fractal and fractional

MDPI

Article

Fractional Divergence of Probability Densities

Aris Alexopoulos

P.O. Box 123AA, Adelaide, SA 5000, Australia; arisa@iinet.net.au

Received: 12 September 2017; Accepted: 19 October 2017; Published: 25 October 2017

Abstract: The divergence or relative entropy between probability densities is examined. Solutions that minimise the divergence between two distributions are usually "trivial" or unique. By using a fractional-order formulation for the divergence with respect to the parameters, the distance between probability densities can be minimised so that multiple non-trivial solutions can be obtained. As a result, the fractional divergence approach reduces the divergence to zero even when this is not possible via the conventional method. This allows replacement of a more complicated probability density with one that has a simpler mathematical form for more general cases.

Keywords: divergence; fractional divergence; probability densities

1. Introduction

The divergence or relative entropy between two probability densities is a measure of dissimilarity between them. The most well known divergence approach is due to Kullback and Leibler which will be discussed in more detail below. Other divergence formulations include the version by Jeffrey which is symmetric for large separations between densities [1]. While Jeffrey's approach is a symmetric f-divergence and is non-negative, it does not obey the triangle inequality. The Jensen–Shannon divergence [2] is essentially a half times the sum of the two separate densities and their respective divergence to their mean. The striking feature about the Jensen–Shannon divergence is that its square root is a true distance metric. That is, it not only displays symmetry but also conforms to the triangle inequality. Closely related to the divergence or relative entropy is entropy itself. Some of the definitions are due to Renyi [3], a one-parameter generalisation of the Shannon entropy and the definition due to Tsallis [4]. Tsallis entropy is in fact the underlying formulation for many other entropy definitions in the literature.

There have been attempts to generalise the concepts of divergence and entropy using fractional calculus. Fractional order mathematics has been applied to many classical areas associated with probability, entropy and divergence. The entropy has been derived in fractional form in [5] and subsequently in [6]. Divergence measures based on the Shannon entropy have been dealt with in [7]. An interpretation of fractional order differentiation in the context of probability has been given in [8]. The role of fractional calculus in probability has been discussed in [9]. In [10], the connection of fractional derivatives and negative probability densities is discussed. One of the first attempts to involve fractional calculus with probability theory is due to Jumarie [11]. The fractional probability measure is discussed, in particular the uniform probability density of fractional order. Comparison of the properties of fractional probabilities to the properties of classical probability theory have been studied in [12–14]. These latter works extend the ideas of Jumarie and give definitions for fractional probability space and fractional probability measure so that a fractional analogue of the classical probability theory is obtained.

The underlying mathematical construct in all of these approaches is the dependence on probability densities or distributions. In many areas of research, there is a requirement to model the statistical behaviour of a physical process by using probability distributions in terms of the cumulative distribution function (CDF) or the probability density function (PDF). Depending on the problem

to be analysed, there is usually a particular distribution that is better suited for the description of the physical process compared to other distributions. The problem is that most distributions contain multiple parameters that must be estimated using such methods as the maximum likelihood approach or method of moments. The estimation of these parameters introduces uncertainty, which translates to performance loss for a particular distribution when used to model physical phenomena. For example, in the detection of signals using the Constant False Alarm Rate (CFAR) approach [15–19], correct estimation of parameters is critical. The estimation of these parameters is almost always not exact and, as a consequence, the detection performance drops because of the loss in accuracy.

The basic requirement is to find a probability density that describes a physical process accurately while possessing a smaller number of parameters. In other words, is there a simpler probability density that can replace a more complicated two or more parameter version? This means that the simpler expression must match the performance of the latter very well for a large solution set. The use of a separation metric is required that will indicate how dissimilar they are. If the separation between them is zero or very close to zero, then the more complicated density can be replaced by the "simpler" density (or approximation). Much work has been done on this problem and two methods have proven to be very useful. The first involves information geometry [20] where the separation is given by the geodesic distance between two probability density-manifolds. The geodesic is obtained via the Fisher–Rao information metric. The geodesic is a true metric because it is symmetric between the densities and obeys the triangle inequality.

The other approach is to consider a class of divergence formulations called f-divergences of which the Kullback–Leibler version belongs to. The Kullback–Leibler divergence is not symmetric for large separations between densities and does not obey the triangle inequality. However, there are a number of ways to make it symmetric for large separations between densities. It is worth noting that there is a mathematical duality between the Kullback–Leibler divergence and the geodesic approach of information geometry. In addition, the latter is more complicated to work with in the mathematical sense because, in many cases, the geodesic must be obtained via the solution of partial differential equations. On the other hand, an f-divergence formulation such as the Kullback–Leibler divergence is relatively easier to implement, requiring the solutions to be obtained via integrals instead.

