On the Kullback-Leibler divergence between location-scale densities

Frank Nielsen
Sony Computer Science Laboratories Inc, Japan

Abstract

We show that the f-divergence between any two densities of potentially different location-scale families can be reduced to the calculation of the f-divergence between one standard density with another location-scale density. It follows that the f-divergence between two scale densities depends only on the scale ratio. We then report conditions on the standard distribution to get symmetric f-divergences. We illustrate this symmetric property with the calculation of the Kullback-Leibler divergence between scale Cauchy distributions. Finally, we show that the minimum f-divergence of any query density of a location-scale family to another location-scale family is independent of the query location-scale parameters.

Keywords: Location-scale family, Kullback-Leibler divergence, location-scale group, Cauchy distributions.

1 Introduction

Let $X \sim p$ be a random variable with cumulative distribution function $F_X$ and probability density $p_X(x)$ on the support $\mathcal{X}$ (usually $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}_{++}$). A location-scale random variable $Y = l + sX$ (for location parameter $l \in \mathcal{X}$ and scale parameter $s > 0$) has distribution $F_Y(y) = F_X(\frac{y - l}{s})$ and density $p_Y(y) = p_X(\frac{y - l}{s})$. The location-scale group $\mathbb{H} = \{(l, s): l \in \mathbb{R} \times \mathbb{R}_{++}\}$ acts on the densities of a location-scale family $\mathcal{F}$. The identity element is $i = (0, 1)$, the group operation $e_1.e_2$ yields $e_1.e_2 = (l_1 + s_1l_2, s_1s_2)$ for $e_1 = (l_1, s_1)$ and $e_2 = (l_2, s_2)$, and the inverse element $e^{-1}$ is $e^{-1} = (-\frac{l}{s}, \frac{1}{s})$ for $e = (l, s)$.

Consider two location-scale families $\mathcal{F}$ sharing the same support $\mathcal{X}$:

$\mathcal{F}_1 = \left\{ p_{l_1,s_1}(x) = \frac{1}{s_1} p\left(\frac{x - l_1}{s_1}\right) : (l_1, s_1) \in \mathbb{H} \right\}$,

and

$\mathcal{F}_2 = \left\{ q_{l_2,s_2}(x) = \frac{1}{s_2} q\left(\frac{x - l_2}{s_2}\right) : (l_2, s_2) \in \mathbb{H} \right\}$,

where $p(x) = p_{0,1}(x)$ and $q(x) = q_{0,1}(x)$ denote the standard densities of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively (also called reduced distributions $\mathcal{F}$).

A location family is a subfamily of a location-scale family, with fixed scale $s_0$. We denote by $p_l = p_{l,s_0}$ the density of a location family. Similarly, a scale family is a subfamily of a location-scale family with prescribed location $l_0$. We denote by $p_s = p_{l_0,s}$ the density of a scale family.

For example, $\mathcal{F}_1$ can be the Cauchy family with standard distribution $p(x) = \frac{1}{\pi(1+x^2)}$ and $\mathcal{F}_2$ the normal family with standard distribution $q(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, both families with support...
The cross-entropy \[ h^\times(p_{t_1,s_1} : q_{t_2,s_2}) \] between a density \( p_{t_1,s_1} \) of \( \mathcal{F}_1 \) and a density \( q_{t_2,s_2} \) of \( \mathcal{F}_2 \) is defined by
\[
h^\times(p_{t_1,s_1} : q_{t_2,s_2}) = - \int_{\mathcal{X}} p_{t_1,s_1}(x) \log q_{t_2,s_2}(x) \, dx.
\] (1)

The differential entropy \[ h \] is the self cross-entropy:
\[
h(p_{t,s}) = h^\times(p_{t,s} : p_{t,s}).
\] (2)

The Kullback-Leibler (KL) divergence is the difference between the cross-entropy and the entropy:
\[
\text{KL}(p_{t_1,s_1} : q_{t_2,s_2}) = h^\times(p_{t_1,s_1} : q_{t_2,s_2}) - h(p_{t_1,s_1}) = \int_{\mathcal{X}} p_{t_1,s_1}(x) \log \frac{p_{t_1,s_1}(x)}{q_{t_2,s_2}(x)} \, dx \geq 0.
\] (3)

Note that the KL divergence between a standard Cauchy distribution and a standard Gaussian distribution is \emph{infinite} since the integral diverges but the KL divergence between a standard Gaussian distribution and a standard Cauchy distribution is finite. Thus the KL divergence between any two arbitrary location-scale families may potentially be infinite and may not admit a closed-form formula using the parameters \((l_1, s_1; l_2, s_2)\).

