On f-divergences between Cauchy distributions

Frank Nielsen, *Senior Member, IEEE,* Kazuki Okamura, *Non-Member*

***Abstract*—We prove that all f-divergences between univariate Cauchy distributions are symmetric. Furthermore, those f-divergences can be calculated as strictly increasing scalar functions of the chi-square divergence. We report a criterion which allows one to expand f-divergences as converging series of power chi divergences, and exemplifies the technique for some f-divergences between Cauchy distributions. In contrast with the univariate case, we show that the f-divergences between multivariate Cauchy densities are in general asymmetric although symmetric when the Cauchy scale matrices coincide. Then we prove that the square roots of the Kullback-Leibler and Bhattacharyya divergences between univariate Cauchy distributions yield complete metric spaces. Finally, we show that the square root of the Kullback-Leibler divergence between univariate Cauchy distributions can be isometrically embedded into a Hilbert space.**

***Index Terms*—Cauchy distributions; Location-scale families; f-divergences; Complex analysis; maximal invariant; Möbius transformations; conditionally negative-definite kernel.**

I. Introduction

The probability density function $p_{l,s}(x)$ of a random variable X following a Cauchy distribution $\mathrm{Cauchy}(l,s)$ [1] is

$$p_{l,s}(x) := \frac{s}{\pi(s^2+(x-l)^2)}, \quad x \in \mathbb{R}.$$

Parameter $l \in \mathbb{R}$ characterizes the median and parameter $s > 0$ is called the probable error [1]. The set of Cauchy distributions forms a location-scale family [2]:

$$\mathcal{C} = \left\{ p_{l,s}(x) = \frac{1}{s} p\left(\frac{x-l}{s}\right) \ : \ (l,s) \in \mathbb{R} \times \mathbb{R}_{++} \right\},$$

where $p(x)$ denotes the Cauchy standard density and $\mathbb{R}_{++}$ denotes the set of positive real numbers:

$$p(x) := \frac{1}{\pi(1+x^2)}. \tag{1}$$

Thus parameters $l \in \mathbb{R}$ and $s > 0$ denote the location parameter and the scale parameter of the Cauchy distribution, respectively. The Cauchy distributions belong to two larger families of distributions: Namely, the t-Student distributions (for $\nu = 1$ degrees of freedom) and the q-normal distributions [3] (for $q = 2$). In Physics, the Cauchy distributions are also sometimes called the Lorentzian distributions [4] or the Breit–Wigner distribution [5].

In information theory, the Cauchy distributions have been used to model and analyze severe non-Gaussian noise with infinite variance ([6], [7], [8]) modeling so-called Cauchy channels in [9], [10]. Cauchy distributions are t-Student distributions with one degree of freedom and thus have heavier tails than Gaussian distributions which make them attractive for analyzing a variety of stochastic phenomena in source/channel coding and quantization [11]. The fat tail property has also been used in machine learning for stochastic low-dimensional embedding by t-SNE [12] and Cauchy distributions were used as a prior in the decoder of a Variational Auto-Encoder (VAE) [13] where it is shown to promote sparse regularization. The Cauchy distributions are q-Gaussians for $q = 2$ [14]: They form a deformed exponential family which plays an important role in non-extensive thermodynamics and anomaly detection [3] and can be characterized as maximum entropy distributions with respect to Tsallis q-entropy [15]. Thus it is important to consider statistical divergences between Cauchy densities for measuring their discrepancies and impact on these above-mentioned tasks [13].

To measure the dissimilarity between any two Cauchy distributions, we consider the class of f-divergences [16], [17]. An f-divergence [18] between any two continuous probability distributions with full support $\mathcal{X} = \mathbb{R}$ is defined by

$$I_f(p:q) := \int_{\mathbb{R}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x, \tag{2}$$

where $f(u)$ is a convex function defined on $(0,\infty)$ called the generator. The generator $f(u)$ of an f-divergence shall be chosen such that

- $f(u)$ is convex to ensure positivity (i.e., $I_f(p,q) \geq 0$) since by Jensen's inequality it comes that $I_f(p:q) \geq f(1)$,
- $f(1) = 0$ so that $I_f(p:q) \geq f(1) = 0$, and
- $f(u)$ is strictly convex at $u = 1$ to ensure reflexivity $I_f(p:q) = 0$ if and only if $p(x) = q(x)$ for all $x \in \mathcal{X}$.

When the integral of (2) diverges, we have $I_f(p:q) = +\infty$. For example, the KLD between a standard Cauchy distribution and a standard normal distribution diverges while the reverse KLD is finite.

The most celebrated f-divergence is the Kullback-Leibler divergence [19] (KLD), an f-divergence obtained for the generator $f_{\mathrm{KL}}(u) = -\log u$. The KLD is also called the relative entropy [19]. The class of f-divergences plays a prominent role in information theory: They are a jointly convex measure of dissimilarities that satisfies the lumping property (also called coarse graining property in [20]). That is,

This paper was presented in part in the Geometric Science of Information (2021), Lecture Notes in Computer Science, Volume 12829, pages 799–807, Springer 2021.

Manuscript received on xxx (*Corresponding author: Frank Nielsen.*)

Frank Nielsen is with Sony Computer Science Laboratories, Inc., Japan (e-mail: Frank.Nielsen@acm.org)

Kazuki Okamura is with Department of Mathematics, Faculty of Science, Shizuoka University, Japan (e-mail: okamura.kazuki@shizuoka.ac.jp). He was supported by JSPS KAKENHI 19K14549.

if $\mathcal{X}' = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ is any partition of $\mathcal{X}$ inducing probability mass functions $p'_i = \int_{x \in \mathcal{X}_i} p(x) \mathrm{d}x$ and $q'_i = \int_{x \in \mathcal{X}_i} q(x) \mathrm{d}x$ for $i \in \{1, \ldots, n\}$, then we have $I_f(p' : q') \leq I_f(p : q)$, where $I_f(p' : q') = \sum_{i=1}^n p'_i f\left(\frac{q'_i}{p'_i}\right)$. See [18], [21].

In general, the f-divergences are oriented dissimilarities: That is, we have $I_f(p : q) \neq I_f(q : p)$ (e.g., the KLD). For example, the KLD between two normal distributions q_{μ_1,σ_1} and q_{μ_2,σ_2} is asymmetric [19]:

$$D_{\mathrm{KL}}(q_{\mu_1,\sigma_1} : q_{\mu_2,\sigma_2}) = \frac{1}{2}\left(\left(\frac{\sigma_1}{\sigma_2}\right)^2 + \frac{(\mu_2-\mu_1)^2}{\sigma_2^2} - 1 + 2\ln\frac{\sigma_2}{\sigma_1}\right) \neq D_{\mathrm{KL}}(q_{\mu_2,\sigma_2} : q_{\mu_1,\sigma_1}).$$

Notice that the normal distributions form a location-scale family with standard density $q(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$ with parameters $\mu \in \mathbb{R}$ and $s \in \mathbb{R}_{++}$ denoting the mean and standard-deviation, respectively.

We have $I_f = I_g$ whenever there exists $\lambda \in \mathbb{R}$ such that $f(u) = g(u) + \lambda(u-1)$. The reverse f-divergence $I_f(q : p)$ can be obtained as a forward f-divergence for the conjugate generator $f^*(u) := uf\left(\frac{1}{u}\right)$ (convex with $f^*(1) = 0$ and strictly convex at 1): $I_f(q : p) = I_{f^*}(p : q)$. Thus an f-divergence is symmetric when there exists a real λ such that $f(u) = uf\left(\frac{1}{u}\right) + \lambda(u-1)$, and f-divergences can always be symmetrized by taking the generator $s_f(u) = \frac{1}{2}(f(u) + uf\left(\frac{1}{u}\right))$.

Remark 1. *Historically, the f-divergences were independently defined and studied by Csiszár [17] and Ali and Silvey [16]. Ali and Silvey [16] further considered a monotonically increasing function g of the f-divergence (i.e., the (f,g)-divergence $g(I_f(p : q))$ which allows them to consider the skewed Bhattacharyya distances as a family of (f,g)-divergences. Csiszár [17] and Ali and Silvey [16] gave definitions of f-divergences with different conventions: While Csiszár defined $I_f^{\mathrm{C}}(p : q) := \int_{\mathbb{R}} q(x) f\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x$, Ali and Silvey defined their "class of coefficient of divergences" [16] by $I_f^{\mathrm{AS}}(p : q) := \int_{\mathbb{R}} p(x) f\left(\frac{q(x)}{p(x)}\right) \mathrm{d}x$. Thus we have $I_f^{\mathrm{C}}(p : q) = I_f^{\mathrm{AS}}(q : p) = I_{f^*}^{\mathrm{C}}(p : q)$.*

Calculating the definite integrals of f-divergences in (2) may be a difficult task: For example, the formula for the KLD between any two Cauchy densities p_{l_1,s_1} and p_{l_2,s_2} was only obtained[1] [23] in 2019:

$$\begin{aligned} D_{\mathrm{KL}}(p_{l_1,s_1} : p_{l_2,s_2}) &:= I_{f_{\mathrm{KL}}}(p : q) \\ &= \int p_{l_1,s_1}(x) \log \frac{p_{l_1,s_1}(x)}{p_{l_2,s_2}(x)} \mathrm{d}x \quad (3) \\ &= \log\left(\frac{(s_1+s_2)^2 + (l_1-l_2)^2}{4s_1s_2}\right). \end{aligned}$$

Let $\lambda_i = (l_i, s_i) \in \mathbb{R} \times \mathbb{R}_{++}, i = 1, 2$. Then we can rewrite the KLD formula of (3) as

$$D_{\mathrm{KL}}(p_{\lambda_1} : p_{\lambda_2}) = \log\left(1 + \frac{1}{2}\chi(\lambda_1, \lambda_2)\right),$$

[1]This result can also be shown by complex analysis. See [22].

where

$$\chi(\lambda_1, \lambda_2) := \frac{(l_1-l_2)^2 + (s_1-s_2)^2}{2s_1s_2} = \frac{\|\lambda_1 - \lambda_2\|^2}{2s_1s_2}. \quad (4)$$

Interestingly, we observe that the KLD between Cauchy distributions is symmetric: $D_{\mathrm{KL}}(p_{l_1,s_1} : p_{l_2,s_2}) = D_{\mathrm{KL}}(p_{l_2,s_2} : p_{l_1,s_1})$. Another important f-divergence is the Pearson (P) chi-square divergence [19] defined between probability density functions $p(x)$ and $q(x)$ as:

$$D_\chi^P(p : q) := \int \frac{(p(x)-q(x))^2}{p(x)} \mathrm{d}x = \int \frac{q^2(x)}{p(x)} \mathrm{d}x - 1.$$

We have $D_\chi^P(p : q) = I_{\chi^P}(p : q)$ where the generator is $\chi^P(u) = (u-1)^2$. The reverse Pearson chi-square divergence is called the Neyman (N) chi-square divergence:

$$D_\chi^N(p : q) := \int \frac{(p(x)-q(x))^2}{q(x)} \mathrm{d}x = \int \frac{p^2(x)}{q(x)} \mathrm{d}x - 1.$$

It is an f-divergence for the generator $\chi^N(u) = (\chi^P)^*(u) = \frac{1}{u}(u-1)^2$: $D_\chi^N(p : q) = I_{\chi^N}(p : q) = D_\chi^P(q : p)$.

The Pearson and Neyman χ^2-divergences for normal distributions could be asymmetric when the integrals do not diverge (see Lemma 1 in [24] and Eq. (86) and (90) in [25]). But the Pearson and Neyman χ^2-divergences between Cauchy densities coincide because they are symmetric [14]:

$$D_\chi(p_{\lambda_1} : p_{\lambda_2}) := D_\chi^N(p_{\lambda_1} : p_{\lambda_2}) = D_\chi^P(p_{\lambda_1} : p_{\lambda_2}) = \chi(\lambda_1, \lambda_2),$$

hence the naming of the function $\chi(\cdot,\cdot)$ in (4). We refer readers to [26], [27], [28] for some results of f-divergences.

In theory, the definite integrals of f-divergences can be reported in closed-form using elementary functions whenever possible using the Risch pseudo-algorithm [29] of symbolic computing. The Risch method is only a pseudo-algorithm because it relies on an oracle which shall answer whether certain expressions are constants or not. Nevertheless, this pseudo-algorithm is implemented with restrictions in many modern computer algebra systems [30]. In practice, we may estimate the f-divergence between two Cauchy densities (or q-Gaussian densities [31]) by the following Monte Carlo technique of sampling n i.i.d. Cauchy random variates $x_1, \ldots, x_n \sim p_{l_1,s_1}$ and estimating the f-divergence by

$$\hat{I}_f^{(n)}[p_{l_1,s_1} : p_{l_2,s_2}] = \frac{1}{n}\sum_{i=1}^n f\left(\frac{p_{l_2,s_2}(x_i)}{p_{l_1,s_1}(x_i)}\right).$$

Under mild conditions, this Monte Carlo estimator is consistent: $\lim_{n\to\infty} \hat{I}_f^{(n)}[p_{l_1,s_1} : p_{l_2,s_2}] = I_f[p_{l_1,s_1} : p_{l_2,s_2}]$.

Contributions and paper outline: In this work, we first prove in §II that all f-divergences between univariate Cauchy distributions are symmetric (Theorem 2) and can be expressed as a strictly increasing scalar function of the chi-squared divergence (Theorems 6 and 10). We illustrate this result by reporting the corresponding functions for the total variation distance, the Kullback-Leibler divergence, the LeCam-Vincze divergence, the squared Hellinger divergence, and the Jensen-Shannon divergence. Further results for the f-divergences between the circular Cauchy, wrapped Cauchy, and log-Cauchy distributions based on the invariance properties of the f-divergences are presented in §III. In §IV, we show that the

symmetric property of f-divergence holds for multivariate location-scale families including the normal and Cauchy families with prescribed matrix scales provided that the standard density is even, but does not hold for the general case of different matrix scales. We report conditions to expand the f-divergences as infinite series of higher-order chi divergences (Theorem 22), and instantiate the results for the Cauchy distributions in §V. Finally, we consider metrizations of the KLD and the Bhattacharyya distances (Theorems 28 and 30) and discuss isometric embedding into a Hilbert space of the KLD (Theorem 33) in §VI.

II. f-Divergences between Cauchy distributions are symmetric

A. Symmetric property of Cauchy f-divergences

Let $\|\lambda\| = \sqrt{\lambda_1^2 + \lambda_2^2}$ denote the Euclidean norm of a 2D vector $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}^2$. We state the main theorem:

Theorem 2. *All f-divergences between univariate Cauchy distributions p_λ and $p_{\lambda'}$ with $\lambda = (l, s)$ and $\lambda' = (l', s')$ are symmetric and can be expressed as*

$$I_f(p_\lambda : p_{\lambda'}) = h_f\left(\chi(\lambda, \lambda')\right)$$

where $h_f : \mathbb{R}_+ \to \mathbb{R}_+$ is a function (with $h_f(0) = 0$).

The proof does not yield explicit closed-form formula for the f-divergences as it can be in general difficult to calculate in closed forms and relies on McCullagh's complex parametrization [32] p_θ of the parameter of the Cauchy density $p_{l,s}$ with $\theta = l + is$:

$$p_\theta(x) = \frac{|\mathrm{Im}(\theta)|}{\pi |x - \theta|^2},$$

since $|x - (l + is)|^2 = ((x - l) + is)((x - l) - is) = (x - l)^2 + s^2$, where $|\cdot|$ denotes the complex modulus (absolute value for pure real numbers). The parameter space θ is the complex upper-half plane $\mathbb{H}$ and the Cauchy distributions are degenerated to Dirac distributions $\delta_l(x)$ whenever $s = 0$. $\mathbb{H}$ denote the complex parameter space of Cauchy distributions.

We make use of the special linear group $\mathrm{SL}(2, \mathbb{R})$ for θ the complex parameter:

$$\mathrm{SL}(2, \mathbb{R}) := \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} \; : \; a, b, c, d \in \mathbb{R}, \quad ad - bc = 1 \right\}.$$

Let $A.\theta := \frac{a\theta + b}{c\theta + d}$ (real linear fractional transformations) be the action of $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2, \mathbb{R})$. McCullagh proved that if $X \sim \mathrm{Cauchy}(\theta)$ then $A.X \sim \mathrm{Cauchy}\left(A.\theta\right)$, where $\theta \in \mathbb{H}$. For example, if $X \sim \mathrm{Cauchy}(is)$ then $\frac{1}{X} \sim \mathrm{Cauchy}(\frac{i}{s})$. Using the $\lambda = (l, s)$ parameterization, we have

$$\begin{aligned} l_A &= \frac{(al + b)(cl + d) + acs^2}{(cl + d)^2 + c^2 s^2}, \\ s_A &= \frac{|s|}{(cl + d)^2 + c^2 s^2}. \end{aligned}$$

We can also define an action of $\mathrm{SL}(2, \mathbb{R})$ to the real line support $\mathcal{X} = \mathbb{R}$ by $x \mapsto \frac{ax + b}{cx + d}, \quad x \in \mathbb{R}$, where we interpret $-\frac{d}{c} \mapsto \frac{a}{c}$ if $c \neq 0$. We remark that $d \neq 0$ if $c = 0$. This map is bijective between $\mathbb{R}$. We have the following invariance:

Lemma 3 (Invariance of Cauchy f-divergence under $\mathrm{SL}(2, \mathbb{R})$)**.** *For any $A \in \mathrm{SL}(2, \mathbb{R})$ and $\theta_1, \theta_2 \in \mathbb{H}$, we have*

$$I_f(p_{A.\theta_1} : p_{A.\theta_2}) = I_f(p_{\theta_1} : p_{\theta_2}).$$

Proof. We prove the invariance by the change of variable in the integral. We have

$$\begin{aligned} & I_f(p_{A.\theta_1} : p_{A.\theta_2}) \\ &= \int_{\mathbb{R}} \frac{\mathrm{Im}(A.\theta_1)}{\pi |x - A.\theta_1|^2} f\left(\frac{\mathrm{Im}(A.\theta_2)|x - A.\theta_1|^2}{\mathrm{Im}(A.\theta_1)|x - A.\theta_2|^2} \right) \mathrm{d}x. \end{aligned}$$

Since $A \in \mathrm{SL}(2, \mathbb{R})$, we have

$$\mathrm{Im}(A.\theta_i) = \frac{\mathrm{Im}(\theta_i)}{|c\theta_i + d|^2}, \quad i \in \{1, 2\}.$$

If $x = A.y$ then $\mathrm{d}x = \frac{\mathrm{d}y}{|cy + d|^2}$, and

$$|A.y - A.\theta_i|^2 = \frac{|y - \theta_i|^2}{|cy + d|^2 \; |c\theta_i + d|^2}, \quad i \in \{1, 2\}.$$

Hence we get:

$$\begin{aligned} & \int_{\mathbb{R}} f\left(\frac{\mathrm{Im}(A.\theta_2)|x - A.\theta_1|^2}{\mathrm{Im}(A.\theta_1)|x - A.\theta_2|^2} \right) \frac{\mathrm{Im}(A.\theta_1)}{\pi |x - A.\theta_1|^2} \mathrm{d}x \\ &= \int_{\mathbb{R}} f\left(\frac{\mathrm{Im}(\theta_2)|y - \theta_1|^2}{\mathrm{Im}(\theta_1)|y - \theta_2|^2} \right) \frac{\mathrm{Im}(\theta_1)}{\pi |y - \theta_1|^2} \mathrm{d}y = I_f\left(p_{\theta_1} : p_{\theta_2}\right). \end{aligned}$$

□

Let us notice that the Cauchy family is the only univariate location-scale family that is also closed by inversion [33]: That is, if $X \sim \mathrm{Cauchy}(l, s)$ then $\frac{1}{X} \sim \mathrm{Cauchy}(l', s')$. Therefore our results are specific to the Cauchy family and cannot be extended to any other location-scale family. However, the characterization by [33] yields some applications. See Appendix in [22] for more details.