The Kullback–Leibler divergence has been used previously to find solutions that allow one density or model to be replaced by another [21–29]. The problem is that the solution sets that give a divergence of zero or close to zero are either unique or trivial in nature. That is, the divergence is not valid for a large set of parameter values. Replacing one model (density) by another only for certain unique or restricted values in their parameters is not very useful for modelling physical processes or systems. Unfortunately, this is the inherent problem associated with the current form of any divergence method. What is required is an approach that extends the solutions, where the divergence is close to zero or zero, beyond the unique and trivial cases. It would then be possible to replace one model with another since there would be a similarity between them for large parameter sets. This idea will be pursued in this paper by making use of fractional calculus to obtain a fractional form for the Kullback–Leibler divergence.

2. Divergence between Two Probability Densities

The divergence between two probability densities considered here is based on the Kullback–Leibler formulation (K-L). This is a pseudo-metric for the distance between the densities because it fails the triangle inequality. The main issue with the K-L formulation is that it is not symmetric unless the metric separation between the densities is small, i.e., probability density $q(x;\vec{\xi}_2)$ is very close in parameter space to density $p(x;\vec{\xi}_1)$: $q(x;\vec{\xi}_2) \approx p(x,\vec{\xi}_1+\delta\vec{\xi}_1)$, where $\vec{\xi}_i$ is the parameter

space of each density $\vec{\xi}_i = (\xi_1, \xi_2, ..., \xi_N)$ and N represents the total number of parameters. The K-L divergence is defined as

$$D(p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) = \int_\Omega p(x;\vec{\xi}_1)\log\left(\frac{p(x;\vec{\xi}_1)}{q(x;\vec{\xi}_2)}\right)dx \tag{1}$$

for some region of integration Ω. It is possible to obtain a symmetric version of (1) that is valid for larger separations and obeys the triangle inequality. One way to do this is using the Jeffrey's formulation as discussed previously:

$$D(p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) = \int_\Omega \left[p(x;\vec{\xi}_1) - q(x;\vec{\xi}_2)\right]\log\left(\frac{p(x;\vec{\xi}_1)}{q(x;\vec{\xi}_2)}\right)dx. \tag{2}$$

It will suffice to consider the divergence as given by (1) in what follows since the approach discussed in this paper is easily applicable to the symmetric Jeffrey's case or other similar formulations. Either way, this does not matter much, since, for almost all cases of interest, small separations dominate. The K–L divergence (1), hereby referred to as the divergence for brevity, is also known as the relative entropy for the following reason. If (1) is re-written as

$$D(p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) = \int_\Omega p(x;\vec{\xi}_1)\log\Big(p(x;\vec{\xi}_1)\Big)dx - \int_\Omega p(x;\vec{\xi}_1)\log\Big(q(x;\vec{\xi}_2)\Big)dx, \tag{3}$$

then the negative of the first integral in (3) is the differential entropy H of the probability density $p(x;\vec{\xi}_1)$. It was first used in statistical physics by Boltzmann and in information theory by Shannon. Both considered the discreet form for a probability mass $p(x_i;\vec{\xi}_1)$

$$H(p(x_i;\vec{\xi}_1)) = -\sum_{i=1}^{N} p(x_i;\vec{\xi}_1)\log\Big(p(x_i;\vec{\xi}_1)\Big). \tag{4}$$

The integral with a positive sign on the right of (3) is the cross-entropy between the densities $p(x;\vec{\xi}_1)$ and $q(x;\vec{\xi}_2)$. Hence, (1) and (3) are also referred to as the relative entropy between two densities. The divergence or relative entropy between probability densities $p(x;\vec{\xi}_1)$ and $q(x;\vec{\xi}_2)$ are interpreted in the following sense. Assume that a physical process or system is known to be accurately represented and modelled by a probability density $p(x;\vec{\xi}_1)$. This density might also represent an ideal or theoretical model. Is there another (perhaps simpler) model with density $q(x;\vec{\xi}_2)$ that is asymptotically close or exact with the former density (model)? If the two densities have a divergence that tends to zero, then the more complicated model can be replaced by the simpler model (approximation) for the given parameters that achieve zero or almost zero divergence. In another sense, the way to understand this is to ask what information is lost if one used the model density $q(x;\vec{\xi}_2)$ compared to the more accurate model density $p(x;\vec{\xi}_1)$. As an example, use of divergence in signal processing is very important—in particular, the detection of targets amongst background noise and clutter. This requires determining if signals (targets) of a given probability density differ from another density that represents the background noise and clutter. The degree of separation above a given threshold determines whether targets are present or not (see Section 6). In fact, the concept of divergence is used in many areas of physics, statistics/mathematics and engineering with a common goal. Ideally, the requirement is to find solutions to (1) in terms of the parameter vectors $\vec{\xi}_1$ and $\vec{\xi}_2$ that make the divergence equal to zero, i.e.,