By making some changes of variable for \( x \) in the cross-entropy integral on the right-hand-side of Eq. 1 we establish the following four basic identities:

**Left scale multiplication.**
\[
h^\times(p_{t_1,\lambda s_1} : q_{t_2,s_2}) = h^\times\left(p_{\lambda \frac{t_1}{\lambda}, s_1} : q_{t_2, s_2} \right) + \log \lambda_1, \quad \forall \lambda_1 \in \mathbb{R}^+. \tag{4}
\]

**Proof.** Make a change of variable in the integral with \( y = \frac{x}{\lambda_1} \) for \( \lambda_1 > 0 \) (or \( x = \lambda_1 y \)) and \( dx = \lambda_1 dy \). Then we have
\[
h^\times(p_{t_1,\lambda s_1} : q_{t_2,s_2}) = - \int \frac{1}{\lambda_1 s_1} p\left(\frac{x-l_1}{\lambda_1 s_1}\right) \log \frac{1}{s_2} q\left(\frac{x-l_2}{s_2}\right) \, dx,
\] (5)
\[
= - \int \frac{1}{s_2} p\left(\frac{y-l_1}{s_1}\right) \log \frac{1}{s_2} \lambda_1 q\left(\frac{y-l_2}{s_2}\right) \, dy,
\] (6)
\[
= - \int p_{\lambda_1 \frac{t_1}{\lambda_1}, s_1}(y) \log \frac{1}{s_2} \lambda_1 q\left(\frac{y-l_2}{s_2 \lambda_1}\right) \, dy + \log \lambda_1 \int p_{\lambda_1 \frac{t_1}{\lambda_1}, s_1}(y) \, dy,
\] (7)
\[
= h^\times\left(p_{\lambda_1 \frac{t_1}{\lambda_1}, s_1} : q_{\frac{t_2}{\lambda_1}, \frac{s_2}{\lambda_1}} \right) + \log \lambda_1. \tag{8}
\]
Left location translation.

\[ h^x(p_{l_1+\alpha_1, s_1} : q_{l_2, s_2}) = h^x(p_{l_1, s_1} : q_{l_2-\alpha_1, s_2}), \quad \forall \alpha_1 \in \mathbb{R}. \]  \hspace{1cm} (9)

Right scale multiplication.

\[ h^x(p_{l_1, s_1} : q_{l_2, \lambda s_2}) = h^x \left( \frac{p_{l_1, s_1}}{\lambda s_2} : \frac{q_{l_2, s_2}}{s_2} \right) + \log \lambda_2, \quad \forall \lambda_2 \in \mathbb{R}_{++}. \]  \hspace{1cm} (10)

Right location translation.

\[ h^x(p_{l_1, s_1} : q_{l_2+\alpha_2, s_2}) = h^x(p_{l_1-\alpha_2, s_1} : q_{l_2, s_2}), \quad \forall \alpha_2 \in \mathbb{R}. \]  \hspace{1cm} (11)

Furthermore, we get the following double-sided scale multiplication identity by a change of variable (can also be obtained by applying the left scale multiplication with parameter \( \lambda_1 = \sqrt{\lambda} \) and then the right scale multiplication with parameter \( \lambda_2 = \sqrt{\lambda} \):

\[ h^x(p_{l_1, \lambda s_1} : q_{l_2, \lambda s_2}) = h^x \left( \frac{p_{l_1, s_1}}{\lambda s_2} : \frac{q_{l_2, s_2}}{s_2} \right) + \log \lambda, \quad \forall \lambda > 0, \]  \hspace{1cm} (12)

and the generic cross-entropy rule by translations:

\[ h^x(p_{l_1+\alpha, s_1} : q_{l_2+\beta, s_2}) = h^x(p_{l_1, s_1} : q_{l_2+\beta-\alpha, s_2}) = h^x(p_{l_1+\alpha-\beta, s_1} : q_{l_2, s_2}), \quad \forall \alpha, \beta \in \mathbb{R}. \]  \hspace{1cm} (13)

By using these “parameter rewriting” rules, we get the following properties:

**Property 1** (Location-scale entropy). We have

\[ h(p_{l, s}) = h(p) + \log s. \]  \hspace{1cm} (14)

That is, the differential entropy of a density \( p_{l, s} \) of a location-scale family is independent of the location and can be calculated from the entropy of the standard density \( p \).