We now prove Theorem 2 using the notion of maximal invariant of Eaton [34] (Chapter 2).

Let us rewrite the function χ with complex arguments as:

$$\chi(z, w) := \frac{|z - w|^2}{2\,\mathrm{Im}(z)\mathrm{Im}(w)}, \quad z, w \in \mathbb{H}. \tag{5}$$

Proposition 4 (McCullagh [32])**.** *The function χ defined in* (5) *is the* maximal invariant *for the action of the special linear group $\mathrm{SL}(2, \mathbb{R})$ to $\mathbb{H} \times \mathbb{H}$ defined by*

$$\mathrm{SL}(2, \mathbb{R}) \times (\mathbb{H} \times \mathbb{H}) \to \mathbb{H} \times \mathbb{H}$$

$$(A, (z, w)) \mapsto (A.z, A.w) = \left(\frac{az + b}{cz + d}, \frac{aw + b}{cw + d} \right),$$

where $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2, \mathbb{R}), \; z, w \in \mathbb{H}$. That is, we have

$$\chi(A.z, A.w) = \chi(z, w), \quad A \in \mathrm{SL}(2, \mathbb{R}), \; z, w \in \mathbb{H},$$

and it holds that for every $z, w, z', w' \in \mathbb{H}$ satisfying that $\chi(z', w') = \chi(z, w)$, there exists $A \in \mathrm{SL}(2, \mathbb{R})$ such that $(A.z, A.w) = (z', w')$.

Proposition 5 (Theorem 2.3 of Eaton [34])**.** *Let X, Y_1 and Y_2 be non-empty sets. Assume that a group G acts on X. Suppose that a map $f : X \to Y_1$ is maximal invariant for the action of G to X. Then, a map $h : X \to Y_2$ is invariant if and only if there exists a function $k : Y_1 \to Y_2$ such that $h(x) = k(f(x))$ for every $x \in X$.*

By Lemma 3 and Proposition 4, $I_f : \mathbb{H} \times \mathbb{H} \to [0, +\infty)$ is invariant and $\chi : \mathbb{H} \times \mathbb{H} \to [0, +\infty)$ is maximal invariant under the action of $SL(2, \mathbb{R})$ to $\mathbb{H} \times \mathbb{H}$. By Proposition 5, there exists a unique function $h_f : [0, \infty) \to [0, \infty)$ such that $h_f(\chi(z, w)) = I_f(p_z : p_w)$ for all $z, w \in \mathbb{H}$.

Therefore,

Theorem 6. *The f-divergence between two univariate Cauchy densities is symmetric and expressed as a function of the chi-squared divergence:*

$$I_f(p_{\theta_1} : p_{\theta_2}) = I_f(p_{\theta_2} : p_{\theta_1}) = h_f(\chi(\theta_1, \theta_2)), \quad \theta_1, \theta_2 \in \mathbb{H}. \tag{6}$$

The function h_f in the above display is uniquely determined.

Thus we have proven that the f-divergences between univariate Cauchy densities are all symmetric. Note that we have $h_f = h_{f^*}$. In general, the f-divergences between two Cauchy mixtures $m(x) = \sum_{i=1}^k w_i p_{l_i, s_i}(x)$ and $m'(x) = \sum_{i=1}^{k'} w'_i p_{l'_i, s'_i}(x)$ are asymmetric (i.e., $I_f(m : m') \neq I_f(m' : m)$) except when $k = k' = 1$.

Let us prove that all f-divergences between two densities of a location family with even standard density are symmetric.

Proposition 7. *Let $\mathcal{L}_p = \{p(x - l) \ : \ l \in \mathbb{R}\}$ denote a (potentially multivariate) location family with even standard density (i.e., $p(-x) = p(x)$) on the support $\mathcal{X} = \mathbb{R}$. Then all f-divergences between two densities p_{l_1} and p_{l_2} of $\mathcal{L}$ are symmetric: $I_f(p_{l_1} : p_{l_2}) = I_f(p_{l_2} : p_{l_1})$.*

Proof. Consider the change of variable $l_1 - x = y - l_2$ (so that $x - l_2 = l_1 - y$) with $\mathrm{d}x = -\mathrm{d}y$ and let us use the property that $p(z - l_1) = p(l_1 - z)$ since p is an even standard density. We have:

$$\begin{aligned}
I_f(p_{l_1} : p_{l_2}) &:= \int_{-\infty}^{+\infty} p(x - l_1) f\left(\frac{p(x - l_2)}{p(x - l_1)}\right) \mathrm{d}x, \\
&= \int_{+\infty}^{-\infty} p(l_1 - x) f\left(\frac{p(x - l_2)}{p(l_1 - x)}\right) (-\mathrm{d}y), \\
&= \int_{-\infty}^{+\infty} p(y - l_2) f\left(\frac{p(x - l_2)}{p(y - l_2)}\right) \mathrm{d}y, \\
&= \int_{-\infty}^{+\infty} p(y - l_2) f\left(\frac{p(l_1 - y)}{p(y - l_2)}\right) \mathrm{d}y, \\
&= \int_{-\infty}^{+\infty} p(y - l_2) f\left(\frac{p(y - l_1)}{p(y - l_2)}\right) \mathrm{d}y, \\
&=: I_f(p_{l_2} : p_{l_1}).
\end{aligned}$$

□

Thus f-divergences between location Cauchy densities are symmetric since $p(x) = p(-x)$ for the even standard Cauchy density of (1).

Remark 8. *Of course, not all statistical divergences between Cauchy densities are symmetric: For example, consider the statistical q-divergence [20] for a scalar $q \in [1, 3)$ between probability density functions $p(x)$ and $r(x)$:*

$$D_q(p : r) := \frac{1}{(1 - q) Z_q(p)} \left(1 - \int p^q(x) r^{1-q}(x) \mathrm{d}x\right),$$

where $Z_q(p) := \int p^q(x) \mathrm{d}x(x)$. Then the 2-divergence between two Cauchy densities p_{λ_1} and p_{λ_2} (with $\lambda_i = (l_i, s_i)$) is available in closed-form as a corresponding Bregman divergence [20]:

$$D_2(p_{\lambda_1} : p_{\lambda_2}) = \frac{\pi}{s_2} \|\lambda_1 - \lambda_2\|^2.$$

Thus $D_2(p_{\lambda_1} : p_{\lambda_2}) \neq D_2(p_{\lambda_2} : p_{\lambda_1})$ when $s_1 \neq s_2$.

Remark 9. *Note that since $I_f(p_{\theta_2} : p_{\theta_1}) = h_f(\chi(\theta_1, \theta_1))$, Lemma 3 can a posteriori be checked for the chi-squared divergence: For any $A \in \mathrm{SL}(2, \mathbb{R})$ and $\theta \in \mathbb{H}$, we have $\chi(A.\theta_1, A.\theta_2) = \chi(\theta_1, \theta_2)$, and therefore for any f-divergence, since we have $I_f(p_{A.\theta_1} : p_{A.\theta_2}) = I_f(p_{\theta_1} : p_{\theta_2})$ since*

$$\begin{aligned}
I_f(p_{A.\theta_2} : p_{A.\theta_1}) &= h_f(\chi(A.\theta_1, A.\theta_1)) \\
&= h_f(\chi(\theta_1, \theta_1)) = I_f(p_{\theta_2} : p_{\theta_1}).
\end{aligned}$$

To prove that $\chi(p_{A.\theta_1} : p_{A.\theta_2}) = \chi(p_{\theta_1} : p_{\theta_2})$, let us first recall that $\mathrm{Im}(A.\theta) = \frac{\mathrm{Im}(\theta)}{|c\theta + d|^2}$ and $|A.\theta_1 - A.\theta_2|^2 = \frac{|\theta_1 - \theta_2|^2}{|c\theta_1 + d|^2 \, |c\theta_2 + d|^2}$. Thus we have

$$\begin{aligned}
\chi(A.\theta_1, A.\theta_2) &= \frac{|A.\theta_1 - A.\theta_2|^2}{2\,\mathrm{Im}(A.\theta_1)\,\mathrm{Im}(A.\theta_2)}, \\
&= \frac{|\theta_1 - \theta_2|^2 \, |c\theta_1 + d|^2 \ |c\theta_2 + d|^2}{|c\theta_1 + d|^2 \ |c\theta_2 + d|^2 \ 2\,\mathrm{Im}(\theta_1)\mathrm{Im}(\theta_2)}, \\
&= \frac{|\theta_1 - \theta_2|^2}{2\,\mathrm{Im}(\theta_1)\,\mathrm{Im}(\theta_2)} = \chi(\theta_1, \theta_2).
\end{aligned}$$

Alternatively, we may also define a bivariate function $g_f(l, s)$ so that using the action of the location-scale group, we have:

$$h_f(\chi(\theta_1, \theta_2)) = g_f\left(\frac{l_1 - l_2}{s_2}, \frac{s_1}{s_2}\right),$$

where $\theta_1 = l_1 + is_1$ and $\theta_2 = l_2 + is_2$.

B. Strictly increasing function h_f

Let us prove now that the function h_f in Theorem 6 is a strictly increasing function.

Theorem 10. *Let $f : (0, \infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$ and $f \in C^1((0, 1)) \cap C^1((1, \infty))$ and $f'(x) < f'(y)$ for every $x < 1 < y$. Let $I_f(p_\lambda : p_{\lambda'})$ be the f-divergence between p_λ and $p_{\lambda'}$, specifically,*

$$I_f(p_\lambda : p_{\lambda'}) = \int_{\mathbb{R}} p_\lambda(x) f\left(\frac{p_{\lambda'}(x)}{p_\lambda(x)}\right) dx.$$

Let χ be the maximal invariant introduced by McCullagh. Let $h_f : (0, \infty) \to [0, \infty)$ be the function such that

$$h_f(\chi(\lambda, \lambda')) = I_f(p_\lambda : p_{\lambda'}), \ \lambda, \lambda' \in \mathbb{H}.$$

Then, h_f is a strictly increasing function.

$$\begin{tikzcd} X \arrow[r, "f"] \arrow[dr, "h"'] & Y_1 \arrow[d, "k"] \\ & Y_2 \end{tikzcd} \qquad \begin{tikzcd} \mathbb{H}\times\mathbb{H} \arrow[r, "\chi"] \arrow[dr, "I_f"'] & \mathbb{R}_+ \arrow[d, "h_f"] \\ & \mathbb{R}_+ \end{tikzcd}$$

TABLE I

COMMUTATIVE DIAGRAMS WHICH EXPLAIN (LEFT) EATON'S MAXIMAL INVARIANT/INVARIANT AND (RIGHT) THE f-DIVERGENCE INVARIANT FUNCTION EXPRESSED AS A FUNCTION OF THE MAXIMAL INVARIANT χ: $I_f = h_f(\chi)$.

The assumption of f is made in order to cover the important case of the total variation distance (an f-divergence for the generator $f(u) = \frac{1}{2}|u-1|$), in such case $f'(1)$ does not exist and $f'(z) = -1 < 1 = f'(w)$ if $z < 1 < w$.

Proof. Let $u \geq 0$. Then, $\chi(i, u+i) = u^2/2$ and hence,

$$h_f\left(\frac{u^2}{2}\right) = I_f(p_i : p_{u+i}).$$

Hence it suffices to show that $F_1(u) := I_f(p_i, p_{u+i})$ is a strictly increasing function. We see that

$$F_1(u) = \int_{\mathbb{R}} \frac{1}{\pi(x^2+1)} f\left(\frac{x^2+1}{(x-u)^2+1}\right) dx.$$

Then,

Lemma 11.

$$F_1'(u) = \int_{\mathbb{R}} \frac{2(x-u)}{\pi((x-u)^2+1)^2} f'\left(\frac{x^2+1}{(x-u)^2+1}\right) dx,\ u > 0,$$

where we let $f'(1) := 0$ for notational consistency if $f'(1)$ does not exist.

The proof of this lemma is given later. We assume this lemma and continue the proof of Theorem 10. It suffices to show that $F_1'(u) > 0$ for every $u > 0$. By the change-of-variable formula,

$$F_1'(u) = \frac{2}{\pi}\int_{\mathbb{R}} \frac{x}{(x^2+1)^2} f'\left(\frac{(x+u)^2+1}{x^2+1}\right) dx,\ u > 0,$$

We also see that

$$\begin{aligned} &\int_{\mathbb{R}} \frac{x}{(x^2+1)^2} f'\left(\frac{(x+u)^2+1}{x^2+1}\right) dx \\ &= \int_0^\infty \frac{x}{(x^2+1)^2} f'\left(\frac{(x+u)^2+1}{x^2+1}\right) dx \\ &+ \int_{-\infty}^0 \frac{x}{(x^2+1)^2} f'\left(\frac{(x+u)^2+1}{x^2+1}\right) dx \\ =&\int_0^\infty \frac{x}{(x^2+1)^2}\left(f'\left(\frac{(x+u)^2+1}{x^2+1}\right) - f'\left(\frac{(x-u)^2+1}{x^2+1}\right)\right) dx. \end{aligned}$$

Since f is convex and $f \in C^1((1,\infty))$,

$$f'\left(\frac{(x+u)^2+1}{x^2+1}\right) \geq f'\left(\frac{(x-u)^2+1}{x^2+1}\right),\quad 0 < x < \frac{u}{2}.$$

By the assumption that $f'(z) < f'(w)$ if $z < 1 < w$, it holds that

$$f'\left(\frac{(x+u)^2+1}{x^2+1}\right) > f'\left(\frac{(x-u)^2+1}{x^2+1}\right),\quad x > \frac{u}{2},$$

Hence,

$$\begin{aligned} &\int_{\mathbb{R}} \frac{x}{(x^2+1)^2} f'\left(\frac{(x+u)^2+1}{x^2+1}\right) dx \\ =&\int_0^\infty \frac{x}{(x^2+1)^2}\left(f'\left(\frac{(x+u)^2+1}{x^2+1}\right) - f'\left(\frac{(x-u)^2+1}{x^2+1}\right)\right) dx \\ &> 0. \end{aligned}$$

Thus we have Theorem 10. □

Proof of Lemma 11. We show this assertion for $u = u_0 > 0$. The following lemma is used for a justification of the differentiation under the integral sign for the computation of $F_1'(u)$.

Lemma 12. *For every $c > 1$,*

$$R_{f,c} := \sup_{1/c<a<b<c} \left|\frac{f(b)-f(a)}{b-a}\right| < +\infty.$$

Proof. Since f is convex,

$$\max_{x\in[1/c,c]} |f'(x)| \leq \max\{|f'(1/c)|, |f'(c)|\}.$$

Assume that $a < b \leq 1$ or $1 \leq a < b$. Then, by the mean-value theorem,

$$\left|\frac{f(b)-f(a)}{b-a}\right| = |f'(\xi)| \leq \max\{|f'(1/c)|, |f'(c)|\}.$$

Finally we assume that $a < 1 < b$. Then,

$$\begin{aligned} \left|\frac{f(b)-f(a)}{b-a}\right| &\leq \left|\frac{f(1)-f(a)}{1-a}\right| + \left|\frac{f(b)-f(1)}{b-1}\right| \\ &\leq 2\max\{|f'(1/c)|, |f'(c)|\}. \end{aligned}$$

□

We return to the proof of Lemma 11. For each fixed $u > 0$,

$$\frac{1}{1+u+u^2} \leq \frac{x^2+1}{(x-u)^2+1} \leq 1+u+u^2.$$

Hence, for some $c_0 > 1$,

$$\begin{aligned} \frac{1}{c_0} &< \inf_{x\in\mathbb{R}, u\in(0,2u_0)} \frac{x^2+1}{(x-u)^2+1} \\ &\leq \sup_{x\in\mathbb{R}, u\in(0,2u_0)} \frac{x^2+1}{(x-u)^2+1} < c_0. \end{aligned}$$

Assume that $0 < |h| < u_0$. Then, by Lemma 12,

$$\begin{aligned} &\frac{1}{x^2+1}\left|\frac{1}{h}\left(f\left(\frac{x^2+1}{(x-u_0-h)^2+1}\right) - f\left(\frac{x^2+1}{(x-u_0)^2+1}\right)\right)\right| \\ &\leq \frac{R_{f,c_0}}{x^2+1}\left|\frac{1}{h}\left(\frac{x^2+1}{(x-u_0-h)^2+1} - \frac{x^2+1}{(x-u_0)^2+1}\right)\right| \end{aligned}$$

$$= R_{f,c_0} \frac{2|x-u_0-h|+u_0}{((x-u_0-h)^2+1)((x-u_0)^2+1)}$$

$$\leq \frac{R_{f,c_0}(1+u_0)}{(x-u_0)^2+1}.$$

We also see that for $x \neq u_0/2$,

$$\lim_{h\to 0} \frac{1}{h}\left(f\left(\frac{x^2+1}{(x-u_0-h)^2+1}\right) - f\left(\frac{x^2+1}{(x-u_0)^2+1}\right)\right)$$

$$= (x^2+1)\frac{2(x-u_0)}{((x-u_0)^2+1)^2} f'\left(\frac{x^2+1}{(x-u_0)^2+1}\right).$$

We have excluded the case that $x = u_0/2$ because the derivative of f at 1 may not exist. Now the lemma follows from the dominated convergence theorem. □

It follows from Theorem 10 that the Chebyshev center [35] (also called the minimax center or circumcenter of the smallest or minimum enclosing ball [36]) of a finite set of n Cauchy distributions $p_{\lambda_1}, \ldots, p_{\lambda_n}$ with respect to a f-divergence :

$$\arg \min_{\lambda \in \mathbb{R}\times\mathbb{R}_{++}} \max_{i\in\{1,\ldots,n\}} I_f(p_\lambda : p_{\lambda_i})$$

does not depend on the generator f since $I_f(p_\lambda : p_{\lambda_i}) = h_f(\chi(\lambda, \lambda_i))$ for a strictly increasing function h_f. Thus we have

$$\arg \min_{\lambda \in \mathbb{R}\times\mathbb{R}_{++}} \max_{i\in\{1,\ldots,n\}} I_f(p_\lambda : p_{\lambda_i})$$

$$= h_f\left(\arg \min_{\lambda \in \mathbb{R}\times\mathbb{R}_{++}} \max_{i\in\{1,\ldots,n\}} h_f(\chi(\lambda, \lambda_i))\right).$$

That is, we can deduce the Chebyshev center of n Cauchy distributions with respect to an arbitrary f-divergence from their Chebyshev center with respect to the chi-square divergence. Similarly the Cauchy Voronoi diagrams with respect to f-divergences all coincide [14] since the bisector of two Cauchy distributions with respect to an f-divergence

$$\mathrm{Bi}_f(\lambda_i, \lambda_j) = \{\lambda \in \mathbb{R}\times\mathbb{R}_{++} \ : \ I_f(\lambda : \lambda_i) = I_f(\lambda : \lambda_j)\}$$

coincides with their bisector with respect to the chi-square divergence:

$$\mathrm{Bi}_\chi(\lambda_i, \lambda_j) = \{\lambda \in \mathbb{R}\times\mathbb{R}_{++} \ : \ \chi(\lambda : \lambda_i) = \chi(\lambda : \lambda_j)\}.$$

That is, we have $\mathrm{Bi}_f(\lambda_i, \lambda_j) = \mathrm{Bi}_\chi(\lambda_i, \lambda_j)$ for all $i, j \in \{1, \ldots, n\}$ and $i \neq j$. Thus the Cauchy Voronoi diagram with respect to a f-divergence amounts to the Cauchy Voronoi diagram with respect to the χ-square divergence [14].

