; pp. 1–7.
48. Cover, T.M. Some advances in broadcast channels. In *Advances in Communication Systems*; Viterbi, A.J., Ed.; Academic Press: New York, NY, USA, 1975; Volume 4, pp. 229–260.
49. Wyner, A.D. Recent results in the Shannon theory. *IEEE Trans. Inf. Theory* **1974**, *20*, 2–9. [CrossRef]
50. Berger, T. *Rate Distortion Theory*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1971.
51. Koshelev, V.N. Estimation of mean error for a discrete successive approximation scheme. *Probl. Inf. Transm.* **1981**, *17*, 20–33.
52. Equitz, W.H.R.; Cover, T.M. Successive refinement of information. *IEEE Trans. Inf. Theory* **1991**, *37*, 269–274. [CrossRef]
53. Kotz, S.; Nadarajah, S. *Multivariate t-Distributions and Their Applications*; Cambridge University Press: Cambridge, UK, 2004.
54. Csiszár, I.; Narayan, P. The secret key capacity of multiple terminals. *IEEE Trans. Inf. Theory* **2004**, *50*, 3047–3061. [CrossRef]
55. Kolmogorov, A.N.; Gnedenko, B.V. *Limit Distributions for Sums of Independent Random Variables*; Addison-Wesley: Reading, MA, USA, 1954.
56. Barron, A.R. Entropy and the central limit theorem. *Ann. Probab.* **1986**, *14*, 336–342. [CrossRef]
57. Artstein, S.; Ball, K.; Barthe, F.; Naor, A. Solution of Shannon's problem on the monotonicity of entropy. *J. Am. Math. Soc.* **2004**, *17*, 975–982. [CrossRef]
58. Tulino, A.M.; Verdú, S. Monotonic decrease of the non-Gaussianness of the sum of independent random variables: A simple proof. *IEEE Trans. Inf. Theory* **2006**, *52*, 4295–4297. [CrossRef]
59. Guo, D.; Shamai, S.; Verdú, S. Mutual information and minimum mean–square error in Gaussian channels. *IEEE Trans. Inf. Theory* **2005**, *51*, 1261–1282. [CrossRef]
60. Guo, D.; Shamai, S.; Verdú, S. Mutual information and conditional mean estimation in Poisson channels. *IEEE Trans. Inf. Theory* **2008**, *54*, 1837–1849. [CrossRef]
61. Jiao, J.; Venkat, K.; Weissman, T. Relations between information and estimation in discrete-time Lévy channels. *IEEE Trans. Inf. Theory* **2017**, *63*, 3579–3594. [CrossRef]
62. Arras, B.; Swan, Y. IT formulae for gamma target: Mutual information and relative entropy. *IEEE Trans. Inf. Theory* **2018**, *64*, 1083–1091. [CrossRef]
63. Pinsker, M.S.; Prelov, V.; Verdú, S. Sensitivity of channel capacity. *IEEE Trans. Inf. Theory* **1995**, *41*, 1877–1888. [CrossRef]
64. Poisson, S.D. Sur la probabilité des résultats moyens des observations. In *Connaisance des Tems, ou des Mouvemens Célestes a l'usage des Astronomes, et des Navigateurs, pour l'an 1827*; Bureau des longitudes: Paris, France, 1824; pp. 273–302.