**Proof.** We have \( h(p_{l, s}) = h^x(p_{l, s} : p_{l, s}) = h^x(p_{0, s} : p_{0, s}) \) (using either the left/right translation rule) and \( h^x(p_{0, s} : p_{0, s}) = h^x(p_{0, 1} : p_{0, 1}) + \log s = h(p) + \log s \) (using either left/right multiplication rule). \( \square \)

**Property 2** (Location-scale cross-entropy). We have

\[ h^x(p_{l_1, s_1} : q_{l_2, s_2}) = h^x \left( \frac{p_{l_1, s_1}}{s_2} : \frac{q_{l_2, s_2}}{s_2} \right) + \log s_2, \]  \hspace{1cm} (15)

\[ = h^x(p : \frac{q_{l_2-1, s_2}}{s_1} : \frac{s_2}{s_1}) + \log s_1. \]  \hspace{1cm} (16)

That is, the cross-entropy between two location-scale densities can be reduced to the calculation of the cross-entropy between a standard density and a density of the other location-scale family.

**Proof.** We have \( h^x(p_{l_1, s_1} : q_{l_2, s_2}) = h^x(p_{l_1, s_1} : q_{l_2, 1}) + \log s_2 \) (right multiplication rule) and \( h^x \left( \frac{p_{l_1, s_1}}{s_2} : \frac{q_{l_2, s_2}}{s_2} \right) = h^x \left( \frac{p_{l_1, s_1}}{s_2} : \frac{q_{l_2, 1}}{s_2} \right) + \log s_2 \) (right translation rule). \( \square \)
Property 3 (Location-scale Kullback-Leibler divergence). We have

\[
\text{KL}(p_{l_1,s_1} : q_{l_2,s_2}) = h^\times \left( p : q_{l_2-l_1}^{s_2/s_1} \right) - h(p) = \text{KL} \left( p : q_{l_2-l_1}^{s_2/s_1} \right),
\]

\[
= h^\times \left( p_{l_1+l_s}^{s_2/s_1} : q \right) - h(p) + \log \frac{s_2}{s_1} = \text{KL}(p_{l_1+l_s}^{s_2/s_1} : q).
\]

Proof. We have \(\text{KL}(p_{l_1,s_1} : q_{l_2,s_2}) = h^\times (p_{l_1,s_1} : q_{l_2,s_2}) - h(p_{l_1,s_1})\). Then we apply Property 2

\[
h^\times (p_{l_1,s_1} : q_{l_2,s_2}) = h^\times \left( p : q_{l_2-l_1}^{s_2/s_1} \right) + \log s_1 \quad \text{and} \quad \text{Property 1} \quad h(p_{l_1,s_1}) = h(p) + \log s_1 \quad \text{to get the result (the terms log } s_1 \text{ cancel out)}.
\]

Similarly, we have the following basic identities for the Kullback-Leibler divergence between any two location-scale densities:

\[
\text{KL}(p_{l_1+l_s}^{s_1} : q_{l_2,s_2}) = \text{KL}(p_{l_1,s_1} : q_{l_2-l_1}^{s_2/s_1}),
\]

\[
\text{KL}(p_{l_1,s_1}^{l_s} : q_{l_2,s_2}) = \text{KL}(p_{l_1,s_1} : q_{l_2}^{s_2/s_1}),
\]

\[
\text{KL}(p_{l_1,s_1} : q_{l_2+l_s}^{s_1}) = \text{KL}(p_{l_1-l_1,s_1} : q_{l_2,s_2}),
\]

\[
\text{KL}(p_{l_1,s_1} : q_{l_2,s_2}^{l_s}) = \text{KL}(p_{l_1,s_1}^{l_s} : q_{l_2}^{s_2/s_1}^{l_s}).
\]

We state the following theorem:

Theorem 1. The Kullback-Leibler divergence between two densities belonging to the same scale family is scale invariant: \(\text{KL}(p_{\lambda s_1} : p_{\lambda s_2}) = \text{KL}(p_{s_1} : p_{s_2})\) for any \(\lambda > 0\).

Proof. We have \(\text{KL}(p_{\lambda s_1} : p_{\lambda s_2}) = \text{KL}(p : p_{\lambda s_2}^{\lambda s_1}) = \text{KL}(p : p_{s_2}^{s_1}) = \text{KL}(p_{s_1} : p_{s_2})\). 

We can define a scalar divergence \(D(s_1 : s_2) := \text{KL}(p_{s_1} : p_{s_2})\) that is scale-invariant: \(D(\lambda s_1 : \lambda s_2) = D(s_1 : s_2)\) for any \(\lambda > 0\). Another common example of scalar divergence is the Itakura-Saito divergence which belongs to the class of Bregman divergences [6].