Table II summarizes the symmetric closed-form f-divergences $I_f(p_\lambda : p_{\lambda'}) = h_f(\chi(p_\lambda : p_{\lambda'}))$ between two univariate Cauchy densities p_λ and $p_{\lambda'}$ that we obtained as a function h_f of the chi-squared divergence $\chi(p_\lambda : p_{\lambda'}) = \frac{\|\lambda-\lambda'\|^2}{2\lambda_2\lambda_2'}$ (with $h_f(0) = 0$). The detailed calculations are reported in the technical report [22].

C. The Chernoff information

We give an application of the symmetry of f-divergences. The Chernoff information [37] between two densities p_1 and p_2 is defined by:

$$C(p_1 : p_2) := -\log \min_{a\in(0,1)} \int p_1(x)^a p_2(x)^{1-a} \mathrm{d}x.$$

The Chernoff information provides an upper bound for the error probabilities of the MAP rule in Bayesian hypothesis testing [19] (Chapter 11). The Bhattacharyya divergence [38] is defined by

$$D_{\mathrm{Bhat}}(p : q) := -\log\left(\int \sqrt{p(x)q(x)}\,\mathrm{d}x\right). \tag{7}$$

Theorem 13. *For the univariate Cauchy location-scale families, the Chernoff information is equal to the Bhattacharyya divergence.*

Proof. Let

$$\Lambda(a) := \log \int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \mathrm{d}x.$$

This is finite for every $a \in \mathbb{R}$, and is in C^∞ class on $\mathbb{R}$.

We see that for every $a \in \mathbb{R}$,

$$\Lambda'(a) = \frac{\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} \mathrm{d}x}{\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \mathrm{d}x}.$$

By the symmetry of f-divergences stated in Theorem 6 where $f(u) = -u^{1-a}\log u$,

$$\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} \mathrm{d}x$$

$$= \int_{\mathbb{R}} p_{\theta_2}(x)^a p_{\theta_1}(x)^{1-a} \log \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \mathrm{d}x.$$

Hence, for $a = 1/2$,

$$\int_{\mathbb{R}} p_{\theta_1}(x)^{1/2} p_{\theta_2}(x)^{1/2} \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} dx = 0.$$

Hence, $\Lambda'(1/2) = 0$. We also see that

$$\Lambda''(a) =$$

$$\frac{\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \left(\log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\right)^2 \mathrm{d}x \int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \mathrm{d}x}{\left(\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \mathrm{d}x\right)^2}$$

$$- \frac{\left(\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \log \frac{p_{\theta_1}(x)}{p_{\theta_2}(x)} \mathrm{d}x\right)^2}{\left(\int_{\mathbb{R}} p_{\theta_1}(x)^a p_{\theta_2}(x)^{1-a} \mathrm{d}x\right)^2}.$$

By the Cauchy-Schwarz inequality, $\Lambda''(a) \geq 0$. Hence $\Lambda(a)$ takes its minimum at $a = 1/2$. □

Thus the Chernoff information between two Cauchy distributions p_{λ_1} and p_{λ_2} can be calculated from the Bhattacharyya coefficient $\mathrm{BC}(p_{\lambda_1} : p_{\lambda_2}) := \int \sqrt{p_{\lambda_1}(x)p_{\lambda_2}(x)}\mathrm{d}x$:

$$C(p_{\lambda_1} : p_{\lambda_2}) = -\log \mathrm{BC}(p_{\lambda_1} : p_{\lambda_2}).$$

f-divergence name	$f(u)$	$I_f(p:q)$	$h_f(u)$
Chi-squared divergence	$(u-1)^2$	$\int \frac{(p(x)-q(x))^2}{p(x)}\mathrm{d}x$	u
Total variation distance	$\frac{1}{2}\lvert u-1\rvert$	$\int \frac{1}{2}\lvert p(x)-q(x)\rvert\mathrm{d}x$	$\frac{2}{\pi}\arctan\left(\sqrt{\frac{u}{2}}\right)$
Kullback-Leibler divergence	$-\log u$	$\int p(x)\log\frac{p(x)}{q(x)}\mathrm{d}x$	$\log(1+\frac{1}{2}u)$
Jensen-Shannon divergence	$\frac{u}{2}\log\frac{2u}{1+u}-\frac{1}{2}\log\frac{1+u}{2}$	$\int\left(p(x)\log\frac{2p(x)}{p(x)+q(x)}+q(x)\log\frac{2q(x)}{p(x)+q(x)}\right)\mathrm{d}x$	$\log\left(\frac{2\sqrt{2+u}}{\sqrt{2+u}+\sqrt{2}}\right)$
Taneja T-divergence	$\frac{u+1}{2}\log\frac{u+1}{2\sqrt{u}}$	$\int\frac{p(x)+q(x)}{2}\log\frac{p(x)+q(x)}{2\sqrt{p(x)q(x)}}\mathrm{d}x$	$\log\left(\frac{1+\sqrt{1+\frac{u}{2}}}{2}\right)$,
LeCam-Vincze divergence	$\frac{(u-1)^2}{1+u}$	$\int\frac{(p(x)-q(x))^2}{p(x)+q(x)}\mathrm{d}x$	$2-4\sqrt{\frac{1}{2(u+2)}}$
squared Hellinger divergence	$\frac{1}{2}(\sqrt{u}-1)^2$	$\frac{1}{2}\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2\mathrm{d}x$	$1-\frac{2K\left(1-\left(1+u+\sqrt{u(2+u)}\right)^{-2}\right)}{\pi\sqrt{1+u+\sqrt{u(2+u)}}}$

TABLE II

CLOSED-FORM f-DIVERGENCES BETWEEN TWO UNIVARIATE CAUCHY DENSITIES EXPRESSED AS A FUNCTION h_f OF THE CHI-SQUARED DIVERGENCE $\chi[p_\lambda:p_{\lambda'}]=\frac{\|\lambda-\lambda'\|^2}{2\lambda_2\lambda_2'}$: $I_f(p_{\lambda_1}:p_{\lambda_2})=h_f(\chi(p_{\lambda_1}:p_{\lambda_2}))$. THE SQUARE ROOT OF THE KLD, LECAM, AND SQUARED HELLINGER DIVERGENCES BETWEEN CAUCHY DENSITIES YIELDS METRIC DISTANCES.

Since the Bhattacharyya coefficient can be recovered from the squared Hellinger divergence:

$$\mathrm{BC}(p_{\lambda_1}:p_{\lambda_2})=1-H^2(p_{\lambda_1}:p_{\lambda_2}),$$

we can use the closed-form of the squared Hellinger divergence in order to recover the closed-form formula of the Bhattacharyya coefficient. See Table II above for the closed-form of the squared Hellinger divergence. The Bhattacharyya and Chernoff divergences are not f-divergences because they are not separable divergences (due to the logarithm function).

III. f-DIVERGENCES OF RELATED DISTRIBUTIONS

There are several distributions which are related to the Cauchy distributions. In this section, we shall make use of the invariance properties of f-divergences to derive results for the circular Cauchy [39], [40], wrapped Cauchy [41] and log-Cauchy [42] families which are all related to the Cauchy distributions via various transformations either on the parameter space or on the observation space. In Appendix C, we consider f-divergences between the truncated Cauchy distributions.

First, consider the family of circular Cauchy distributions parameterized by complex parameters w belonging to the unit disk $\mathbb{D}=\{w\in\mathbb{C}\ :\ |w|<1\}$. A Circular Cauchy distribution (CC) is an angular distribution [40] playing an important role in circular and directional statistics [43] with the following probability density function:

$$p_w^{\mathrm{cc}}(\phi):=\frac{1}{2\pi}\frac{1-|w|^2}{|e^{i\phi}-w|^2},\quad \phi\in[-\pi,\pi).$$

Let $w=\rho e^{i\phi_0}$ be the polar form of w. The circular Cauchy density can be rewritten [39] as:

$$p_{\rho,\phi_0}^{\mathrm{cc}}(\phi)=\frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\phi-\phi_0)},\quad \phi\in[-\pi,\pi).$$

McCullagh [1] noticed that if $X\sim\mathrm{Cauchy}(\theta)$ then $Y=\frac{1+iX}{1-iX}$ follows $\mathrm{CCauchy}\left(\frac{1+i\theta}{1-i\theta}\right)$ with parameter complex $w=\frac{1+i\theta}{1-i\theta}$. Denote the complex parameter reciprocal conversion functions $\theta\leftrightarrow w$ by $w(\theta)=\frac{1+i\theta}{1-i\theta}$ and $\theta(w)=i\frac{1-w}{(1+w)}$. Let us write $w=a+ib$ for $a,b\in\mathbb{R}$.

Consider the subgroup of Möbius transformations $\mathrm{SL}_2(\mathbb{C})$ that maps $\mathbb{D}$ onto itself via the following transformations:

$$w\mapsto t_{\phi,a}(w):=e^{i\phi}\frac{w+a}{\bar{a}w+1},\quad \phi\in[-\pi,\pi),a\in\mathbb{C}.$$

The following invariance of f-divergences with respect to non-degenerate holomorphic mappings $t_{\phi,a}$ of parameters holds:

Proposition 14. *We have* $I_f(p_{w_1}^{\mathrm{cc}}:p_{w_2}^{\mathrm{cc}})=I_f\left(p_{t_{\phi,a}(w_1)}^{\mathrm{cc}}:p_{t_{\phi,a}(w_2)}^{\mathrm{cc}}\right)$ *for all* $\phi\in[-\pi,\pi)$ *and* $a\in\mathbb{C}$.

This is a consequence of Proposition 3, the invariance of f-divergences to diffeomorphisms, and the relationship between Cauchy and circular Cauchy distributions via complex parameterization. This proposition relies on the fact that $I_f(p_{\theta_1}:p_{\theta_2})=I_f(p_{\eta_1}:p_{\eta_2})$ for any smooth invertible transformations $\eta(\theta)$ (with smooth inverse $\theta(\eta)$), and the fact that f-divergences are invariant to diffeomorphisms on the sample space [44]. Thus by combining these diffeomorphism properties with the relationship of Cauchy with circular Cauchy distributions represented by their complex parameterizations and Proposition 4, we get Proposition 14.

The transformation $t_{\phi,a}(w)$ can be composed together to form a group of transformations of the unit complex disk $\mathbb{D}$ onto itself called the holomorphic automorphism group $\mathrm{Aut}(\mathbb{D})$ [45], [46]. Informally speaking, these transformations $\{t_{\phi,a}(w)\}_{\phi,a}$ represent the hyperbolic motions (i.e., translations and rotations).

Theorem 15 (f-divergences between circular Cauchy distributions)**.** *The f-divergence between two circular Cauchy distributions amounts to the f-divergence between two corresponding Cauchy distributions:* $I_f(p_{w_1}^{\mathrm{cc}}:p_{w_2}^{\mathrm{cc}})=I_f(p_{\theta(w_1)}:p_{\theta(w_2)})$. *It follows that all f-divergences between circular Cauchy distributions are symmetric and can be expressed as scalar functions of the chi-square divergence.*

This theorem follows from the invariance of f-divergences [20], [47] and Theorem 2. That is, let $Y=m(X)$ for a diffeomorphism m between the domains of continuous random variables X and Y. Denote by p_X and q_Y the probability densities functions with support $\mathcal{X}$. It is a key

property of f-divergences that f-divergences are invariant under diffeomorphic transformations [44], [48]:

$$I_f(p_{X_1} : p_{X_2}) = I_f(q_{Y_1} : q_{Y_2}).$$

This invariance of f-divergences further holds for non-deterministic mappings called sufficiency of stochastic kernels [21]. This result is related to the result obtained for the Kullback-Leibler divergence in [49] (Lemma 5.1). It is worth noting that the circular Cauchy distribution can be interpreted as the exit distribution of a Brownian motion starting at $w \in \mathbb{D}$ when reaching the unit boundary circle, see [1].

Next, consider the wrapped Cauchy distributions (WC) [41] with probability density functions:

$$p^{\mathrm{wc}}_{\mu,\gamma}(\phi) = \sum_{n=-\infty}^{\infty} \frac{\gamma}{\pi\left(\gamma^2 + (\phi - \mu + 2\pi n)^2\right)}, \quad -\pi \leq \phi < \pi,$$

where $\mu \in \mathbb{R}$ denotes the peak position of the unwrapped distribution and $\gamma > 0$ the scale parameter. Let $\eta = \mu + i\gamma$.

The density can be rewritten equivalently as

$$p^{\mathrm{wc}}_{\mu,\gamma}(\phi) = \frac{1}{2\pi} \frac{\sinh(\gamma)}{\cosh(\gamma) - \cos(\phi - \mu)}.$$

Since we have the following identity:

$$p^{\mathrm{cc}}_{w}(\phi) = p^{\mathrm{wc}}\left(\phi, \eta(w)\right), \quad \eta(w) = \frac{w - i}{w + i},$$

it follows the following theorem:

Theorem 16 (f-divergences between wrapped Cauchy distributions)**.** *The f-divergence between two wrapped Cauchy distributions amounts to the f-divergence between two corresponding Cauchy distributions: $I_f(p^{\mathrm{wc}}_{\eta_1} : p^{\mathrm{wc}}_{\eta_2}) = I_f(p_{\theta(\eta_1)} : p_{\theta(\eta_2)})$. It follows that the f-divergence between wrapped Cauchy distributions is symmetric and can be expressed as a scalar function of the chi-square divergence.*

Finally, consider the family $\mathcal{LC}$ of Log-Cauchy (LC) distributions (see [42], p. 443) and [50], p. 329):

$$\mathcal{LC} := \left\{ p^{\mathrm{lc}}_{\mu,\sigma}(y) = \frac{1}{y\pi}\left(\frac{\sigma}{(\log y - \mu)^2 + \sigma^2}\right) : \mu > 0, \sigma > 0 \right\},$$

defined on the positive real support $\mathcal{Y} = \mathbb{R}_{++}$.

If $X \sim \mathrm{Cauchy}(l, s)$ is a random variable following a Cauchy distribution then $Y = \exp(X)$ is a random variable following a log-Cauchy distribution with $\mu = l$ and $\sigma = s$. Reciprocally, if Y follows a log-Cauchy distribution $\mathrm{LogCauchy}(\mu, \sigma)$, then $X = \log(Y)$ follows a Cauchy distribution with $l = \mu$ and $s = \sigma$. In particular, if $Y \sim \mathrm{LogCauchy}(0, 1)$ then $X = \log(Y) \sim \mathrm{Cauchy}(0, 1)$.

We state the symmetric property of f-divergences between log-Cauchy distributions:

Theorem 17. *The f-divergences between two Log-Cauchy distributions $\mathrm{LogCauchy}(\mu_1, \sigma_1)$ and $\mathrm{LogCauchy}(\mu_2, \sigma_2)$ amount to the f-divergences between the two corresponding Cauchy distributions: $I_f(p^{\mathrm{lc}}_{\mu_1,\sigma_1} : p^{\mathrm{lc}}_{\mu_2,\sigma_2}) = I_f(p_{\mu_1,\sigma_1} : p_{\mu_2,\sigma_2})$. It follows that the f-divergences between two Log-Cauchy distributions are symmetric and can be expressed as a scalar function of the chi-square divergence.*

Proof. First, let us recall that the generic relationships between the probability density functions p_X and q_Y with corresponding real-valued random variables satisfying $Y = m(X)$ for a differentiable and invertible function m with $m'(x) \neq 0$ is

$$\begin{aligned} p_X(x) &= m'(x) \times q_Y(m(x)) = m'(x) \times q_Y(y), \\ q_Y(y) &= (m^{-1})'(y) \times p_X(m^{-1}(y)) = (m^{-1})'(y) \times p_X(x). \end{aligned}$$

Now consider the case $y = m(x) = \exp(x)$ with $m^{-1}(y) = \log(y)$, and $m'(x) = \exp(x)$ and $(m^{-1})'(y) = 1/y$. Let us make a change of variable in the f-divergence integral with $y = \exp(x)$ and $\mathrm{d}y = \exp(x)\mathrm{d}x$. We have $p_{l,s}(x)\mathrm{d}x = p^{\mathrm{lc}}_{\mu,\sigma}(y)\mathrm{d}y$, with $\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{y}$ and $\frac{\mathrm{d}y}{\mathrm{d}x} = e^y$. Let $q_{Y_i} \sim \mathrm{LogCauchy}(\mu_i, \sigma_i)$ and $p_{X_i} \sim \mathrm{Cauchy}(\mu_i, \sigma_i)$ for $i \in \{1, 2\}$. By a change of variable, we have:

$$\begin{aligned} I_f(q_{Y_1} : q_{Y_2}) &:= \int_{\mathbb{R}_{++}} q_{Y_1}(y) f\left(\frac{q_{Y_2}(y)}{q_{Y_1}(y)}\right) \mathrm{d}y \\ &= \int_{\mathbb{R}_{++}} (m^{-1})'(y) p_{X_1}(m^{-1}(y)) \\ &\qquad \times f\left(\frac{(m^{-1})'(y) p_{X_2}(m^{-1}(y))}{(m^{-1})'(y) p_{X_1}(m^{-1}(y))}\right) \mathrm{d}y \\ &= \int_{\mathbb{R}} p_{X_1}(x) f\left(\frac{p_{X_2}(x)}{p_{X_1}(x)}\right) \mathrm{d}x \\ &=: I_f(p_{X_1} : p_{X_2}). \end{aligned}$$

Then we use the symmetric property of the f-divergences of the Cauchy distributions to deduce the symmetry of the f-divergences between log-Cauchy distributions: $I_f(p^{\mathrm{lc}}_{\mu_1,\sigma_1} : p^{\mathrm{lc}}_{\mu_2,\sigma_2}) = I_f(p^{\mathrm{lc}}_{\mu_2,\sigma_2} : p^{\mathrm{lc}}_{\mu_1,\sigma_1})$. It follows that we have $I_f(p^{\mathrm{lc}}_{\mu_1,\sigma_1} : p^{\mathrm{lc}}_{\mu_2,\sigma_2}) = h_f(\chi((\mu_1, \sigma_1), (\mu_2, \sigma_2)))$. □

IV. f-Divergences between multivariate Cauchy distributions

For a symmetric positive-definite $d \times d$ matrix $P \succ 0$ and a d-dimensional location vector μ, the density of a random variable [48] $X_{\mu,P} := PX + \mu$ with $X \sim p(x)$ (standard density) is

$$p_{\mu,P}(x) := |P|^{-1}\, p(P^{-1}(x - \mu)). \tag{8}$$

A d-dimensional location-scale family is formed by the set of densities $\{p_{\mu,P}(x) : P \succ 0, \mu \in \mathbb{R}^d\}$. For example, the set of multivariate normal distributions (MVNs) forms a multidimensional location-scale family [48].

The probability density function of a d-dimensional Multi-Variate Cauchy distribution [51] (MVCs) with parameters $\mu \in \mathbb{R}^d$ and $\Sigma \succ 0$ be a $d \times d$ positive-definite symmetric matrix is defined by:

$$p_{\mu,\Sigma}(x) := \frac{C_d}{(\det \Sigma)^{1/2}} \left(1 + (x - \mu)^\top \Sigma^{-1} (x - \mu)\right)^{-(d+1)/2}, \tag{9}$$

$x \in \mathbb{R}^d$, where $C_d = \Gamma\left(\frac{d+1}{2}\right) / \pi^{\frac{d+1}{2}}$ is a normalizing constant, and $\Gamma(\cdot)$ denotes the gamma function. The MVCs form a multivariate location-scale family with standard density:

$$p(x) := C_d \left(1 + x^\top x\right)^{-(d+1)/2},$$

where matrix parameter $P = \Sigma^{\frac{1}{2}}$ in (8) denotes the square root of symmetric the positive-definite matrix of $\Sigma \succ 0$.