The result presented for the KL divergence holds in the more general setting of Csiszár’s \(f\)-divergences [5,13]:

\[
I_f(p : q) = \int_X p(x) f \left( \frac{q(x)}{p(x)} \right) \, dx,
\]

for a positive convex function \(f\), strictly convex at 1, with \(f(1) = 0\). The KL divergence is a \(f\)-divergence for the generator \(f(u) = -\log u\).

Theorem 2. The \(f\)-divergence between two location-scale densities \(p_{l_1,s_1}\) and \(q_{l_2,s_2}\) can be reduced to the calculation of the \(f\)-divergence between one standard density with another location-scale density:

\[
I_f(p_{l_1,s_1} : q_{l_2,s_2}) = I_f \left( p : q_{l_2-l_1}^{s_2/s_1} \right) = I_f \left( p_{l_2-l_1}^{s_2/s_1} : q \right).
\]
Proof. The proofs follow from changes of the variable $x$ in the integral: Consider $y = \frac{x-l_1}{s_1}$ with $dx = s_1dy$, $x = s_1y + l_1$ and $\frac{x-l_2}{s_2} = \frac{s_1y+l_1-l_2}{s_2} = \frac{y-\frac{l_2-l_1}{s_1}}{s_1}$.

$$I_f(p_{l_1, s_1} : q_{l_2, s_2}) := \int_X p_{l_1, s_1}(x) f \left( \frac{q_{l_2, s_2}(x)}{p_{l_1, s_1}(x)} \right) dx,$$

$$= \int Y \frac{1}{s_1} p(y) f \left( \frac{\frac{1}{s_2} q \left( \frac{y-\frac{l_2-l_1}{s_1}}{s_2} \right)}{s_1 p(y)} \right) s_1 dy,$$

$$= \int p(y) f \left( \frac{\frac{q_{l_2-l_1, \frac{s_1}{s_2}}(y)}{p(y)}}{p(y)} \right) dy,$$

$$= I_f \left( p : q_{l_2-l_1, \frac{s_1}{s_2}} \right).$$

The proof for $I_f(p_{l_1, s_1} : q_{l_2, s_2}) = I_f(p_{l_1-l_2, \frac{s_1}{s_2}} : q)$ is similar, or one can use the adjoint generator $f^*(u) = uf(\frac{1}{u})$ which yields the reverse $f$-divergence: $I_{f^*}(p : q) = I_f(q : p)$.}

Note that $f$-divergences are invariant under any diffeomorphism $y = t(x)$ of the sample space $X$. The $f$-divergences are called invariant divergences in information geometry [1]. In particular, this invariance property includes the diffeomorphism defined by the group action of the location-scale group.

Thus the $f$-divergences between scale densities amount to a scale-invariant scalar distance:

$$D_f(s_1 : s_2) := I_f(p_{s_1} : q_{s_2}) = I_f \left( p : q_{\frac{s_2}{s_1}} \right) =: D_f \left( \frac{s_2}{s_1} : 1 \right),$$

$$= I_f \left( p_{\frac{s_2}{s_1}} : q \right) =: D_f \left( \frac{s_1}{s_2} : 1 \right).$$

2 The KL divergence between Cauchy location-scale distributions

In this section, we consider a working example for the scale Cauchy family. Surprisingly, the formula has not been widely reported in the literature (an erratum [1] corrects the formula given in [15]). Note that the Cauchy scale family can also be interpreted as a $q$-Gaussian family for $q = 2$ [1] and a $\alpha$-stable family [15] for $\alpha = 1$.

Consider the cross-entropy between two Cauchy distributions $p_1$ and $p_2$. Using Property [8], we can assume without loss of generality that the distribution $p_2$ is the standard Cauchy distribution $p$, and focus on calculating the following cross-entropy:

$$h^\times (p_{l,s} : p) = -\int_{-\infty}^{\infty} p_{l,s}(x) \log p(x) dx,$$
with location \( l = \frac{t_1 - t_2}{s_1} \) and scale \( s = \frac{s_1}{s_2} \), where
\[
p(x) = \frac{1}{\pi(1 + x^2)}, \quad p_{l,s}(x) = \frac{s}{\pi(s^2 + (x-l)^2)}.
\] (32)

The scale Cauchy distributions form a subfamily with \( l = 0 \). We shall use the following result on definite integrals (listed under the logarithmic forms of definite integrals in many handbooks of formulas and tables)\(^2\)
\[
A(a, b) = \int_{-\infty}^{\infty} \frac{\log(a^2 + x^2)}{b^2 + x^2} \, dx = \frac{2\pi}{b} \log(a + b), \quad a, b > 0.
\] (33)

We get the cross-entropy between two scale Cauchy distributions \( p_{s_1} \) and \( p_{s_2} \) as follows:
\[
h^\times(p_{s_1} : p_{s_2}) = h^\times(p_s : p) + \log s_2,
\] (34)
\[
= \frac{s}{\pi} \int \frac{1}{s^2 + x^2} \log(1 + x^2) \, dx + \log \pi + \log s_2,
\] (35)
\[
= \log \pi s_2 + \frac{s}{\pi} I(1, s),
\] (36)
\[
= \log \pi \frac{(s_1 + s_2)^2}{s_2}.
\] (37)

The differential entropy is obtained for \( s_1 = s_2 = s \):
\[
h(p_s) = h^\times(p_s : p_s) = \log 4\pi s,
\] (38)
in accordance with \[^5\] (p. 68). Thus the Kullback-Leibler between two scale Cauchy distributions is:
\[
\text{KL}(p_{s_1} : p_{s_2}) = h^\times(p_{s_1} : p_{s_2}) - h(p_{s_1}) = 2 \log \left( \frac{s_1 + s_2}{2\sqrt{s_1s_2}} \right),
\] (39)
\[
= 2 \log \left( \frac{1 + \frac{s_2}{s_1}}{2\sqrt{\frac{s_2}{s_1}}} \right) = 2 \log \left( \frac{1 + \frac{s_1}{s_2}}{2\sqrt{\frac{s_1}{s_2}}} \right),
\] (40)

Notice that \( A(s_1, s_2) = \frac{s_1 + s_2}{2} \) is the arithmetic mean of the scales, and \( G(s_1, s_2) = \sqrt{s_1s_2} \) is the geometric mean of the scales. Thus the KL divergence can be rewritten as \( \text{KL}(p_{s_1} : p_{s_2}) = 2 \log \frac{A(s_1, s_2)}{G(s_1, s_2)} \). Since we have the arithmetic-geometric inequality \( A \geq G \) (and \( \frac{A}{G} \geq 1 \)), it follows that \( \text{KL}(p_{s_1} : p_{s_2}) \geq 0 \).

Let us notice that the KL divergence between two Cauchy scale distributions is symmetric: \( \text{KL}(p_{s_1} : p_{s_2}) = \text{KL}(p_{s_2} : p_{s_1}) \). For exponential families \[^{11}\], the KL divergence is provably symmetric only for the location (multivariate/elliptical) Gaussian family since the KL divergence amount to a Bregman divergence, and the only symmetric Bregman divergences are the squared Mahalanobis distances \[^2\]. Not all scale families are symmetric: For example, the Rayleigh distributions form a scale family (and also an exponential family \[^{11}\]) but the KL divergence between two Rayleigh distributions amount to an Itakura-Saito divergence \[^{11}\] that is asymmetric.

\[^2\] Also listed online at [https://en.wikipedia.org/wiki/List_of_definite_integrals](https://en.wikipedia.org/wiki/List_of_definite_integrals)
Proposition 1. The differential entropy, cross-entropy and Kullback-Leibler divergence between two scale Cauchy densities $p_{s_1}$ and $p_{s_2}$ are:

$$h(p_s) = \log 4\pi s,$$

$$h^\times(p_{s_1}:p_{s_2}) = \log \pi \frac{(s_1 + s_2)^2}{s_2},$$

$$\text{KL}(p_{s_1}:p_{s_2}) = 2\log \left(\frac{s_1 + s_2}{2\sqrt{s_1 s_2}}\right).$$

Corollary 1. The Kullback-Leibler divergence between two Cauchy scale distributions is scale invariant.

Proof. Theorem 1 already proves this property for any scale family including the Cauchy scale family. However, here we shall directly use the property of homogeneous means. Since for all $\lambda > 0$, we have $A(\lambda s_1, \lambda s_2) = \lambda A(s_1, s_2)$ and $G(\lambda s_1, \lambda s_2) = \lambda G(s_1, s_2)$, it follows that $A(s_1^2, \lambda s_2^2) = \lambda A(s_1, s_2)$, and we have $\text{KL}(p_{s_1^2}:p_{s_2^2}) = \text{KL}(p_{s_1}:p_{s_2})$. □