In this section, we shall prove that the f-divergences between any two densities of a multidimensional location-scale family with prescribed scale root matrix P and even standard density (i.e., $p(x) = p(-x)$) is symmetric, and then shows that the KLD between bivariate Cauchy distributions is asymmetric in general.

First, let us consider the case $\Sigma = I$: The corresponding set of multivariate Cauchy distributions yields a multivariate location subfamily $\{p_\mu(x) = p_{\mu,I}(x) \ : \ \mu \in \mathbb{R}^d\}$ with standard distribution $p(x) = p_{0,I}(x) = C_d \left(1 + x^\top x\right)^{-(d+1)/2}$. Since the standard density is even (i.e., $p(x) = p(-x)$), we can extend straightforwardly the result of Proposition 7 using a multidimensional change of variable in the integrals of f-divergences:

Proposition 18. *The f-divergences between any two densities of the multivariate location Cauchy family are symmetric: $I_f(p_{\mu_1} : p_{\mu_2}) = I_f(p_{\mu_2} : p_{\mu_1})$.*

Next, we consider the case of MVC location subfamilies with the prescribed matrix Σ (or equivalently P).

Proposition 19. *The f-divergences between any two densities of the multivariate location Cauchy family $\{p_{\mu,\Sigma} \ : \ \mu \in \mathbb{R}^d\}$ with prescribed matrix Σ is symmetric: $I_f(p_{\mu_1,\Sigma} : p_{\mu_2,\Sigma}) = I_f(p_{\mu_2,\Sigma} : p_{\mu_1,\Sigma})$.*

Proof. We shall use the following identities of f-divergences arising from the location-scale family group structure [48] (affine group):

$$I_f\left(p_{l_1,P_1} : p_{l_2,P_2}\right) = I_f\left(p : p_{P_1^{-1}(l_2-l_1),P_1^{-1}P_2}\right)$$
$$= I_f\left(p_{P_2^{-1}(l_1-l_2),P_2^{-1}P_1} : p\right).$$

Thus for the MVCs, we have:

$$I_f\left(p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}\right) = I_f\left(p : p_{\Sigma_1^{-\frac{1}{2}}(\mu_2-\mu_1),\Sigma_1^{-\frac{1}{2}}\Sigma_2^{\frac{1}{2}}}\right)$$
$$= I_f\left(p_{\Sigma_2^{-\frac{1}{2}}(\mu_1-\mu_2),\Sigma_2^{-\frac{1}{2}}\Sigma_1^{\frac{1}{2}}} : p\right).$$

It follows that when $\Sigma_1 = \Sigma_2 = \Sigma$, we get:

$$I_f\left(p_{\mu_1,\Sigma} : p_{\mu_2,\Sigma}\right) = I_f\left(p : p_{\Sigma^{-\frac{1}{2}}(\mu_2-\mu_1),I}\right)$$
$$= I_f\left(p_{\Sigma^{-\frac{1}{2}}(\mu_1-\mu_2),I} : p\right).$$

Recasting the equalities using the multivariate location Cauchy family, we obtain:

$$I_f\left(p_{\mu_1,\Sigma} : p_{\mu_2,\Sigma}\right) = I_f\left(p : p_{\Sigma^{-\frac{1}{2}}(\mu_2-\mu_1)}\right)$$
$$= I_f\left(p_{\Sigma^{-\frac{1}{2}}(\mu_1-\mu_2)} : p\right).$$

Since we proved in Proposition 18 for the multivariate Cauchy location family that $I_f(p_{\mu_1}, p_{\mu_2}) = I_f(p_{\mu_2}, p_{\mu_1})$ (with $p_\mu(x) := p_{\mu,I}(x)$), it follows that we have:

$$I_f\left(p_{\mu_1,\Sigma} : p_{\mu_2,\Sigma}\right) = I_f\left(p : p_{\Sigma^{-\frac{1}{2}}(\mu_2-\mu_1)}\right)$$
$$= I_f\left(p_{\Sigma^{-\frac{1}{2}}(\mu_2-\mu_1)} : p\right) = I_f\left(p_{\mu_2,\Sigma} : p_{\mu_1,\Sigma}\right).$$

□

However, contrary to the family of univariate Cauchy distributions, we have the following result:

Proposition 20. *There exist two bivariate Cauchy densities p_{μ_1,Σ_1} and p_{μ_2,Σ_2} such that*

$$D_{\mathrm{KL}}\left(p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}\right) \neq D_{\mathrm{KL}}\left(p_{\mu_2,\Sigma_2} : p_{\mu_1,\Sigma_1}\right).$$

Proof. We let $d = 2$. In this case, by recalling (9),

$$p_{\mu,\Sigma}(x) = \frac{C_2}{\det(\Sigma)}\left(\frac{1}{1 + (x-\mu)^\top \Sigma^{-1}(x-\mu)}\right)^{3/2}, \quad (10)$$

for $x \in \mathbb{R}^2$, where $C_2 = 1/(2\pi)$. By the change of variable in the integral [48], we have

$$D_{\mathrm{KL}}\left(p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}\right) =$$
$$D_{\mathrm{KL}}\left(p_{0,I_2} : p_{\Sigma_1^{-1/2}(\mu_2-\mu_1),\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2}}\right),$$

where I_2 denotes the unit 2×2 matrix.

Let

$$\mu_1 = 0, \Sigma_1 = I_2, \ \mu_2 = (0, \sqrt{n})^\top, \Sigma_2 = \begin{bmatrix} n & 0 \\ 0 & \frac{1}{n} \end{bmatrix},$$

where n is a natural number. We will show that $D_{\mathrm{KL}}\left(p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}\right) \neq D_{\mathrm{KL}}\left(p_{\mu_2,\Sigma_2} : p_{\mu_1,\Sigma_1}\right)$ for sufficiently large n.

By (10), we see that for $x = (x_1, x_2) \in \mathbb{R}^2$,

$$p_{\mu_1,\Sigma_1}(x) = p_{0,I_2}(x) = C_2\left(\frac{1}{1 + x_1^2 + x_2^2}\right)^{3/2},$$

and

$$p_{\mu_2,\Sigma_2}(x) = C_2\left(\frac{1}{1 + x_1^2/n + n(x_2 - \sqrt{n})^2}\right)^{3/2}.$$

Then,

$$\log\left(\frac{p_{\mu_1,\Sigma_1}(x)}{p_{\mu_2,\Sigma_2}(x)}\right) =$$
$$\frac{3}{2}\left(\log\left(1 + x_1^2/n + n(x_2 - \sqrt{n})^2\right) - \log\left(1 + x_1^2 + x_2^2\right)\right).$$

Therefore,

$$D_{\mathrm{KL}}\left(p_{\mu_1,\Sigma_1} : p_{\mu_2,\Sigma_2}\right) =$$
$$\frac{3C_2}{2}\int_{\mathbb{R}^2}\frac{\log(1 + x_1^2/n + n(x_2 - \sqrt{n})^2) - \log(1 + x_1^2 + x_2^2)}{(1 + x_1^2 + x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2$$

and

$$D_{\mathrm{KL}}\left(p_{\mu_2,\Sigma_2} : p_{\mu_1,\Sigma_1}\right) = D_{\mathrm{KL}}\left(p_{0,I_2} : p_{-\Sigma_2^{-1/2}\mu_2,\Sigma_2^{-1}}\right)$$
$$= \frac{3C_2}{2}\int_{\mathbb{R}^2}\frac{\log(1 + nx_1^2 + (x_2 + n)^2/n) - \log(1 + x_1^2 + x_2^2)}{(1 + x_1^2 + x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2$$

$$= \frac{3C_2}{2}\int_{\mathbb{R}^2}\frac{\log(1 + (x_1 + n)^2/n + nx_2^2) - \log(1 + x_1^2 + x_2^2)}{(1 + x_1^2 + x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2.$$

For ease of notation, let

$$F_n(x_1, x_2) := \log\left(1 + x_1^2/n + n(x_2 - \sqrt{n})^2\right) - \log\left(1 + x_1^2 + x_2^2\right).$$

Then, it suffices to show that

$$\lim_{n\to\infty}\int_{\mathbb{R}^2}\frac{F_n(x_1,x_2)}{(1+x_1^2+x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2=+\infty \tag{11}$$

Let

$$A_n:=\left\{(x_1,x_2):\frac{1+x_1^2/n+n(x_2-\sqrt{n})^2}{(1+(x_1+n)^2/n+nx_2^2)}>\sqrt{n}\right\}.$$

Then, $F_n(x_1,x_2)\geq(\log n)/2$ on A_n, and, for every $(x_1,x_2)\in\mathbb{R}^2$, there exists N_1 such that for every $n\geq N_1$, $(x_1,x_2)\in A_n$. By Fatou's lemma [52] (p. 93),

$$\lim_{n\to\infty}\int_{A_n}\frac{\mathrm{d}x_1\mathrm{d}x_2}{(1+x_1^2+x_2^2)^{3/2}}=\int_{\mathbb{R}^2}\frac{\mathrm{d}x_1\mathrm{d}x_2}{(1+x_1^2+x_2^2)^{3/2}}=\frac{1}{C_2}. \tag{12}$$

Hence, there exists N_2 such that for every $n\geq N_2$,

$$\int_{A_n}\frac{F_n(x_1,x_2)}{(1+x_1^2+x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2\geq\frac{\log n}{4C_2}.$$

We see that for every $n\geq 1$,

$$\inf_{(x_1,x_2)\in\mathbb{R}^2}F_n(x_1,x_2)\geq-\log 4-2\log(n+1).$$

Hence,

$$\int_{A_n^c}\frac{F_n(x_1,x_2)}{(1+x_1^2+x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2$$
$$\geq-(\log 4+2\log(n+1))\int_{A_n^c}\frac{\mathrm{d}x_1\mathrm{d}x_2}{(1+x_1^2+x_2^2)^{3/2}}.$$

By this and (12), there exists $N_3>N_2$ such that for every $n\geq N_3$,

$$-(\log 4+2\log(n+1))\int_{A_n^c}\frac{\mathrm{d}x_1\mathrm{d}x_2}{(1+x_1^2+x_2^2)^{3/2}}\geq-\frac{\log n}{8C_2}.$$

Thus we have (11). □

Remark 21. *We now give numerical computations for*

$$G(n):=\int_{\mathbb{R}^2}\frac{F_n(x_1,x_2)}{(1+x_1^2+x_2^2)^{3/2}}\mathrm{d}x_1\mathrm{d}x_2.$$

appearing in (11). *By using Mathematica, we obtain that* $G(100)=20.2759$, $G(10^4)=49.1063$, *and* $G(10^6)=78.0951$. *As we will see, the divergence is slow.*

V. Expressing f-divergences as converging power chi divergence series

In this section, we aim at rewriting the f-divergences as converging infinite series of power chi divergences [24], [53]. The Pearson power chi divergence $D^P_{\chi,k}$ of order k (for any integer $k\in\{2,\dots,\}$) is a dissimilarity obtained for the generator $f^P_{\chi,k}(u)=(u-1)^k$ which generalizes the Pearson χ_2-divergence ($k=2$):

$$\begin{aligned}D^P_{\chi,k}(p:q)&=\int p(x)f^P_{\chi,k}\left(\frac{q(x)}{p(x)}\right)\mathrm{d}\mu(x),\\&=\int p(x)\left(\frac{q(x)}{p(x)}-1\right)^k\mathrm{d}\mu(x),\\&=\int\frac{(q(x)-p(x))^k}{p(x)^{k-1}}\mathrm{d}\mu(x).\end{aligned}$$

We have $D^P_{\chi,2}(p:q)=D^P_{\chi}(p:q):=\int\frac{(p(x)-q(x))^2}{p(x)}\mathrm{d}\mu(x)$. For even integers $k\geq 4$, the Pearson power chi divergence are non-negative dissimilarities since $f^P_{\chi,k}(u)$ is strictly convex (we have ${f^P_{\chi,k}}''(u)=k(k-1)(u-1)^{k-2}\geq 0$). For odd integers $k\geq 3$, the Pearson power chi divergence may be negative. Similarly, we can define the Neyman power chi divergence $D^N_{\chi,k}$ of order k:

$$D^N_{\chi,k}(p:q)=D_{\chi,k}(q:p)=\int\frac{(p(x)-q(x))^k}{q(x)^{k-1}}\mathrm{d}\mu(x).$$

We have $D^N_{\chi,2}(p:q)=D^N_{\chi}(p:q):=\int\frac{(p(x)-q(x))^2}{q(x)}\mathrm{d}\mu(x)$. When k is even it is an f-divergence, otherwise $D^N_{\chi,k}$ may fail the positive-definiteness property of f-divergences. We note $D_{\chi,k}(p:q)=D^P_{\chi,k}(p:q)$ below.

We first state a general framework to obtain power chi divergence expansions of f-divergences.

Theorem 22. *Let X be a topological space and μ be a Borel measure on X with full support. Let $\{p_\theta(x)\}_\theta$ be a family of probability density functions on (X,μ). Assume that for each θ, $p_\theta(x)$ is positive and continuous with respect to x. We also assume that for each θ_1 and θ_2 there exists $C=C(\theta_1,\theta_2)$ such that $p_{\theta_1}(x)\leq Cp_{\theta_2}(x)$ for every $x\in X$. Let $f(z)=\sum_{n=1}^\infty a_n(z-1)^n$ be an analytic function ($f\in C^\omega$), and denote by r_f be the convergence radius of f. Assume that $r_f\geq 1$. Let I_f be the induced f-divergence. Then,*
(i) If $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}<1+r_f$ for every x, then,

$$\begin{aligned}I_f(p_{\theta_1}:p_{\theta_2})&=\sum_{n=2}^\infty a_n\int_X\left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}-1\right)^n p_{\theta_1}(x)\mathrm{d}\mu(x)\\&=\sum_{n=2}^\infty a_nD_{\chi,n}(p_{\theta_1}:p_{\theta_2}).\end{aligned}$$

(ii) If $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}>1+r_f$ for some x, then, the infinite sum $\sum_{n=2}^\infty a_n\int_X\left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}-1\right)^n p_{\theta_1}(x)\mu(dx)$ diverges.

In the above, the sums begin in $n=2$, because the term for $n=1$ vanishes.

Proof. (i) By the assumption and $r_f\geq 1$, $\inf_{x\in X}\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}>1-r_f$. Hence, $\sup_{x\in X}\left|\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}-1\right|<r_f$. Thus we have the Taylor series:

$$f\left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}\right)=\sum_{n=2}^\infty a_n\left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}-1\right)^n,$$

and the convergence is uniform with respect to x. By noting that $p_\theta(x)$ is a probability density function, we have the assertion.

(ii) Since $\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}$ is continuous with respect to x, there exist $\delta_0>0$ and an open set U_0 such that

$$\inf_{x\in U_0}\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)}\geq\delta_0+1+r_f\geq\delta_0+2.$$

Then,

$$a_n \int_{\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \geq 1} \left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} - 1 \right)^n p_{\theta_1}(x)\mu(dx)$$

$$\geq a_n (\delta_0 + r_f)^n \int_{U_0} p_{\theta_1}(x)\mu(dx).$$

Since $r_f \geq 1$,

$$a_n \int_{\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} < 1} \left| \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} - 1 \right|^n p_{\theta_1}(x)\mu(dx)$$

$$\leq a_n \left(1 - \inf_{x \in \mathbb{R}} \frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} \right)^n \to 0, \ n \to \infty.$$

By the assumptions, $\int_{U_0} p_{\theta_1}(x)\mu(dx) > 0$. Thus we see that

$$\lim_{n \to \infty} a_n \int_X \left(\frac{p_{\theta_2}(x)}{p_{\theta_1}(x)} - 1 \right)^n p_{\theta_1}(x)\mu(dx) = +\infty.$$

□

Now we deal with the particular case of Cauchy distributions. We first remark that for every (l_1, s_1) and (l_2, s_2),

$$\max_{x \in \mathbb{R} \cup \{\pm\infty\}} \frac{p_{l_2,s_2}(x)}{p_{l_1,s_1}(x)} = \max_{x \in \mathbb{R} \cup \{\pm\infty\}} \frac{p_{l_1,s_1}(x)}{p_{l_2,s_2}(x)},$$

because there exists $A \in \mathrm{SL}(2, \mathbb{R})$ such that $\theta_1 = A.\theta_2$ and $\theta_2 = A.\theta_1$ where $\theta_j = \ell_j + is_j, \ j \in \{1, 2\}$.

We first deal with the case when the convergence radius is 1. We denote the Kullback-Leibler, α-divergence, Jensen-Shannon, and squared Hellinger divergences by D_{KL}, I_α, D_{JS} and D_H^2, respectively.

Corollary 23. *(i) If* $l^2 + \left(s - \frac{4}{5} \right)^2 < \frac{9}{16}$, *then,* $\sup_{x \in \mathbb{R}} \frac{p_{0,1}(x)}{p_{l,s}(x)} < 2$, *and hence,*

$$\begin{aligned}
D_{\mathrm{KL}}(p_{l,s} : p_{0,1}) &= \sum_{n=2}^{\infty} \frac{(-1)^n}{n} D_{\chi,n}(p_{l,s} : p_{0,1}), \\
I_\alpha(p_{l,s} : p_{0,1}) &= \sum_{n=2}^{\infty} \frac{-4}{1-\alpha^2} \binom{\frac{1+\alpha}{2}}{n} D_{\chi,n}(p_{l,s} : p_{0,1}), \\
D_{\mathrm{JS}}(p_{l,s} : p_{0,1}) &= \sum_{n=2}^{\infty} \frac{(-1)^n (2^{n-1} - 1)}{n(n-1)2^{n-1}} D_{\chi,n}(p_{l,s} : p_{0,1}), \\
D_H^2(p_{l,s} : p_{0,1}) &= \sum_{n=2}^{\infty} \frac{(-1)^n (2n-3)!!}{2^{n-1} n!} D_{\chi,n}(p_{l,s} : p_{0,1}),
\end{aligned}$$

where we used the generalized binomial coefficient [54] for the α*-divergences.*
(ii) If $l^2 + \left(s - \frac{4}{5} \right)^2 > \frac{9}{16}$, *then,* $\sup_{x \in \mathbb{R}} \frac{p_{0,1}(x)}{p_{l,s}(x)} > 2$, *and hence, all of the infinite sums in (i) diverge.*

We now deal with the case when the convergence radius is 2. Let $D_{\mathrm{HM}}(p : q) := \int \frac{2p(x)q(x)}{p(x) + q(x)} \mathrm{d}x$ be the harmonic (mean) divergence [55], [56] (not a regular divergence since $D_{\mathrm{HM}}(p : p) = 1$).