Let us mention the generic formula 3 for the Kullback-Leibler divergence between Cauchy location-scale density $p_{l_1,s_1}$ and $p_{l_2,s_2}$ is

$$\text{KL}(p_{l_1,s_1}:p_{l_2,s_2}) = \log \frac{(s_1 + s_2)^2 + (l_1 - l_2)^2}{4s_1 s_2}. \quad (41)$$

3 Conditions on the standard density for symmetric KL divergences

Let us study when the KL divergence between location-scale families is symmetric by characterizing the standard distribution: $\text{KL}(p_{l_1,s_1}:p_{l_2,s_2}) = \text{KL}(p_{l_2,s_2}:p_{l_1,s_1})$. Since $\text{KL}(p_{l_1,s_1}:p_{l_2,s_2}) = \text{KL}(p:p_{l,s})$ (with $s = \frac{s_1}{s_2}$ and $l = \frac{l_1 - l_2}{s_1}$), we consider the case where

$$\text{KL}(p:p_{l,s}) = \text{KL}(p_{l,s}:p) = \text{KL}\left(p:p_{\frac{1}{2},-\frac{l}{2}}\right). \quad (42)$$

The equality $\text{KL}(p:p_{l,s}) = \text{KL}(p:p_{\frac{1}{2},-\frac{l}{2}})$ yields the following equivalent condition:

$$\int_{\mathcal{X}} p(x) \log \frac{p_{\frac{1}{2},-\frac{l}{2}}(x)}{p_{l,s}(x)} dx = 0, \quad \forall l \in \mathbb{R}, s > 0. \quad (43)$$

Assume a location family (i.e., $s = 1$), then we have the condition

$$\int_{\mathcal{X}} p(x) \log \frac{p_{l}(x)}{p_{1}(x)} dx = 0, \quad \forall l \in \mathbb{R}, s > 0. \quad (44)$$

Since $p_{-l}(x) = p(x + l)$ and $p_{l}(x) = p(x - l)$, we end up with the condition

$$\int_{\mathcal{X}} p(x) \log \frac{p(x + l)}{p(x - l)} dx = 0, \quad \forall l \in \mathbb{R}. \quad (45)$$

For example, the location normal distribution has symmetric KL divergence because it satisfies Eq. 45. Indeed, for normal location distributions, we have $\int_{\mathcal{X}} p(x) \log \frac{p_{l}(x)}{p_{0}(x)} dx = 2l \int_{\mathcal{X}} 2x p(x) = 2l E[x] = 0$ since $p(x)$ for the standard Gaussian density is an even function.
Consider now the scale family (with \( l = 0 \)), then we find the following condition
\[
\int_X p(x) \log \frac{p(\frac{x}{s})}{p(sx)} \, dx = 2 \log s, \quad \forall s \in \mathbb{R}^+.
\]  
(46)

For example, the Cauchy scale distribution has symmetric KL divergence because the Cauchy standard distribution satisfies Eq. (46)

\[
\int_X p(x) \log \frac{p(\frac{x}{s})}{p(sx)} \, dx = \frac{1}{\pi} \log \frac{\pi(1 + s^2x^2)}{\pi 1 + \frac{x^2}{2}},
\]  
(47)
\[
= \frac{1}{\pi} \left( A(s, 1) - A \left( \frac{1}{s}, 1 \right) \right),
\]  
(48)
\[
= 2 \log \frac{1 + s}{1 + \frac{1}{s}} = 2 \log s.
\]  
(49)

Similarly, the \( f \)-divergence between two densities \( p \) and \( q \) is symmetric if and only if:
\[
\int_X \left( p(x)f \left( \frac{q(x)}{p(x)} \right) - q(x)f \left( \frac{p(x)}{q(x)} \right) \right) \, dx = 0.
\]  
(50)

### 4 Kullback-Leibler minimizations between location-scale families

Consider the density manifold \( M \) (Fréchet manifold), and two densities \( p \) and \( q \) of \( M \). We can generate the location-scale families/submanifolds \( P = \{ \frac{1}{s} p \left( \frac{x-l}{s} \right) : (l, s) \in \mathbb{H} \} \) and \( Q = \{ \frac{1}{s} q \left( \frac{x-l}{s} \right) : (l, s) \in \mathbb{H} \} \), where \( \mathbb{H} = \mathbb{R} \times \mathbb{R}^+ \) is the open half-space of 2D location-scale parameters.

Consider the following Kullback-Leibler minimization problem:
\[
\text{KL}(p_{l_1,s_1} : Q) := \min_{(l_2,s_2) \in \mathbb{H}} \text{KL}(p_{l_1,s_1} : q_{l_2,s_2})
\]  
(51)
\[
\equiv \min_{(l_2,s_2) \in \mathbb{H}} \text{KL}(p : q_{l_2-l_1,s_2})
\]  
(52)
\[
\equiv \min_{(l,s) \in \mathbb{H}} \text{KL}(p : q_{l,s}) := \text{KL}(p : Q),
\]  
(53)
(54)

with \( l = \frac{l_2-l_1}{s_1} \) and \( s = \frac{s_2}{s_1} \). Once the best parameters \((l^*, s^*)\) have been calculated for a query density \( p_{l_1,s_1} \), we get the minimizer on the other location-scale family as \( l^*_2 = s_1 l^* + l_1 \) and \( s^*_2 = s^* s_1 \).

We have \( \text{KL}(p : q_{l,s}) = h^*(p : q_{l,s}) - h(p) \), and therefore \( \min_{(l,s) \in \mathbb{H}} \text{KL}(p : q_{l,s}) \) amount to \( \max_{(l,s) \in \mathbb{H}} \int p(x) \log q_{l,s}(x) \, d\mu(x) \).

**Theorem 3.** The minimum KL divergence \( \text{KL}(p_{l_1,s_1} : q_{l_1^*,s_1^*}) \) induced by the right-sided KL minimization of \( p_{l_1,s_1} \) with \( Q \) is independent of the location-scale query parameter \((l_1, s_1)\). Similarly, the KL divergence \( \text{KL}(p_{l_2,s_2}^* : q_{l_2,s_2}) \) induced by the left-sided KL minimization of \( q_{l_2,s_2} \) with \( P \) is independent of the location-scale query parameter \((l_2, s_2)\).

Notice that in general \( \text{KL}(p_{l_1,s_1} : q_{l_1^*,s_1^*}) \neq \text{KL}(p_{l_2,s_2}^* : q_{l_2,s_2}) \). The theorem is a statement of the property mentioned without proof in [16].

The proof extends easily to \( f \)-divergences as follows:
**Theorem 4.** The f-divergence \( I_f(p_{l_1, s_1} : q_{l_1, s_1}^*) \) induced by the right-sided f-divergence minimization of \( p_{l_1, s_1} \) with \( Q \) is independent of \((l_1, s_1)\). Similarly, the f-divergence \( I_f(p_{l_2, s_2}^* : q_{l_2, s_2}) \) induced by the left-sided f-divergence minimization of \( q_{l_2, s_2} \) with \( P \) is independent of \((l_2, s_2)\).

**Proof.** Without loss of generality, consider the left-sided f-divergence minimization problem (right-sided density query). We have

\[
I_f(P : q_{l_2, s_2}) := \min_{(l_1, s_1) \in \mathbb{H}} I_f(p_{l_1, s_1} : q_{l_2, s_2}) = \min_{(l_1, s_1) \in \mathbb{H}} I_f\left(p \mid q_{l_2, s_2} \right).
\]

Let \( l = \frac{l_2 - l_1}{s_1} \) and \( s = \frac{s_2}{s_1} \). Then the minimization problem becomes:

\[
\min_{(l_1, s_1) \in \mathbb{H}} I_f(p_{l_1, s_1} : q_{l_2, s_2}) = \min_{(l, s) \in \mathbb{H}} I_f(p : q_{l, s}) := I_f(p : Q).
\]

Once the optimal parameter \( l^* \) and \( s^* \) have been calculated, we recover the density \( p_{l_1^*, s_1^*} \in P \) that minimizes \( I_f(p_{l_1, s_1} : q_{l_2, s_2}) \) as \( p_{l_1^*, s_1^*} \) with

\[
s_1^* = \frac{s_2}{s^*},
\]

\[
l_1^* = l_2 - l^* s_1^*.
\]

\( \square \)

Let us remark that these f-divergence minimization problems between a query density and a location-scale family can be interpreted as information projections [10] of a query density onto a location-scale manifold.

Let us rework the example originally reported in [10]: Consider \( p(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \) and \( q(x) = \exp(-x) \) be the standard density of the half-normal distribution and the standard density of the exponential distribution defined over the support \( \mathcal{X} = [0, \infty) \), respectively. We consider the scale families \( P = \{ p_{s_1}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) : s_1 > 0 \} \) and \( Q = \{ q_{s_2}(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) : s_2 > 0 \} \). Using symbolic computing detailed in Appendix A, we find that

\[
\text{KL}(p_{s_1} : q_{s_2}) = \frac{1}{2} \left( 2 \log \frac{s_2}{s_1} + \log \frac{2}{\pi} - 1 \right) + \sqrt{\frac{2}{\pi}} \frac{s_1}{s_2}.
\]

Let \( r = \frac{s_1}{s_2} \). Then \( \text{KL}(p_{s_1} : q_{s_2}) = \sqrt{\frac{2}{\pi}} r - \log r + \log \frac{\sqrt{2}}{\pi} - \frac{1}{2} \). That is, the KL between the scale families depends only on the scale ratio as proved earlier.