Corollary 24. *(i) If* $l^2 + \left(s - \frac{5}{3} \right)^2 < \frac{16}{9}$, *then,* $\sup_{x \in \mathbb{R}} \frac{p_{0,1}(x)}{p_{l,s}(x)} < 3$ *and hence,*

$$\begin{aligned}
&D_{\mathrm{HM}}(p_{l,s} : p_{0,1}) \\
&= \sum_{n=2}^{\infty} \frac{(-1)^{n+1}}{2^n} \int_{\mathbb{R}} \left(\frac{p_{0,1}(x)}{p_{l,s}(x)} - 1 \right)^n p_{l,s}(x) \mathrm{d}x \\
&= \sum_{n=2}^{\infty} \frac{(-1)^{n+1}}{2^n} D_{\chi,n}(p_{l,s} : p_{0,1}).
\end{aligned}$$

(ii) If $l^2 + \left(s - \frac{5}{3} \right)^2 > \frac{16}{9}$, *then,* $\sup_{x \in \mathbb{R}} \frac{p_{0,1}(x)}{p_{l,s}(x)} > 3$ *and hence, the infinite sum in (i) diverges.*

Other expansions are available in Table 3 of [53] (e.g., Jeffreys' divergence). We refer to Appendix D for implementation of the calculation of f-divergences using these series.

We finally consider the total variation distance between the Cauchy distributions. Then, we *cannot* expect power chi expansions.

Proposition 25. *Let* $f(u) := \frac{|u-1|}{2}$. *Then, for every* $a_1, \cdots, a_n$,

$$\lim_{(l,s) \to (l_0,s_0)} \frac{I_f(p_{l,s} : p_{l_0,s_0}) - \sum_{j=2}^n a_j C_j}{|C_n|} = +\infty,$$

where we let $C_j := \int_{\mathbb{R}} \left(\frac{p_{l,s}(x)}{p_{l_0,s_0}(x)} - 1 \right)^j p_{l_0,s_0}(x) dx$ *for* $2 \leq j \leq n$.

Proof. Let us begin by proving the following lemma:

Lemma 26. *As* $(l, s) \to (l_0, s_0)$,

$$\sup_{x \in \mathbb{R}} \left| \frac{p_{l,s}(x)}{p_{l_0,s_0}(x)} - 1 \right| = O\left(\sqrt{(l - l_0)^2 + (s - s_0)^2} \right).$$

Proof. We see that

$$\frac{p_{l,s}(x)}{p_{l_0,s_0}(x)} - 1$$

$$= \frac{s}{s_0} - 1 + \left(\frac{s}{s_0} - 1 \right) \left(\frac{(x - l_0)^2 + s_0^2}{(x - l)^2 + s^2} - 1 \right) + \frac{(x - l_0)^2 + s_0^2}{(x - l) + s^2} - 1.$$

Since

$$\frac{(x - l_0)^2 + s_0^2}{(x - l)^2 + s^2} - 1 = \frac{2(l - l_0)(x - l) + (l - l_0)^2 + s_0^2 - s^2}{(x - l)^2 + s^2}$$

$$= O\left(\sqrt{(l - l_0)^2 + (s - s_0)^2} \right),$$

we have the assertion. □

By this lemma, we see that

$$\sum_{j=2}^n a_j C_j = O\left((l - l_0)^2 + (s - s_0)^2 \right).$$

On the other hand,

$$I_f(p_{l,s} : p_{l_0,s_0}) = \frac{2}{\pi} \arctan\left(\frac{1}{2} \sqrt{\frac{(l - l_0)^2 + (s - s_0)^2}{s s_0}} \right).$$

Hence,

$$\lim_{(l,s)\to(l_0,s_0)} \frac{I_f(p_{l,s}:p_{l_0,s_0})}{(l-l_0)^2+(s-s_0)^2} = +\infty.$$

Thus we see that

$$\lim_{(l,s)\to(l_0,s_0)} \frac{I_f(p_{l,s}:p_{l_0,s_0}) - \sum_{j=2}^n a_j C_j}{(l-l_0)^2+(s-s_0)^2} = +\infty.$$

By Lemma 26, we see that for $n \geq 2$,

$$C_n = O\left(\left((l-l_0)^2+(s-s_0)^2\right)^{n/2}\right),\ (l,s)\to(l_0,s_0).$$

Thus we have the assertion. □

Remark 27. *Consider the exponential family of exponential distributions* $\{p_\lambda(x) = \lambda\exp(-\lambda x),\ \lambda \in \mathbb{R}_{++}\}$ *defined on the positive half-line support* $\mathcal{X} = \mathbb{R}_{++}$. *The criterion* $\frac{p_{\theta_2}}{p_{\theta_1}} < 1+r_f$ *is satisfied for* $\lambda_1 < \lambda_2 < (1+r_f)\lambda_1$. *Moreover the Pearson order-k power chi divergences are available in closed form for integers $k > 1$ since $\lambda_1 < \lambda_2$ by adapting Lemma 3 of [24] (i.e., when $\lambda_1 < \lambda_2$, it is enough to have conic natural parameter spaces instead of affine spaces). Thus we can calculate the KLD between p_{λ_1} and p_{λ_2} as converging Taylor chi series. In this case, the KLD is also known to be in closed-form as a Bregman divergence for exponential distributions:*

$$D_{\mathrm{KL}}(p_{\lambda_1}:p_{\lambda_2}) = \frac{\lambda_2}{\lambda_1} - \log\frac{\lambda_2}{\lambda_1} - 1.$$

However, if we choose the exponential family of normal distributions, we cannot bound their density ratio; therefore, the Taylor chi series diverge.

Notice that even if the series diverges, the f-divergences may be finite (e.g., when the ratio of densities fails to be bounded by $1+r_f$). In that case, we cannot represent I_f by a Taylor series. By truncating the distributions, we may potentially find a validity range where to apply the Taylor expansion.

VI. Metric spaces induced by the square roots of the Kullback-Leibler and Bhattacharyya divergences

Recall that f-divergences can always be symmetrized by taking the generator $s(u) = f(u)+f^*(u) = f(u)+uf(1/u)$. Metrizing f divergences consists in finding the largest exponent $\alpha > 0$ such that I_s^α is a metric distance satisfying the triangle inequality [57], [58], [59]. For example, the square root of the Jensen-Shannon divergence [60] yields a metric distance which is moreover Hilbertian [61], i.e., meaning that there is an embedding $\phi(\cdot)$ into a Hilbert space $\mathcal{H}$ such that $D_{\mathrm{JS}}(p,q) = \|\phi(p)-\phi(q)\|_{\mathcal{H}}$.

We shall show that the square roots of the Kullback-Leibler divergence and the Bhattacharyya divergence between univariate Cauchy distributions are distances in Theorems 28 and 30 below respectively. In Theorem 33 we also show that the square roots of the Kullback-Leibler divergence are isometrically embeddable into a Hilbert space. In this section we adopt the complex parametrization of the Cauchy distribution ((5)).

Considering the metrization of divergences is important for designing efficient algorithms relying on techniques of computational geometry [62]. A metric distance is symmetric and satisfies the triangle inequality. The triangle inequality can be used to speed up proximity queries. Indeed, the closest point to a query point with respect to a divergence D or its metrization D^α (for some $\alpha > 0$) is the same. Indeed, the nearest neighbor of a point p to a given point set with respect to D or any strictly increasing monotonous function of D does not change: The Voronoi diagrams with respect to D or $m(D)$ coincide for any strictly increasing monotonous function m. For example, Lloyd's batched k-means clustering heuristic [63] uses proximity queries with respect to the Euclidean distance but update the clustering center with respect to the squared Euclidean distance which is not a metric but which yields the center of mass as the centroid. However, metrized divergences satisfy the triangle inequality and this triangle inequality can be used to efficiently prune distance comparison tests. See, for example, the vantage-point tree data structure which is designed for metric spaces [64]. The triangle inequality can also be used to accelerate k-means clustering [65]. Metrizations of symmetrized Bregman divergences have been studied in [61] for clustering algorithms. Proximity queries are also useful in statistics like computing the Chernoff information between a set of Cauchy distributions (multiple hypothesis testing). Indeed, the Chernoff information of a finite set of Cauchy distributions is calculated as the minimum pairwise Chernoff information [66], [67] of these Cauchy distributions. This minimum pairwise Chernoff information can be calculated as nearest neighbor queries between the square root of the symmetric Kullback-Leibler divergence.

A. Metrizations of the Kullback-Leibler and Bhattacharyya divergences

Here we give full details of the proof of Theorem 3 in [14], which lead to an extension of it. See Remark 29 below.

Theorem 28 (Theorem 3 in [14])**.** *The square root of the Kullback-Leibler divergence between the location-scale family of the univariate Cauchy densities is a distance on* $\mathbb{H}$.

Proof. We proceed as in [14] by letting $t(u) := \log\left(\frac{1+\cosh(\sqrt{2}u)}{2}\right), u \geq 0$. Let us consider the properties of $F_2(u) := t(u)^{1/2}/u$. The proof of Theorem 3 in [14] shows that if the function $F_2(u)$ is decreasing, then, the assertion holds. Below we show that $F_2(u)$ is decreasing. We see that

$$F_2'(u) = -2\frac{t(u)^{-1/2}}{u^2}G\left(\frac{u}{\sqrt{2}}\right),$$

where

$$G_2(w) := (2+e^{2w}+e^{-2w})\log\left(\frac{e^w+e^{-w}}{2}\right) - \frac{w}{2}(e^{2w}-e^{-2w}).$$

If we let $x := e^w$, then,

$$G_2(w) = (x+x^{-1})\left((x+x^{-1})\log\left(\frac{x^2+1}{2x}\right) - \frac{x-x^{-1}}{2}\log x\right).$$

Let

$$H_2(x) := x\left((x+x^{-1})\log\left(\frac{x^2+1}{2x}\right) - \frac{x-x^{-1}}{2}\log x\right).$$

Then, $H_2(1)=0$ and

$$H_2'(x) = 4\left(x\log\left(\frac{x^2+1}{2}\right) - \frac{3}{2}x\log x + \frac{x^3}{x^2+1} - \frac{x}{2}\right).$$

Let

$$I_2(x) := x\log\left(\frac{x^2+1}{2}\right) - \frac{3}{2}x\log x + \frac{x^3}{x^2+1} - \frac{x}{2}.$$

Then, $I_2(1)=0$ and

$$I_2'(x) = \log\left(\frac{x^2+1}{2}\right) - \frac{3}{2}\log x + \frac{x^2(3x^2+5)}{(x^2+1)^2} - 2.$$

Let

$$J_2(x) := (x^2+1)^2\log\left(\frac{x^2+1}{2}\right)$$
$$-\frac{3}{2}(x^2+1)^2\log x + x^2(3x^2+5) - 2(x^2+1)^2.$$

Then, $J(1)=0$. If we let $y := x^2$, then,

$$J(x) = (y+1)^2\log\left(\frac{y+1}{2}\right) - \frac{3}{4}(y+1)^2\log y + (y^2+y-2).$$

Let $K_2(y) := J_2(\sqrt{y})$. Then,

$$K_2'(y) = 2(y+1)\left(\log\left(\frac{y+1}{2}\right)+1\right)$$
$$-\frac{3}{2}(y+1)\log y - \frac{3(y+1)^2}{4y} + (2y+1)$$
$$= y + (y+1)\left(2\log(y+1) - \frac{3}{2}\log y + \frac{9}{4} - \frac{3}{4y} - 2\log 2\right).$$

If $y>1$, then, $2\log(y+1) > \frac{3}{2}\log y$ and $\frac{9}{4} - \frac{3}{4y} - 2\log 2 > \frac{3}{2} - 2\log 2 > 0$. Then, $J_2(x) > J_2(1) = 0$ for every $x>1$. Hence, $I_2(x) > I_2(1) = 0$ for every $x>1$. Hence, $G_2(w)>0$ for every $w>0$. Hence, $F_2'(u) < 0$ for every $u>0$. This means that F_2 is strictly decreasing on $[0,\infty)$. Thus we proved that $D_{\mathrm{KL}}(p_{\theta_1}:p_{\theta_2})^{1/2}$ gives a distance on $\mathbb{H}$. □

Remark 29. *By modifying the above proof slightly, we can show that for $\alpha > 1/2$, $D_{\mathrm{KL}}(p_{\theta_1}:p_{\theta_2})^{\alpha}$ is not a distance on $\mathbb{H}$. See [22] for more details.*

Recall the definition of the Bhattacharyya divergence in (7). The term $\int\sqrt{p(x)q(x)}\mathrm{d}x$ is called the Bhattacharyya coefficient. It is easy to see that $D_{\mathrm{Bhat}}(p:q)=0$ iff $p=q$, and $D_{\mathrm{Bhat}}(p:q) = D_{\mathrm{Bhat}}(q:p)$.

Theorem 30. *The square root of the Bhattacharyya divergence between the location-scale family of the univariate Cauchy densities is a distance on $\mathbb{H}$.*

For exponential families, see Proposition 2 in [14] and [68]. The proof of [14, Proposition 2] is *not* applicable to the proof of Theorem 30 above, because it cannot be a Bregman divergence.

Proof. We show the triangle inequality. We follow the idea in the proof of Theorem 3 in [14] dealing with the Kullback-Leibler divergence. We construct the metric transform $t_{\mathrm{FR}\to\mathrm{Bhat}}$, and show that $t_{\mathrm{FR}\to\mathrm{Bhat}}(s)$ is increasing and $\sqrt{t_{\mathrm{FR}\to\mathrm{Bhat}}(s)}/s$ is decreasing. Let ρ_{FR} denote the Fisher-Rao distance. Let $F_3(s) := \cosh(\sqrt{2}s) - 1$. Then, by following the argument in the proof of [14, Theorem 3], $\chi(z,w) = F_3(\rho_{\mathrm{FR}}(z,w))$. Let

$$I_3(z,w) := \int\sqrt{p_z(x)p_w(x)}\,\mathrm{d}x.$$

Then, by the invariance of the f-divergences, $I_3(A.z, A.w) = I_3(z,w)$. Hence we have that for some function J_3, $J_3(\chi(z,w)) = I_3(z,w)$. Hence,

$$\sqrt{D_{\mathrm{Bhat}}(p_{\theta_1}:p_{\theta_2})} = \sqrt{-\log J_3\left(F_3(\rho_{\mathrm{FR}}(\theta_1,\theta_2))\right)}.$$

We have that $t_{\mathrm{FR}\to\mathrm{Bhat}}(s) = -\log J_3(F_3(s))$. It also holds that for every $a\in(0,1)$, $J_3\left(\chi(ai,i)\right)) = I_3(ai,i)$.

By the change-of-variable $x = \tan\theta$ in the integral of $I(ai,i)$, it is easy to see that

$$I_3(ai,i) = \frac{2\sqrt{a}\,\mathbf{K}(1-a^2)}{\pi},$$

where $\mathbf{K}$ is the elliptic integral of the first kind [69]. It is defined by[2]

$$\mathbf{K}(t) := \int_0^{\pi/2}\frac{1}{\sqrt{1-t\sin^2\theta}}d\theta,\ 0\le t<1.$$

Hence,

$$J_3\left(\frac{(1-a)^2}{2a}\right) = \frac{2\sqrt{a}\,\mathbf{K}(1-a^2)}{\pi}.$$

Since

$$F_3(s) = \cosh(\sqrt{2}s) - 1 = \frac{(1-e^{-\sqrt{2}s})^2}{2e^{-\sqrt{2}s}},$$

we have that

$$J_3(F_3(s)) = \frac{2e^{-s/\sqrt{2}}\,\mathbf{K}(1-e^{-2\sqrt{2}s})}{\pi}.$$

Since the above function is decreasing with respect to s, $t_{\mathrm{FR}\to\mathrm{Bhat}}(s)$ is increasing.

Furthermore, we have that

$$\frac{\sqrt{t_{\mathrm{FR}\to\mathrm{Bhat}}(s)}}{s} = \sqrt{-\frac{1}{s^2}\log\left(\frac{2e^{-s/\sqrt{2}}\,\mathbf{K}(1-e^{-2\sqrt{2}s})}{\pi}\right)}. \tag{13}$$

This function is decreasing with respect to s. See Figure 1. We can show this fact by using the results for the complete elliptic integrals. See Section A in Appendix. □

[2]This definition is adopted by WolframAlpha, but is a little different from the usual definition. The usual one is $\mathbf{K}(t) = \int_0^{\pi/2}\frac{1}{\sqrt{1-t^2\sin^2\theta}}d\theta$.

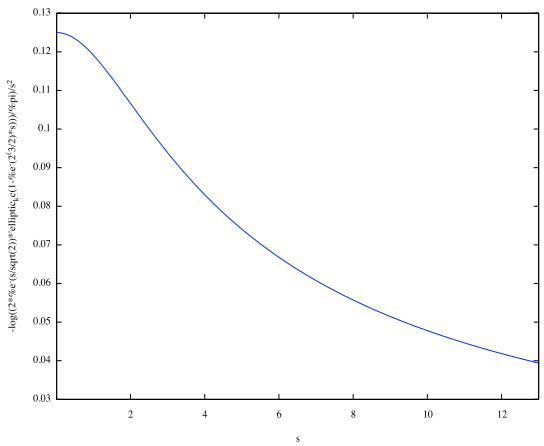

Fig. 1. Graph of $\frac{\sqrt{t_{\mathrm{FR}\to\mathrm{Bhat}}(s)}}{s}$

B. Geometric properties of the metrizations of f-divergences

If a divergence D is given, then we can define an associated Riemannian metric g_D on the parameter space by following Eguchi [70], [71]. Specifically, by regarding D is a smooth function on $M \times M$ where M is the space of parameters, we let

$$(g_D)_r(X_r, Y_r) := -X_p Y_q \, D(p,q)|_{p=q=r}, \ r \in M,$$

where X, Y are vector fields on M.

It is known that if D is the Kullback-Leibler divergence (or a standard f-divergence [20] with $f''(1) = 1$), then, g_D is the Fisher metric. If D is not the Kullback-Leibler divergence, then, we are not sure whether g_D is the Fisher metric. However, g_D is the Fisher metric for every smooth f-divergence between the Cauchy distribution.

Proposition 31. *Let I_f be the f-divergence between the location-scale family of the univariate Cauchy densities. Let F be a function such that*

$$I_f(p_{\theta_1} : p_{\theta_2}) = F(\chi(\theta_1, \theta_2)), \ \theta_1, \theta_2 \in \mathbb{H}.$$

Assume that F is in $C^2([0,\infty))$. Then, the Riemannian metric g_D is $F'(0)\rho$, where ρ is the Poincaré metric on $\mathbb{H}$.

We remark that $\sqrt{2}\rho_{\mathrm{FR}}$ is identical with the Poincaré distance on $\mathbb{H}$.

Proposition 32. *The square roots of the Kullback-Leibler and Bhattacharyya divergences between the location-scale family of the univariate Cauchy densities are both complete distances on $\mathbb{H}$.*

Proof. Assume that $(z_n)_n$ is a Cauchy sequence with respect to $\sqrt{D_{\mathrm{KL}}}$. Since $\sqrt{D_{\mathrm{KL}}(p_z : p_w)}$ is increasing as a function of $\chi(z,w)$, we see that $\chi(z_n, z_m) \to 0, n, m \to \infty$. We see that $\chi(z,w) \leq \delta$ if and only if

$$|w - (\mathrm{Re}(z) + i(1+\delta)\mathrm{Im}(z))| \leq \sqrt{\delta(\delta+2)}\mathrm{Im}(z).$$

Hence $(z_n)_n$ is bounded. Let z be an accumulation point of $(z_n)_n$. Then, $z_{k_n} \to z, \ n \to \infty$ with respect to the Euclidean distance. Hence, $\chi(z_{k_n}, z) \to 0, n \to \infty$. Hence, $\sqrt{D_{\mathrm{KL}}(p_{z_{k_n}} : p_z)} \to 0, n \to \infty$. Since $(z_n)_n$ is a Cauchy sequence with respect to $\sqrt{D_{\mathrm{KL}}}$, we see that $\sqrt{D_{\mathrm{KL}}(p_{z_n} : p_z)} \to 0, n \to \infty$. □

We finally consider an isometric embedding of the kernel defined by the square root of the Kullback-Leibler divergence into a Hilbert space. The following is a significant extension of Theorem 4 in [14].

Theorem 33. *The square root of the Kullback-Leibler divergence between the location-scale family of the univariate Cauchy densities is isometrically embeddable into a Hilbert space.*

The method of the proof is completely different from the proof of Theorem 4 in [14], because it cannot be a Bregman divergence. See also [61]. Furthermore, Theorem 1 in [60], which gives a criteria whether a kernel is isometrically embeddable into a Hilbert space, will not be applicable to this setting. We give an elementary long proof of this assertion in Appendix B in along with the following proof.

Proof. By [72], it suffices to show that $D_{\mathrm{KL}}(p_z : p_w)$ is a conditionally negative definite kernel on $\mathbb{H}$, that is, for every $(c_1, \cdots, c_n)$ such that $\sum_{i=1}^n c_i = 0$ and every $z_1, \cdots, z_n \in \Theta$,

$$\sum_{i,j=1}^n c_i c_j D_{\mathrm{KL}}(p_{z_i} : p_{z_j}) \leq 0. \tag{14}$$

We consider transformations of the parameter spaces from $\mathbb{H}$ to the hyperbolic plane in $\mathbb{R}^3$. Let

$$\mathbb{L} := \{(x,y,z) \in \mathbb{R}^3 : z > 0, x^2 + y^2 - z^2 = -1\}.$$

Let

$$d_{\mathbb{L}}\left((x_1,y_1,z_1),(x_2,y_2,z_2)\right)$$
$$:= \cosh^{-1}\left(z_1 z_2 - x_1 x_2 - y_1 y_2\right), (x_1,y_1,z_1),(x_2,y_2,z_2) \in \mathbb{L}.$$

Let $\phi_1 : \mathbb{L} \to \mathbb{D}$ be the map defined by $\phi_1(x,y,z) := \left(\frac{x}{1+z}, \frac{y}{1+z}\right)$. Let $\phi_2 : \mathbb{D} \to \mathbb{H}$ be the map defined by $\phi_2(x,y) := \left(-\frac{2y}{(1-x)^2+y^2}, \frac{1-x^2-y^2}{(1-x)^2+y^2}\right)$. Then, ϕ_1 and ϕ_2 are both bijective. Hence $\phi_2 \circ \phi_1$ is a bijection between $\mathbb{H}$ and $\mathbb{L}$.

Hence, in order to show (14), it suffices to show that for $(x_1,y_1,z_1), \cdots, (x_n,y_n,z_n) \in \mathbb{L}$,

$$\sum_{i,j=1}^n c_i c_j \log\left(1 + \frac{\chi\left(\phi_2(\phi_1(x_i,y_i,z_i)), \phi_2(\phi_1(x_j,y_j,z_j))\right)}{2}\right) \leq 0.$$

Since

$$\chi(\phi_2(w_1), \phi_2(w_2)) = \frac{2|w_1 - w_2|^2}{(1-|w_1|^2)(1-|w_2|^2)}, \quad w_1, w_2 \in \mathbb{D},$$

we see that for $(x_1,y_1,z_1),(x_2,y_2,z_2) \in \mathbb{L}$,

$$\chi\left(\phi_2(\phi_1(x_1,y_1,z_1)), \phi_2(\phi_1(x_2,y_2,z_2))\right)$$
$$= \cosh\left(d_{\mathbb{L}}\left((x_1,y_1,z_1),(x_2,y_2,z_2)\right)\right) - 1.$$

Hence, in order to show (14), it suffices to show that for $(x_1,y_1,z_1), \cdots, (x_n,y_n,z_n) \in \mathbb{L}$,

$$\sum_{i,j=1}^n c_i c_j \log\left(\frac{1 + \cosh\left(d_{\mathbb{L}}\left((x_i,y_i,z_i),(x_j,y_j,z_j)\right)\right)}{2}\right) \leq 0.$$

Since $2(\cosh(x/2))^2 = 1 + \cosh(x), x \in \mathbb{R}$, in order to show (14), it suffices to show that for $(x_1, y_1, z_1), \cdots, (x_n, y_n, z_n) \in \mathbb{L}$,

$$\sum_{i,j=1}^{n} c_i c_j 2 \log\left(\cosh\left(\frac{d_{\mathbb{L}}\left((x_i, y_i, z_i), (x_j, y_j, z_j)\right)}{2}\right)\right) \leq 0. \tag{15}$$

The last inequality follows from Theorem 7.5 in Faraut-Harzallah [73].

Theorem 34 (Theorem 7.5 in Faraut-Harzallah [73])**.** *Let $d \geq 2$. For $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$ and $y = (y_1, \cdots, y_d) \in \mathbb{R}^d$, let*

$$r(x, y) := \cosh^{-1}\left(\sqrt{\left(1 + \sum_{i=1}^{d} x_i^2\right)\left(1 + \sum_{i=1}^{d} y_i^2\right)} - \sum_{i=1}^{d} x_i y_i\right)$$

and

$$Q(x, y) := \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})} \times \int_0^{\pi} \log\left(\cosh(r(x, y)) - \cos\theta \sinh(r(x, y))\right) \sin^{d-2}\theta d\theta.$$

Then, Q is a conditionally negative definite kernel on $\mathbb{R}^d$, specifically, for every $(c_1, \cdots, c_n)$ such that $\sum_{i=1}^{n} c_i = 0$ and every $z_1, \cdots, z_n \in \mathbb{R}^d$,

$$\sum_{i,j=1}^{n} c_i c_j Q(z_i, z_j) \leq 0.$$

We apply this theorem to the case that $d = 2$. Then, we see that for $x = (x_1, x_2)$ and $y = (y_1, y_2)$,

$$r(x, y) = d_{\mathbb{L}}\left(\left(x_1, x_2, \sqrt{1 + x_1^2 + x_2^2}\right), \left(y_1, y_2, \sqrt{1 + y_1^2 + y_2^2}\right)\right),$$

and

$$Q(x, y) = \frac{1}{\pi}\int_0^{\pi} \log\left(\cosh(r(x, y)) - \cos\theta \sinh(r(x, y))\right) d\theta.$$

The formula no.4.224.9 in [74] states that

$$2 \log\cosh\left(\frac{r}{2}\right) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left(\cosh(r) + \cos\theta \sinh(r)\right) d\theta$$
$$= \frac{1}{\pi}\int_0^{\pi} \log\left(\cosh(r) + \cos\theta \sinh(r)\right) d\theta, r \geq 0.$$

Hence,

$$Q(x, y) = 2 \log\cosh\left(\frac{r(x, y)}{2}\right).$$

Thus (15) holds. □

Remark 35. *(i) Faraut-Harzallah [73] often uses terminologies of representation theory. The proof of Theorem 7.5 in Faraut-Harzallah [73] heavily depends on Takahashi's long paper [75] in representation theory. Another paper of Faraut-Harzallah [76] gave another derivation of Theorem 7.5 in [73]. However, [76] heavily depends on Helgason's long paper [77] in representation theory. By following the outline of [76], we give an elementary, long proof of Theorem 33 in Appendix B without using any terminologies of representation theory.*
(ii) Theorem 8.1 in [73] proved some results on positive and negative definite kernels defined on a real infinite dimensional hyperbolic space. It is similar to Theorem 7.5 in [73], but is different and we cannot apply the result to our case. See Chapter5, Section 5 in Berg, Christensen and Ressel [78] for details. [73] gives some interesting examples of conditionally negative definite kernels on Euclidian spaces of every dimension.

Remark 36. *We can show that neither $(\mathbb{H}, \sqrt{D_{\text{Bhat}}})$ nor $(\mathbb{H}, \sqrt{D_{\text{KL}}})$ is a geodesic metric space. Now by Hopf-Rinow's theorem (see [79, Theorem 16]) and Proposition 32, neither $\sqrt{D_{\text{Bhat}}}$ or $\sqrt{D_{\text{KL}}}$ is a Riemannian distance. Furthermore, we can also show that neither $(\mathbb{H}, \sqrt{D_{\text{Bhat}}})$ nor $(\mathbb{H}, \sqrt{D_{\text{KL}}})$ is Gromov-hyperbolic. Now we see that both of the metrics $\sqrt{D_{\text{Bhat}}}$ and $\sqrt{D_{\text{KL}}}$ are locally similar to the Poincaré metric (recall Proposition 31), however, in global, they are completely different from the Poincaré metric. See [22] for details.*

VII. Conclusion and Discussion

In this work, we first proved that all f-divergences [18] between univariate Cauchy distributions are symmetric by handling the parameters of Cauchy distributions as complex numbers and considering the action of the Möbius group [80]. We showed that those f-divergences can be calculated as strictly increasing scalar functions of the chi-square divergence by using the notion of maximal invariant [34], [1]. However, we showed that the f-divergences between multivariate Cauchy densities are in general asymmetric by reporting an example, but on the other hand, proved that f-divergences between multivariate Cauchy divergences are symmetric when the Cauchy scale matrices coincide. We also reported a criterion for expanding f-divergences as converging series of power chi divergences, and illustrated the technique for some common f-divergences between Cauchy distributions. Finally, we proved that the square roots of the Kullback-Leibler and the Bhattacharyya divergences between univariate Cauchy distributions yield both complete metric spaces, and the square root of the Kullback-Leibler divergence between univariate Cauchy distributions is isometrically embeddable into a Hilbert space.

When divergences are symmetric, the induced divergence-based information α-geometry [20] all coincide with the Fisher-Rao geometry (i.e., the 0-geometry) (equivalently, the Amari-Chentsov symmetric cubic tensor vanishes). When parametric families belong to exponential families, the only families which yield symmetric KL divergences are provably location normal distributions. Indeed, the KLD between two densities of an exponential family amounts to an equivalent Bregman divergence and the only symmetric Bregman divergences are squared Mahalanobis divergences [81]. In this paper, we proved that all f-divergences between Cauchy densities are symmetric. Once such a family is found, we may consider diffeomorphisms $y = m(x)$ which keeps the f-divergences invariant but yield other parametric families of distributions (e.g., circular, wrapped, or log-Cauchy families).

Notice that the Cauchy densities are infinitely divisible distributions and can thus be realized as scale mixtures of Gaussian density functions (see https://betanalpha.github.io/assets/case_studies/fitting_the_cauchy.html).

To conclude, let us state an interesting open problem raised by our work:

Problem 37. *Characterize all parametric distribution families $\{p_\theta\}_\theta$ (and transformational models [82]) for which the Kullback-Leibler divergence (or more generally any f-divergence) is symmetric.*

Appendix A
Complete elliptic integrals

This section is devoted to the details of the proof of (13) in the proof of Theorem 30. Let $\mathbf{E}$ be the complete elliptic integral of the second kind. We let[3]

$$\mathbf{E}(t) := \int_0^{\pi/2} \sqrt{1 - t\sin^2\theta} d\theta.$$

Proof. Let

$$F_4(u) := \frac{-\log\left(2e^{-u/4}\mathbf{K}(1-e^{-u})/\pi\right)}{u^2}.$$

We consider the derivative.

$$F_4'(u) = \frac{-1}{u^2}\left(\frac{1}{4} + e^{-u}\frac{\mathbf{K}'(1-e^{-u})}{\mathbf{K}(1-e^{-u})} - \frac{2}{u}\log\left(\frac{2}{\pi}\mathbf{K}(1-e^{-u})\right)\right).$$

Now it suffices to show that for every $u > 0$,

$$\frac{1}{4} + e^{-u}\frac{\mathbf{K}'(1-e^{-u})}{\mathbf{K}(1-e^{-u})} - \frac{2}{u}\log\left(\frac{2}{\pi}\mathbf{K}(1-e^{-u})\right) > 0.$$

Let $x := 1 - e^{-u}$. Then, it suffices to show that for every $x \in (0,1)$,

$$\frac{1}{4} + (1-x)\frac{\mathbf{K}'(x)}{\mathbf{K}(x)} + \frac{2}{\log(1-x)}\log\left(\frac{2}{\pi}\mathbf{K}(x)\right) > 0.$$

Let

$$G_4(x) := \log\left(\frac{2}{\pi}\mathbf{K}(x)\right) + (\log(1-x))\left(\frac{1}{8} + \frac{1-x}{2}\frac{\mathbf{K}'(x)}{\mathbf{K}(x)}\right).$$

It suffices to show that $G_4(x) < 0$ for every $x \in (0,1)$.

We see that $G(0) = 0$. Hence it suffices to show that $G'(x) < 0$ for every $x \in (0,1)$. By Lemma 42 below,

$$G_4(x) = \log\left(\frac{2}{\pi}\mathbf{K}(x)\right) + (\log(1-x))\left(\frac{3}{8} + \frac{1}{4x}\left(\frac{\mathbf{E}(x)}{\mathbf{K}(x)} - 1\right)\right).$$

By Lemmas 42 and 43 below,

$$G_4'(x) = -\frac{H_4(x)}{8x^2(1-x)},$$

where we let

$$H_4(x) := (x(2-x) + (x-1)\log(1-x))\mathbf{K}(x)^2 - 2x\mathbf{K}(x)\mathbf{E}(x) + \log(1-x)\mathbf{E}(x)^2.$$

Then it suffices to show that $H_4(x) > 0$ for every $x \in (0,1)$. Since $-2x < 0$ and $\log(1-x) < 0$, by noting Lemma 44 below, it holds that

$$\frac{\mathbf{H}(x)}{\mathbf{K}(x)^2} \geq (x(2-x) + (x-1)\log(1-x)) - 2xI_4(x) + \log(1-x)I_4(x)^2,$$

where we let

$$I_4(x) := \frac{1}{2} - \frac{x}{4} + \frac{\sqrt{1-x}}{2}.$$

Our main idea is to use different estimates for $H(x)/\mathbf{K}(x)^2$ on a neighborhood of 1 and on the complement of it.

Lemma 38. *For $x \leq 0.998$,*

$$(x(2-x) + (x-1)\log(1-x)) > 2xI_4(x) + \log(1-x)I_4(x)^2.$$

Proof. Let $y := \sqrt{1-x}$. Then,

$$(x(2-x) + (x-1)\log(1-x)) > 2xI(x) + \log(1-x)I(x)^2$$

is equivalent with

$$\log y > 4\frac{y^2-1}{y^2+6y+1}.$$

Let

$$P_4(y) := \log y - 4\frac{y^2-1}{y^2+6y+1}.$$

Then, $P_4(1) = 0$. By considering the derivative of P_4, it is increasing $y < 5 - 2\sqrt{6}$ and decreasing $y > 5 - 2\sqrt{6}$.

We see that $P_4(y) > 0 \quad y > 0.041$. Now the assertion follows from the fact that $0.998 < 1 - (0.041)^2$. □

Now it suffices to show that $H_4(x) > 0$ for $x > 0.998$.

Lemma 39.

$$x(2-x) + (x-1)\log(1-x) \geq 1, \quad x \in (0.998, 1).$$

Proof. Let $g_4(x) := x(2-x) + (x-1)\log(1-x)$. Then, $g(1) = 1$ and $g_4'(x) = 3 - 2x + \log(1-x)$. This is negative if $x > 0.9$. □

Lemma 40.

$$2x\frac{\mathbf{E}(x)}{\mathbf{K}(x)} < \frac{1}{2}, \quad x \in (0.998, 1).$$

Proof. We see that

$$\frac{d}{dx}\left(x\frac{\mathbf{E}(x)}{\mathbf{K}(x)}\right) \leq 2\frac{\mathbf{E}(x)}{\mathbf{K}(x)} - \frac{1}{2}.$$

By Lemma 43 below and the fact that

$$\frac{\mathbf{E}(0.995)}{\mathbf{K}(0.995)} < \frac{1}{4},$$

we see that

$$2\frac{\mathbf{E}(x)}{\mathbf{K}(x)} \leq \frac{1}{2}, \quad x > 0.995.$$

Hence,

$$2x\frac{\mathbf{E}(x)}{\mathbf{K}(x)} < 2\frac{\mathbf{E}(0.995)}{\mathbf{K}(0.995)} < \frac{1}{2}.$$

[3] This is also a little different from the usual definition. The usual one is $\mathbf{E}(t) = \int_0^{\pi/2}\sqrt{1-t^2\sin^2\theta}d\theta$.

□

Lemma 41.

$$-\log(1-x)\left(\frac{\mathbf{E}(x)}{\mathbf{K}(x)}\right)^2 < \frac{1}{2}, x \in (0.998, 1).$$

Proof. We use Lemma 45 below. It suffices to show that

$$\frac{2x^{1/2}}{\log(1+x^{1/2})-\log(1-x^{1/2})} \leq \sqrt{\frac{1}{-2\log(1-x)}}$$

for $x \in (0.998, 1)$. This is equivalent with

$$h_4(x) := \left(\log(1+x^{1/2})-\log(1-x^{1/2})\right)^2 + 8x\log(1-x) \geq 0$$

for $x \in (0.998, 1)$. We see that

$$-\frac{h_4'(x)}{2} =$$

$$\frac{\log(1-\sqrt{x})-\log(1+\sqrt{x})+2\sqrt{x}(x+(x-1)\log(1-x))}{(1-x)\sqrt{x}}.$$

It is easy to see that

$$\log(1-\sqrt{x})-\log(1+\sqrt{x})+2\sqrt{x}(x+(x-1)\log(1-x)) < 0$$

for $x \in (0.998, 1)$. Hence h_4 is increasing at least on $(0.998, 1)$. Now use the fact that $h_4(0.998) > 0$. □

By Lemmas 39, 40 and 41, we see that $H_4(x) > 0$ for $x > 0.998$. The proof of (13) is completed. □

1) Some Lemmas concerning the complete elliptic integrals: We collect standard results about the complete elliptic integrals.

Lemma 42.

$$\mathbf{K}'(x) = -\frac{\mathbf{K}(x)}{2x} + \frac{\mathbf{E}(x)}{2x(1-x)}.$$

Lemma 43.

$$\frac{d}{dx}\left(\frac{\mathbf{E}(x)}{\mathbf{K}(x)}\right) = -\frac{1}{2x} + \frac{1}{x}\frac{\mathbf{E}(x)}{\mathbf{K}(x)} - \frac{1}{2x(1-x)}\left(\frac{\mathbf{E}(x)}{\mathbf{K}(x)}\right)^2 \leq 0.$$

In particular, $\mathbf{E}/\mathbf{K}$ *is strictly decreasing.*

Lemma 44.

$$\frac{\mathbf{E}(x)}{\mathbf{K}(x)} \leq \frac{1}{2} - \frac{x}{4} + \frac{\sqrt{1-x}}{2}, \ x \in [0, 1).$$

The following is due to Anderson, Vamanamurthy, and Vuorinen [83].

Lemma 45 ([83, Theorem 3.6])**.**

$$\frac{\mathbf{E}(x)}{\mathbf{K}(x)} \leq \frac{2x^{1/2}}{\log(1+x^{1/2})-\log(1-x^{1/2})}, \ x \in [0, 1).$$

Appendix B
Negative definiteness of the KLD between Cauchy densities

In this section, we give an elementary but long proof of Theorem 33. We first give an outline of the alternative proof. Our proof follows the strategy of [76] and consists of three steps. Contrarily to the proof of Theorem 33 given in Section VI, we do not need to introduce the hyperboloid space $\mathbb{L}$.