We KL divergence is minimized when \( -\frac{1}{r} + \sqrt{\frac{2}{\pi}} = 0 \). That is, when \( r = \sqrt{\frac{\pi}{2}} \). We find that \( \text{KL}(p_{s_1} : Q) = \frac{1}{2} + \log \frac{2}{\pi} \) is independent of \( s_1 \), as expected.

## 5 Concluding remarks

The canonical structure of the densities of the location-scale families make it possible to get various identities for the cross-entropy, the differential entropy, and the Kullback-Leibler divergence, by making change of variables in the corresponding integrals. In particular, the Kullback-Leibler divergence (or more generally any f-divergence) between location-scale densities can be reduced
to the calculation of the Kullback-Leibler divergence between one standard density with another transformed location-scale density. It follows that the Kullback-Leibler divergence between scale densities depends only on the scale ratio. We illustrated our approach by computing the Kullback-Leibler divergence between scale Cauchy distributions which is symmetric. More generally, we reported a condition on the standard density of a location-scale family which yields symmetric Kullback-Leibler divergences. We then proved that the minimum f-divergence between a query density of a location-scale family with any member of another location-scale family does not depend on the query location-scale parameters. To conclude, let us mention that we can derive similar identities for information-theoretic measures from change of variables in integrals for location-dispersion families [14].

A Symbolic calculation using MAXIMA

We use the computer algebra system MAXIMA to calculate Eq. 59

```plaintext
pe(x) := sqrt(2/%pi)*exp(-x*x/(2.0));
qe(x) := exp(-x);
p(x,s1) := (1/s1)*pe(x/s1);
q(x,s2) := (1/s2)*qe(x/s2);
assume(s1>0);
assume(s2>0);
integrate(p(x,s1)*log(p(x,s1)/q(x,s2)),x,0,inf);
ratsimp(%);
```

References

[1] S. Amari. Information Geometry and Its Applications. Applied Mathematical Sciences. Springer Japan, 2016.

[2] Jean-Daniel Boissonnat, Frank Nielsen, and Richard Nock. Bregman Voronoi diagrams. Discrete & Computational Geometry, 44(2):281–307, 2010.

[3] Frédéric Chyzak and Frank Nielsen. A closed-form formula for the Kullback-Leibler divergence between Cauchy distributions. arXiv preprint arXiv:1905.10965, 2019.

[4] Thomas M Cover and Joy A Thomas. Elements of information theory. John Wiley & Sons, 2012.

[5] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizit von markoffschen ketten. Magyar. Tud. Akad. Mat. Kutató Int. Közl, 8:85–108, 1963.

[6] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural computation, 21(3):793–830, 2009.

[^1]: http://maxima.sourceforge.net/
[7] John D Lafferty. The density manifold and configuration space quantization. *Transactions of the American Mathematical Society*, 305(2):699–741, 1988.

[8] Joseph Victor Michalowicz, Jonathan M Nichols, and Frank Bucholtz. *Handbook of differential entropy*. CRC Press, 2013.

[9] M. K. Murray and J. W. Rice. *Differential Geometry and Statistics*, volume 48. CRC Press, 1993.

[10] Frank Nielsen. What is an information projection? * Notices of the AMS*, 65(3):321–324.

[11] Frank Nielsen and Vincent Garcia. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

[12] Frank Nielsen and Richard Nock. Entropies and cross-entropies of exponential families. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 3621–3624. IEEE, 2010.

[13] Frank Nielsen and Richard Nock. On the chi square and higher-order chi distances for approximating $f$-divergences. *IEEE Signal Processing Letters*, 21(1):10–13, 2013.

[14] Toshio Ohnishi and Takemi Yanagimoto. Conjugate location-dispersion families. *Journal of the Japan Statistical Society*, 37(2):307–325, 2007.

[15] George Tzagkarakis and Panagiotis Tsakalides. A statistical approach to texture image retrieval via alpha-stable modeling of wavelet decompositions. In *International Workshop on Image Analysis for Multimedia Interactive Services*, pages 21–23, 2004.

[16] Cristiano Villa. A Property of the Kullback-Leibler Divergence for Location-scale Models. *ArXiv e-prints*, April 2016.