Outline of Proof of Theorem 33. By [72], it suffices to show that $D_{\mathrm{KL}}(p_z : p_w)$ is a conditionally negative definite kernel on $\mathbb{H}$, that is, for every $(c_1, \cdots, c_n)$ such that $\sum_{i=1}^n c_i = 0$ and every $z_1, \cdots, z_n \in \Theta$,

$$\sum_{i,j=1}^n c_i c_j D_{\mathrm{KL}}(p_{z_i} : p_{z_j}) = \sum_{i,j=1}^n c_i c_j \log\left(1 + \frac{\chi(z_i, z_j)}{2}\right) \leq 0.$$

If we find a positive-definite kernel H_s on $\mathbb{H}$ for each $s > 0$ such that

$$\log\left(1 + \frac{\chi(z,w)}{2}\right) = \lim_{s \to +0} \frac{1 - H_s(z,w)}{s}, \quad z, w \in \mathbb{H},$$

then, for each $s > 0$, $1 - H_s(z,w)$ is a conditionally negative definite kernel on $\mathbb{H}$ and hence $\log\left(1 + \frac{\chi(z,w)}{2}\right)$ is also a conditionally negative definite kernel on $\mathbb{H}$.

Step 1. Let d be the Poincaré distance on $\mathbb{H}$, which is equal to $\sqrt{2}\rho_{\mathrm{FR}}$. Then, $\cosh(d(z,w)) = 1 + \chi(z,w)$ and

$$2\log\cosh\left(\frac{d(z,w)}{2}\right) = \log\left(1 + \frac{\chi(z,w)}{2}\right).$$

We see that for every $r \geq 0$,

$$2\log\cosh\left(\frac{r}{2}\right)$$

$$= \lim_{s \to +0} \frac{1}{s}\left(1 - \frac{1}{2\pi}\int_{-\pi}^{\pi} (\cosh(r) + \cos\theta\sinh(r))^{-s}\, d\theta\right).$$

Hence it suffices to show that

$$H_s(z,w) :=$$

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} (\cosh(d(z,w)) + \cos\theta\sinh(d(z,w)))^{-s}\, d\theta, z, w \in \mathbb{H},$$

is positive definite for every $s \in (0, 1)$.

Step 2. Let

$$P(z,x) := \frac{\mathrm{Im}(z)}{|x-z|^2}(x^2+1), \ z \in \mathbb{H}, x \in \mathbb{R},$$

and,

$$\mu(dx) := \frac{dx}{\pi(x^2+1)}, \ x \in \mathbb{R}.$$

Then we see that

$$H_s(z,w) = \int_{\mathbb{R}} P(z,x)^s P(w,x)^{1-s}\mu(dx) =$$

$$C_s \int_{\mathbb{R}^2} (P(w,x)P(z,y))^{1-s}\left(\frac{(x-y)^2}{(x^2+1)(y^2+1)}\right)^{-s}\mu(dx)\mu(dy),$$

where C_s is a positive constant depending only on s.

Step 3. Let $z_1, \cdots, z_n \in \mathbb{H}$ and $c_1, \cdots, c_n \in \mathbb{R}$ with $\sum_{i=1}^n c_i = 0$. Let

$$\varphi_s(x) := \sum_{i=1}^n c_i P(z_i, x)^{1-s}, \ x \in \mathbb{R}, \tag{16}$$

which is continuous on $\mathbb{R}$. Let

$$k_s(x,y) := \left(\frac{(x-y)^2}{(x^2+1)(y^2+1)} \right)^{-s}, \tag{17}$$

which is a positive definite kernel on $\mathbb{R}$.

Thus we see that

$$\sum_{i.j=1}^n c_i c_j H_s(z_i, z_j) = \frac{C_s}{\pi^2} \iint_{\mathbb{R}^2} \frac{\varphi_s(x)\varphi_s(y) k_s(x,y)}{(x^2+1)(y^2+1)} dxdy \geq 0.$$

In order to show the last inequality, we use a certain approximation for k_s. □

Now we proceed to the full proof.

Proof of Theorem 33. Step 1. It is known that (see formula no.4.224.9 in [74])

$$2 \log \cosh \left(\frac{r}{2} \right)$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \log \left(\cosh(r) + \cos\theta \sinh(r) \right) d\theta, \ r \geq 0.$$

We see that for $r \geq 0$,

$$|\log(\cosh(r) + \cos\theta \sinh(r))| \leq r.$$

Since for $t > 0$, $\lim_{s \to +0} \frac{1-t^{-s}}{s} = \log t$ and $\left| \frac{1-t^{-s}}{s} \right| \leq |\log t|$,

$$\int_{-\pi}^{\pi} \log \left(\cosh(r) + \cos\theta \sinh(r) \right) d\theta$$

$$= \lim_{s \to +0} \int_{-\pi}^{\pi} \frac{1 - (\cosh(r) + \cos\theta \sinh(r))^{-s}}{s} d\theta, r > 0,$$

by the Lebesgue convergence theorem. This convergence also holds for $r = 0$. This completes Step 1.

Step 2. This part is the longest.

Lemma 46.

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} (\cosh(r) + \cos\theta \sinh(r))^{-s} d\theta = \int_{\mathbb{R}} P(e^r i, x)^s \mu(dx).$$

Proof. Let $x = \tan \frac{\theta}{2}$. Then, $d\theta = \frac{2}{1+x^2} dx$ and

$$\cosh(r) + \cos\theta \sinh(r) = \frac{e^{2r} + x^2}{e^r(1+x^2)} = \frac{1}{P(e^r i, x)}.$$

□

Lemma 47. *For $A \in \mathrm{SO}(2)$ and $z \in \mathbb{H}$,*

$$\int_{\mathbb{R}} P(A.z, x)^s \mu(dx) = \int_{\mathbb{R}} P(z, x)^s \mu(dx).$$

Proof. Let $A = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$. Let $y \in \mathbb{R}$ such that $x = A.y$. Then, $P(A.z, A.y) = P(z, y)$ and

$$\mu(dx) = \frac{1}{\pi} \frac{1}{(A.y)^2 + 1} \frac{dx}{dy} dy = \frac{1}{\pi} \frac{1}{y^2+1} dy = \mu(dy).$$

□

Now we introduce a group structure on $\mathbb{H}$. For $z = z_1 + iz_2$ and $w = w_1 + iw_2$, let $zw := (z_1 + z_2 w_1) + iz_2 w_2$. This gives a group structure on $\mathbb{H}$. It holds that

$$w^{-1} = \frac{-w_1 + i}{w_2}, \ w = w_1 + iw_2$$

and the unit element is the imaginary unit i.

We see that $\chi(w^{-1}z, i) = \chi(z, w), z, w \in \mathbb{H}$ and hence

$$d(w^{-1}z, i) = d(z, w), z, w \in \mathbb{H}. \tag{18}$$

Lemma 48. *For $z, w \in \mathbb{H}$,*

$$H_s(z, w) = \int_{\mathbb{R}} P(w^{-1}z, x)^s \mu(dx).$$

Proof. By (18), we can assume that $w = i$. Then there exists $A \in SO(2)$ such that $e^{d(z,i)} i = A.z$. Now the assertion follows from Lemmas 46 and 47. □

For $w = w_1 + iw_2 \in \mathbb{H}$ and $x \in \mathbb{R}$, we let $wx := w_2 x + w_1$.

Lemma 49.

$$P(w^{-1}z, x) P(w, wx) = P(z, wx), \quad z, w \in \mathbb{H}, x \in \mathbb{R}.$$

Proof. Since

$$w^{-1}z = \frac{z_1 - w_1 + iz_2}{w_2}, \quad z = z_1 + iz_2, w = w_1 + iw_2,$$

we see that

$$P(w^{-1}z, x) = \frac{z_2 w_2}{(z_1 - w_1 - w_2 x)^2 + z_2^2} (x^2 + 1).$$

We also see that

$$P(z, wx) = \frac{z_2((w_2 x + w_1)^2 + 1)}{(z_1 - w_1 - w_2 x)^2 + z_2^2}$$

and

$$P(w, wx) = \frac{(w_2 x + w_1)^2 + 1}{w_2(x^2+1)}.$$

The assertion follows from these identities. □

Proposition 50.

$$H_s(z, w) = \int_{\mathbb{R}} P(z, x)^s P(w, x)^{1-s} \mu(dx), \quad z, w \in \mathbb{H}.$$

Proof. By Lemmas 48 and 49,

$$H_s(z, w) = \int_{\mathbb{R}} P(z, wx)^s P(w, wx)^{-s} \mu(dx).$$

Let $y = wx = w_2 x + w_1$. Then, $\mu(dx) = \frac{w_2}{\pi |y - w|^2} dy$. Hence,

$$\int_{\mathbb{R}} P(z, wx)^s P(w, wx)^{-s} \mu(dx) = \int_{\mathbb{R}} P(z, y)^s P(w, y)^{1-s} \mu(dy).$$

□

Lemma 51. *For every $s \in (0,1/2)$, there exists a positive constant C_s such that for every $a \in \mathbb{R}$*

$$(1+a^2)^{-s} = \frac{C_s}{\pi}\int_{\mathbb{R}} \frac{|x+a|^{-2s}}{(1+x^2)^{1-s}}dx.$$

Proof. Let $x = \tan\theta, |\theta| < \pi/2$. Then, $d\theta = \cos^2\theta dx = \frac{1}{1+x^2}dx$ and $\frac{(x+a)^2}{1+x^2} = (\sin\theta + a\cos\theta)^2$. Hence,

$$\int_{\mathbb{R}} \frac{|x|^{-2s}}{(1+(x-a)^2)^{1-s}}dx = \int_{-\pi/2}^{\pi/2} |\sin\theta + a\cos\theta|^{-2s}\, d\theta.$$

By symmetry,

$$\int_{-\pi/2}^{\pi/2} |\sin\theta + a\cos\theta|^{-2s}\, d\theta = \frac{1}{2}\int_{-\pi}^{\pi} |\sin\theta + a\cos\theta|^{-2s}\, d\theta$$

$$= \pi(1+a^2)^{-s}\int_{-\pi}^{\pi} |\cos\theta|^{-2s}\, d\theta.$$

The assertion holds if we let $C_s := \left(\int_{-\pi}^{\pi} |\cos\theta|^{-2s}\, d\theta\right)^{-1}$. □

The following is a crucial part of the proof.

Lemma 52 (intertwining formula)**.** *For every $s \in (0,1/2)$, $w \in \mathbb{H}$ and $y \in \mathbb{R}$,*

$$P(w,y)^s = C_s \int_{\mathbb{R}} P(w,x)^{1-s}\left(\frac{(x-y)^2}{(x^2+1)(y^2+1)}\right)^{-s}\mu(dx). \tag{19}$$

Proof. Let $\xi := w - y$ and $t := x - y$. Then, (19) holds if and only if

$$\left(\frac{\mathrm{Im}(\xi)}{|\xi|^2}\right)^s = \frac{C_s}{\pi}\int_{\mathbb{R}} \left(\frac{\mathrm{Im}(\xi)}{|\xi - t|^2}\right)^{1-s} |t|^{-2s}dt. \tag{20}$$

Let $u := (t - \mathrm{Re}(\xi))/\mathrm{Im}(\xi)$. Then,

$$\int_{\mathbb{R}} \left(\frac{\mathrm{Im}(\xi)}{|\xi - t|^2}\right)^{1-s} |t|^{-2s}dt$$

$$= (\mathrm{Im}(\xi))^{-s}\int_{\mathbb{R}} \left(\frac{1}{1+u^2}\right)^{1-s}\left|u + \frac{\mathrm{Re}(\xi)}{\mathrm{Im}(\xi)}\right|^{-2s} du.$$

Hence (20) holds if and only if

$$\left(\left(\frac{\mathrm{Re}(\xi)}{\mathrm{Im}(\xi)}\right)^2 + 1\right)^{-s}$$

$$= \frac{C_s}{\pi}\int_{\mathbb{R}} \left(\frac{1}{1+u^2}\right)^{1-s}\left|u + \frac{\mathrm{Re}(\xi)}{\mathrm{Im}(\xi)}\right|^{-2s} du,$$

which follows from Lemma 51. □

By Proposition 50 and Lemma 52,

Proposition 53. *For every $s \in (0,1/2)$ and $z, w \in \mathbb{H}$,*

$$H_s(z,w) =$$

$$C_s \int_{\mathbb{R}^2} (P(w,x)P(z,y))^{1-s}\left(\frac{(x-y)^2}{(x^2+1)(y^2+1)}\right)^{-s}\mu(dx)\mu(dy).$$

This completes Step 2. Figure 2 below describes the dependencies between the assertions in Step 2.

Step 3. We first recall the definition of k_s given in (17).

Lemma 54. *$k_s(x,y)$ is a positive definite kernel on $\mathbb{R}$.*

Proof. For $r \in (0,1)$, let

$$k_s^{(r)}(x,y) := \left(1 - r\frac{(xy+1)^2}{(x^2+1)(y^2+1)}\right)^{-s}.$$

Since $(x,y) \mapsto \frac{1}{(x^2+1)(y^2+1)}$ and $(x,y) \mapsto (xy)^2 + 2xy + 1$ are both positive definite kernels on $\mathbb{R}$, $(x,y) \mapsto \frac{(xy+1)^2}{(x^2+1)(y^2+1)}$ is also a positive definite kernel on $\mathbb{R}$. By the Taylor expansion, $(1-x)^{-s} = \sum_{n=0}^{\infty} a_n x^n,\ |x| < 1$, for $a_n \geq 0, n = 0, 1, \cdots$. Hence $k_s^{(r)}(x,y)$ is a positive definite kernel on $\mathbb{R}$. Since

$$\lim_{r\to 1-0} k_s^{(r)}(x,y) = k_s(x,y),$$

$k_s(x,y)$ is also a positive definite kernel on $\mathbb{R}$. □

We second recall the definition of φ_s given in (16). By this and the quadrature rule for the Riemannian integral for continuous functions, it holds that for every $a < b$ and $r \in (0,1)$,

$$\iint_{[a,b]^2} \frac{\varphi_s(x)\varphi_s(y)k_s^{(r)}(x,y)}{(x^2+1)(y^2+1)}dxdy \geq 0.$$

Since $0 \leq k_s^{(r)}(x,y) \leq k_s(x,y)$,

$$\iint_{\mathbb{R}^2} \frac{|\varphi_s(x)\varphi_s(y)|k_s^{(r)}(x,y)}{(x^2+1)(y^2+1)}dxdy$$

$$\leq \iint_{\mathbb{R}^2} \frac{|\varphi_s(x)\varphi_s(y)|k_s(x,y)}{(x^2+1)(y^2+1)}dxdy$$

$$\leq \sum_{i.j=1}^{n} |c_i||c_j|H_s(z_i,z_j) < +\infty.$$

By the Lebesgue convergence theorem, we see that for every $r \in (0,1)$,

$$\iint_{\mathbb{R}^2} \frac{\varphi_s(x)\varphi_s(y)k_s^{(r)}(x,y)}{(x^2+1)(y^2+1)}dxdy$$

$$= \lim_{n\to\infty}\iint_{[-n,n]^2} \frac{\varphi_s(x)\varphi_s(y)k_s^{(r)}(x,y)}{(x^2+1)(y^2+1)}dxdy \geq 0.$$

and furthermore,

$$\iint_{\mathbb{R}^2} \frac{\varphi_s(x)\varphi_s(y)k_s(x,y)}{(x^2+1)(y^2+1)}dxdy$$

$$= \lim_{r\to 1-0}\iint_{\mathbb{R}^2} \frac{\varphi_s(x)\varphi_s(y)k_s^{(r)}(x,y)}{(x^2+1)(y^2+1)}dxdy \geq 0.$$

This completes Step 3 and the proof. □

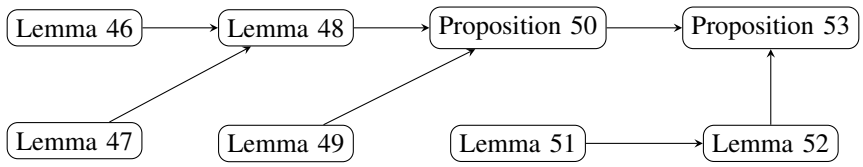

Fig. 2. A diagram showing relations between assertions

Remark 55. *It is easy to see that $H_{1/2}(z,w)$ is positive definite, because*

$$H_{1/2}(z,w) = \int_{\mathbb{R}} \sqrt{p_z(x)}\sqrt{p_w(x)}\mathrm{d}x.$$

See Remark 35 (ii).

Appendix C
f-divergences between the truncated Cauchy distribution

Now we define the *truncated* distribution of univariate location-scale families. Let $p(x)$ be a positive Borel measurable function on $\mathbb{R}$ such that $\int_{\mathbb{R}} p(x)\mathrm{d}x = 1$. Let $M > 0$. Let

$$p^{\mathrm{t}}(x) := \begin{cases} c_M p(x) & |x| \le M \\ 0 & |x| > M \end{cases}$$

where $c_M := 1/\int_{-M}^{M} p(x)\mathrm{d}x$. For $l \in \mathbb{R}$ and $s > 0$, let

$$p^{\mathrm{t}}_{l,s}(x) := \frac{1}{s} p^{\mathrm{t}}\left(\frac{x-l}{s}\right).$$

We consider the Kullback-Leibler divergence between truncated distributions.

Proposition 56. *Assume that $p(x)$ is continuous on $\mathbb{R}$ and $-Ms_2 + l_2 < -Ms_1 + l_1 < Ms_1 + l_1 < Ms_2 + l_2$. Then, $D_{\mathrm{KL}}(p^{\mathrm{t}}_{l_1,s_1} : p^{\mathrm{t}}_{l_2,s_2}) = +\infty$, and, $D_{\mathrm{KL}}(p^{\mathrm{t}}_{l_2,s_2} : p^{\mathrm{t}}_{l_1,s_1}) < +\infty$.*

Proof. By the assumption, we have that $\{x : p^{\mathrm{t}}_{l_1,s_1}(x) > 0\} \subsetneq \{x : p^{\mathrm{t}}_{l_2,s_2}(x) > 0\}$. Let $U := \{x : p^{\mathrm{t}}_{l_2,s_2}(x) > 0 = p^{\mathrm{t}}_{l_1,s_1}(x)\}$. Then,

$$U = [-Ms_2 + l_2, -Ms_1 + l_1) \cup (Ms_1 + l_1, Ms_2 + l_2]$$

and

$$-\log\left(\frac{p^{\mathrm{t}}_{l_1,s_1}(x)}{p^{\mathrm{t}}_{l_2,s_2}(x)}\right) = +\infty \quad x \in U.$$

Since U has inner points,

$$\int_U -\log\left(\frac{p^{\mathrm{t}}_{l_1,s_1}(x)}{p^{\mathrm{t}}_{l_2,s_2}(x)}\right) p^{\mathrm{t}}_{l_2,s_2}(x)\mathrm{d}x = +\infty.$$

On the other hand, by the assumption that p is continuous, $-\log\left(\frac{p^{\mathrm{t}}_{l_1,s_1}(x)}{p^{\mathrm{t}}_{l_2,s_2}(x)}\right) p^{\mathrm{t}}_{l_2,s_2}(x)$ is bounded on

$$\{x : p^{\mathrm{t}}_{l_2,s_2}(x) > 0\} \setminus U = \{x : p^{\mathrm{t}}_{l_1,s_1}(x) > 0\}$$
$$= [-Ms_1 + l_1, Ms_1 + l_1].$$

Hence,

$$\int_{\{x : p^{\mathrm{t}}_{l_2,s_2}(x) > 0\} \setminus U} -\log\left(\frac{p^{\mathrm{t}}_{l_1,s_1}(x)}{p^{\mathrm{t}}_{l_2,s_2}(x)}\right) p^{\mathrm{t}}_{l_2,s_2}(x)\mathrm{d}x < +\infty.$$

Hence,

$$D_{\mathrm{KL}}(p^{\mathrm{t}}_{l_1,s_1} : p^{\mathrm{t}}_{l_2,s_2})$$
$$= \int_{\{x : p^{\mathrm{t}}_{l_2,s_2}(x) > 0\}} -\log\left(\frac{p^{\mathrm{t}}_{l_1,s_1}(x)}{p^{\mathrm{t}}_{l_2,s_2}(x)}\right) p^{\mathrm{t}}_{l_2,s_2}(x)\mathrm{d}x = +\infty.$$

Since $-\log\left(\frac{p^{\mathrm{t}}_{l_2,s_2}(x)}{p^{\mathrm{t}}_{l_1,s_1}(x)}\right) p^{\mathrm{t}}_{l_1,s_1}(x)$ is bounded on $\{x : p^{\mathrm{t}}_{l_1,s_1}(x) > 0\} = [-Ms_1 + l_1, Ms_1 + l_1]$. Hence,

$$D_{\mathrm{KL}}(p^{\mathrm{t}}_{l_2,s_2} : p^{\mathrm{t}}_{l_1,s_1})$$
$$= \int_{\{x : p^{\mathrm{t}}_{l_1,s_1}(x) > 0\}} -\log\left(\frac{p^{\mathrm{t}}_{l_2,s_2}(x)}{p^{\mathrm{t}}_{l_1,s_1}(x)}\right) p^{\mathrm{t}}_{l_1,s_1}(x)\mathrm{d}x < +\infty.$$

□

The truncated Cauchy distribution and related issues have been investigated by many papers (e.g. [84], [85], [86], [87]). Let $M > 0$. Let

$$p^{\mathrm{tc}}(x) := \begin{cases} \dfrac{c_M}{1+x^2} & |x| \le M \\ 0 & |x| > M \end{cases}$$

where $c_M := 2\arctan(M)$. For $l \in \mathbb{R}$ and $s > 0$, let

$$p^{\mathrm{tc}}_{l,s}(x) := \frac{1}{s} p^{\mathrm{tc}}\left(\frac{x-l}{s}\right) = \begin{cases} \dfrac{c_M s}{s^2 + (x-l)^2} & |x-l| \le Ms \\ 0 & |x-l| > Ms \end{cases}.$$

Contrarily to the Cauchy distribution, this distribution has moments of every order.

By Proposition 56, the Kullback-Leibler divergence between truncated Cauchy distributions can be infinite.

Proposition 57. *The Pearson chi-square divergence between the scale family of the truncated Cauchy distribution can be asymmetric.*

Proof. Let $M = 10$, $(l_1, s_1) = (0, 1)$ and $(l_2, s_2) = (0, 2)$. Then,

$$D^P_\chi(p^{\mathrm{tc}}_{l_1,s_1} : p^{\mathrm{tc}}_{l_2,s_2}) = \int_{x : p^{\mathrm{tc}}_{l_1,s_1}(x) > 0} \frac{p^{\mathrm{tc}}_{l_2,s_2}(x)^2}{p^{\mathrm{tc}}_{l_1,s_1}(x)}\mathrm{d}x - 1$$
$$= c_{10}\int_{-10}^{10} \frac{4(x^2+1)}{(x^2+4)^2}\mathrm{d}x - 1,$$

and

$$D_\chi^P(p_{l_2,s_2}^{\mathrm{tc}} : p_{l_1,s_1}^{\mathrm{tc}}) = \int_{x:p_{l_2,s_2}(x)>0} \frac{p_{l_1,s_1}^{\mathrm{tc}}(x)^2}{p_{l_2,s_2}^{\mathrm{tc}}(x)} \mathrm{d}x - 1$$

$$= c_{10} \int_{-20}^{20} \frac{(x^2+4)}{2(x^2+1)^2} 1_{|x|\leq 10} \mathrm{d}x - 1$$

$$= c_{10} \int_{-10}^{10} \frac{(x^2+4)}{2(x^2+1)^2} \mathrm{d}x - 1.$$

We see that

$$\int_{-10}^{10} \frac{4(x^2+1)}{(x^2+4)^2} \mathrm{d}x = \frac{5}{52}(-3+26\arctan(5)) \fallingdotseq 3.14504$$

and

$$\int_{-10}^{10} \frac{x^2+4}{2(x^2+1)^2} \mathrm{d}x = \frac{15}{101} + \frac{5}{2}\arctan(10) \fallingdotseq 3.8263$$

Hence $D_\chi^P(p_{l_1,s_1}^{\mathrm{tc}} : p_{l_2,s_2}^{\mathrm{tc}}) \neq D_\chi^P(p_{l_2,s_2}^{\mathrm{tc}} : p_{l_1,s_1}^{\mathrm{tc}})$. □

Appendix D
Code snippet for Taylor expansions of f-divergences

We provide below a code using the Maxima[4] symbolic computing software to calculate the truncated Taylor series of f-divergences between two Cauchy distributions (instantiated for the Kullback-Leibler divergence).

```
Cauchy(x,l,s):=(s/(%pi*((x-l)**2+s**2)));
KLCauchy(l1,s1,l2,s2):=log(((s1+s2)**2+(l1-l2)**2)/(4*s1*s2));
l1:0;
s1:1;
l2:0.6;
s2:6/5;
k:40;
testcond: (9/16)-(l2**2+(s2-(4/5))**2);
print("Is condition>0 for Taylor expansion?:",testcond);
Cauchy1:Cauchy(x,l1,s1);
Cauchy2:Cauchy(x,l2,s2);
print("Exact KL");
KLCauchy(l1,s1,l2,s2);
ExactKL:float(%);
print("KL numerical integration:");
kla: quad_qagi( Cauchy1*log(Cauchy1/Cauchy2), x, minf,
inf,'epsrel=1d-10);
NumKL:float(kla[1]);

for i:2 while (i<=k)
do( r[i]: quad_qagi( (Cauchy1-Cauchy2)**i/Cauchy2**(i-1),
x, minf, inf,'epsrel=1d-10), print(i,r[i][1]));

print("KL Taylor truncated series:");
TaylorKL: sum( (((-1)**i)/i)*r[i][1], i, 2, k);
print("Exact:",ExactKL,"Numerical:",NumKL,
"Trunc.Taylor",TaylorKL);
print("Error |Taylor-Exact|",abs(TaylorKL-ExactKL));
```

Acknowledgment

The authors would like to express their gratitude to two anonymous referees for many fruitful comments.

[4]https://maxima.sourceforge.io/

References

[1] P. McCullagh, "Conditional inference and Cauchy models," *Biometrika*, vol. 79, no. 2, pp. 247–259, 1992.
[2] R. E. Kass and P. W. Vos, *Geometrical foundations of asymptotic inference*. John Wiley & Sons, 2011, vol. 908.
[3] J. Naudts, *Generalised thermostatistics*. Springer Science & Business Media, 2011.
[4] L. Filipovic and S. Selberherr, "A Two-Dimensional Lorentzian Distribution for an Atomic Force Microscopy Simulator," in *Monte Carlo Methods and Applications*. De Gruyter, 2012, pp. 97–104.
[5] S. Širca, "Special continuous probability distributions," in *Probability for Physicists*. Springer, 2016, pp. 65–91.
[6] J. Miller and J. Thomas, "Detectors for discrete-time signals in non-Gaussian noise," *IEEE Transactions on Information Theory*, vol. 18, no. 2, pp. 241–250, 1972.
[7] E. S. Sousa, "Performance of a spread spectrum packet radio network link in a Poisson field of interferers," *IEEE Transactions on Information Theory*, vol. 38, no. 6, pp. 1743–1754, 1992.
[8] N. Merhav, "Optimum estimation via gradients of partition functions and information measures: A statistical-mechanical perspective," *IEEE transactions on information theory*, vol. 57, no. 6, pp. 3887–3898, 2011.
[9] J. Fahs and I. Abou-Faycal, "A Cauchy input achieves the capacity of a Cauchy channel under a logarithmic constraint," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 3077–3081.
[10] I. Valero-Toranzo, S. Zozor, and J.-M. Brossier, "Generalization of the de bruijn identity to general ϕ-entropies and ϕ-fisher informations," *IEEE Transactions on Information Theory*, vol. 64, no. 10, pp. 6743–6758, 2017.
[11] N. Farvardin and J. Modestino, "Optimum quantizer performance for a class of non-gaussian memoryless sources," *IEEE Transactions on Information Theory*, vol. 30, no. 3, pp. 485–497, 1984.
[12] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.
[13] T. Koike-Akino and Y. Wang, "AutoVAE: Mismatched Variational Autoencoder with Irregular Posterior-Prior Pairing," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1689–1694.
[14] F. Nielsen, "On Voronoi diagrams on the information-geometric Cauchy manifolds," *Entropy*, vol. 22, no. 7, p. 713, 2020.
[15] C. Tsallis, "Introduction to nonextensive statistical mechanics: approaching a complex world," *Springer*, vol. 1, no. 1, pp. 2–1, 2009.
[16] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.
[17] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.
[18] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
[19] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.
[20] S.-i. Amari, *Information Geometry and Its Applications*, ser. Applied Mathematical Sciences. Springer Japan, 2016.
[21] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.
[22] F. Nielsen and K. Okamura, "On f-divergences between Cauchy distributions," arXiv preprint arXiv:2101.12459, Tech. Rep., 2021.
[23] F. Chyzak and F. Nielsen, "A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions," *arXiv preprint arXiv:1905.10965*, 2019.
[24] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2013.
[25] F. Nielsen, "On the Jensen-Shannon symmetrization of distances relying on abstract means," *Entropy*, vol. 21, no. 5, 2019. [Online]. Available: http://www.mdpi.com/1099-4300/21/5/485
[26] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory." *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.
[27] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.
[28] J.-F. Collet, "An exact expression for the gap in the data processing inequality for f-divergences," *IEEE Transactions on Information Theory*, vol. 65, no. 7, pp. 4387–4391, 2019.

[29] R. H. Risch, "The problem of integration in finite terms," *Transactions of the American Mathematical Society*, vol. 139, pp. 167–189, 1969.

[30] M. Bronstein, *Symbolic integration I: transcendental functions*. Springer Science & Business Media, 2005, vol. 1.

[31] K. P. Nelson and W. J. Thistleton, "Comments on "Generalized Box-Müller Method for Generating q-Gaussian Random Deviates"," *IEEE Transactions on Information Theory*, vol. 67, no. 10, pp. 6785–6789, 2021.

[32] P. McCullagh, "On the distribution of the Cauchy maximum-likelihood estimator," *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, vol. 440, no. 1909, pp. 475–479, 1993.

[33] F. B. Knight, "A characterization of the Cauchy type," *Proceedings of the American Mathematical Society*, vol. 55, no. 1, pp. 130–135, 1976.

[34] M. L. Eaton, *Group invariance applications in statistics*. Institute of Mathematical Statistics Hayward, California, 1989.

[35] Ç. Candan, "Chebyshev center computation on probability simplex with α-divergence measure," *IEEE Signal Processing Letters*, vol. 27, pp. 1515–1519, 2020.

[36] E. Welzl, "Smallest enclosing disks (balls and ellipsoids)," in *New results and new trends in computer science*. Springer, 1991, pp. 359–370.

[37] F. Nielsen, "An information-geometric characterization of Chernoff information," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 269–272, 2013.

[38] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bull. Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[39] S. Kato, M. Jones *et al.*, "An extended family of circular distributions related to wrapped Cauchy distributions via Brownian motion," *Bernoulli*, vol. 19, no. 1, pp. 154–171, 2013.

[40] A. Pewsey, M. Neuhäuser, and G. D. Ruxton, *Circular statistics in R*. Oxford University Press, 2013.

[41] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped Cauchy distribution," *Journal of Applied Statistics*, vol. 15, no. 2, pp. 247–254, 1988.

[42] A. W. Marshall and I. Olkin, *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*. Springer, 2007.

[43] K. V. Mardia and P. E. Jupp, *Directional statistics*. John Wiley & Sons, 2009, vol. 494.

[44] Y. Qiao and N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3884–3890, 2010.

[45] A. A. Ungar, "The holomorphic automorphism group of the complex disk," *Aequationes mathematicae*, vol. 47, no. 2-3, pp. 240–254, 1994.

[46] T. Needham, *Visual complex analysis*. Oxford University Press, 1998.

[47] F. Nielsen, "An elementary introduction to information geometry," *Entropy*, vol. 22, no. 10, p. 1100, 2020.

[48] ——, "On information projections between multivariate elliptical and location-scale families," arXiv, Tech. Rep., 2021, 2101.03839.

[49] Y. Akaoka, K. Okamura, and Y. Otobe, "Bahadur efficiency of the maximum likelihood estimator and one-step estimator for quasi-arithmetic means of the cauchy distribution," *Annals of the Institute of Statistical Mathematics*, vol. 74, no. 5, pp. 895–923, 2022.

[50] D. J. Olive, *Statistical theory and inference*. Springer, 2014.

[51] S. J. Press, "Multivariate stable distributions," *Journal of Multivariate Analysis*, vol. 2, no. 4, pp. 444–462, 1972.

[52] S. Kesavan, *Measure and Integration*. Springer, 2019.

[53] F. Nielsen and G. Hadjeres, "On power chi expansions of f-divergences," *arXiv preprint arXiv:1903.05818*, 2019.

[54] R. L. Graham, D. E. Knuth, O. Patashnik, and S. Liu, "Concrete mathematics: a foundation for computer science," *Computers in Physics*, vol. 3, no. 5, pp. 106–107, 1989.

[55] K. Jain and A. Srivastava, "On symmetric information divergence measures of Csiszar's f-divergence class," *Journal of Applied Mathematics, Statistics and Informatics (JAMSI)*, vol. 3, no. 1, pp. 85–102, 2007.

[56] S. S. Dragomir *et al.*, "A refinement of Jensen's inequality with applications for f-divergence measures," *Taiwanese Journal of Mathematics*, vol. 14, no. 1, pp. 153–164, 2010.

[57] P. Kafka, F. Österreicher, and I. Vincze, "On powers of f-divergences defining a distance," *Studia Sci. Math. Hungar*, vol. 26, no. 4, pp. 415–422, 1991.

[58] F. Österreicher and I. Vajda, "A new class of metric divergences on probability spaces and its applicability in statistics," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 3, pp. 639–653, 2003.

[59] I. Vajda, "On metric divergences of probability measures," *Kybernetika*, vol. 45, no. 6, pp. 885–900, 2009.

[60] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," in *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.* IEEE, 2004, p. 31.

[61] S. Acharyya, A. Banerjee, and D. Boley, "Bregman divergences and triangle inequality," in *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 2013, pp. 476–484.

[62] M. d. Berg, M. v. Kreveld, M. Overmars, and O. Schwarzkopf, *Computational geometry*. Springer, 1997.

[63] S. Lloyd, "Least squares quantization in PCM," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[64] P. N. Yianilos, "Data structures and algorithms for nearest neighbor," in *Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms*, vol. 66. SIAM, 1993, p. 311.

[65] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, 2003, pp. 147–153.

[66] M. B. Westover, "Asymptotic geometry of multiple hypothesis testing," *IEEE transactions on information theory*, vol. 54, no. 7, pp. 3327–3329, 2008.

[67] K. Li, "Discriminating quantum states: The multiple Chernoff distance," *The Annals of Statistics*, vol. 44, no. 4, pp. 1661–1679, 2016.

[68] F. Nielsen and S. Boltz, "The Burbea-Rao and Bhattacharyya centroids," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5455–5466, 2011.

[69] B. C. Carlson and J. L. Gustafson, "Asymptotic expansion of the first elliptic integral," *SIAM Journal on Mathematical Analysis*, vol. 16, pp. 1072–1092, 1985.

[70] S. Eguchi, "Second order efficiency of minimum contrast estimators in a curved exponential family," *The Annals of Statistics*, pp. 793–803, 1983.

[71] ——, "Geometry of minimum contrast," *Hiroshima Mathematical Journal*, vol. 22, no. 3, pp. 631–647, 1992.

[72] I. J. Schoenberg, "Metric spaces and positive definite functions," *Transactions of the American Mathematical Society*, vol. 44, no. 3, pp. 522–536, 1938.

[73] J. Faraut and K. Harzallah, "Distances hilbertiennes invariantes sur un espace homogène," *Ann. Inst. Fourier (Grenoble)*, vol. 24, no. 3, pp. xiv, 171–217, 1974.

[74] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 8th ed. Academic press, 2015.

[75] R. Takahashi, "Sur les représentations unitaires des groupes de Lorentz généralisés," *Bull. Soc. Math. France*, vol. 91, pp. 289–433, 1963.

[76] J. Faraut and K. Harzallah, "Fonctions sphériques de type positif sur les espaces hyperboliques," *C. R. Acad. Sci. Paris Sér. A-B*, vol. 274, pp. A1396–A1398, 1972.

[77] S. Helgason, "A duality for symmetric spaces with applications to group representations," *Advances in Math.*, vol. 5, pp. 1–154, 1970.

[78] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic analysis on semigroups. Theory of positive definite and related functions*. Springer, Cham, 1984, vol. 100.

[79] P. Petersen, *Riemannian geometry*, 2nd ed., ser. Graduate Texts in Mathematics. Springer, New York, 2006, vol. 171.

[80] P. McCullagh, "Möbius transformation and Cauchy parameter estimation," *Annals of statistics*, vol. 24, no. 2, pp. 787–808, 1996.

[81] J.-D. Boissonnat, F. Nielsen, and R. Nock, "Bregman Voronoi diagrams," *Discrete & Computational Geometry*, vol. 44, no. 2, pp. 281–307, 2010.

[82] O. E. Barndorff-Nielsen, P. Blæsild, and P. S. Eriksen, *Decomposition and invariance of measures, and statistical transformation models*. Springer Science & Business Media, 2012, vol. 58.

[83] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, "Functional inequalities for hypergeometric functions and complete elliptic integrals," *SIAM journal on mathematical analysis*, vol. 23, no. 2, pp. 512–524, 1992.

[84] S. Nadarajah and S. Kotz, "A truncated t distribution," *The Mathematical Scientist*, vol. 29, no. 2, pp. 122–126, 2004.

[85] ——, "A truncated bivariate Cauchy distribution," *Bulletin of the Malaysian Mathematical Sciences Society. Second Series*, vol. 30, no. 2, pp. 185–193, 2007.

[86] S. F. Ateya and E. A. Madhagi, "On multivariate truncated generalized Cauchy distribution," *Statistical Papers*, vol. 54, no. 3, pp. 879–897, 2013.

[87] M. A. Aldahlan, F. Jamal, C. Chesneau, M. Elgarhy, and I. Elbatal, "The truncated Cauchy power family of distributions with inference and applications," *Entropy*, vol. 22, no. 3, p. article no. 346, 2020.

Frank Nielsen Frank Nielsen was awarded his PhD on adaptive computational geometry (1996) from University of Côte d'Azur (France). He is a fellow of Sony Computer Science Laboratories Inc. (Japan) where he currently conducts research on the fundamentals and practice of geometric machine learning and intelligence with applications in visual computing. Frank Nielsen co-organizes with Frédéric Barbaresco the biannual conference Geometric Science of Information (GSI).

Kazuki Okamura Kazuki Okamura was awarded his PhD on mathematical sciences (2015) from the University of Tokyo (Japan). He is a lecturer in Department of Mathematics in Faculty of Science at Shizuoka University, Japan. His research focuses mainly on probability theory and its applications.