$$\int_\Omega p(x;\vec{\xi}_1)\log\left(\frac{p(x;\vec{\xi}_1)}{q(x;\vec{\xi}_2)}\right)dx = 0, \tag{5}$$

or from (3) when the entropy term is equal to the cross entropy term. The problem is, since the two densities are different and with different parameters, it is not possible to achieve zero divergence between them except perhaps for particular or unique solutions such as solutions pertaining to their intersections. In some cases, the solutions are trivial such as when the two densities are of the exact mathematical form, which means a divergence of zero is possible since the parameters of one can be made to take on the same values as those of the other. For example, for two Exponential densities with parameters λ_1 and λ_2, it is trivial to show by inspection or by using the divergence (1) that

$$p(x;\lambda_1) = \lambda_1 e^{-\lambda_1 x} \qquad \text{and} \qquad q(x;\lambda_2) = \lambda_2 e^{-\lambda_2 x} \tag{6}$$

have a divergence of zero *everywhere* only when $\lambda_1 \equiv \lambda_2$. In fact, forcing the divergence to be zero as in (5) may not necessarily give solutions that achieve zero divergence. In such cases, it is also possible that the solutions become complex, which does not make sense when applied to a real physical problem. In what follows, it will be shown that it is possible to extend the domain of validity of solutions that give zero divergence beyond the trivial or unique cases. This can be done via the transformation of one or more of the parameters appearing in the divergence equations using fractional calculus. The method will be applied to two important densities used in many fields of research: the Exponential density and a well known power form, namely, the Pareto density. The first step is to obtain the conventional and fractional divergences for the Exponential-Pareto case and then to do the same for the Exponential–Exponential case.

3. Conventional Divergence of Exponential and Pareto Densities

The Exponential and Pareto distributions have been used to model a large number of problems. For example, the Pareto distribution is critical in the analysis of radar clutter. For this reason, a fractional-order Pareto distribution has been presented in [30] in order to more accurately model sea clutter in microwave radar. Consider *i.i.d.* random variables belonging to the Exponential density $X_i \sim Exp(\lambda)$ as well as the Pareto density $X_i \sim Pa(x_0, \beta)$. That is,

$$p(x;\lambda) = \lambda e^{-\lambda x}, \tag{7}$$

where the parameter space contains only one parameter, λ, which is usually related to the expectation μ of the random variables by $\lambda = 1/\mu$. The Pareto density has parameter space $\vec{\xi}_2 = (x_0, \beta)$ where x_0 is the scale parameter and β is the shape parameter:

$$q(x;x_0,\beta) = \beta x_0^{\beta} x^{-(\beta+1)}. \tag{8}$$

The idea here is to replace the two parameter Pareto density with the one parameter and simpler Exponential density. On this basis, this can only be true for certain solutions where the divergence between them is zero or close to zero. For brevity, the densities will be written as $p(x)$ and $q(x)$. The divergence expression between the Exponential and Pareto densities is obtained from

$$D(p(x)||q(x)) = \int_{\Omega} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx, \tag{9}$$

where the log-function in the integrand is simplified to

$$\log\left(\frac{p(x)}{q(x)}\right) = \log\left(\frac{\lambda}{\beta x_0^{\beta}}\right) - \lambda x + (\beta+1)\log(x). \tag{10}$$

Substituting into (9) and taking the integration domain to be the interval $\Omega = [0, \infty)$, the divergence now becomes,

$$D(p(x)||q(x)) = \int_0^\infty \lambda e^{-\lambda x} \left[\log\left(\frac{\lambda}{\beta x_0^\beta}\right) - \lambda x + (\beta+1)\log(x) \right] dx. \tag{11}$$

The first term in the integrand is trivial since the second axiom of probability states that the integral of the density in the interval is unity: $\int_0^\infty p(x)dx = 1$. The other terms can be completed by using integration by parts to finally arrive at the following expression for the divergence between the two densities

$$D(p(x)||q(x)) = \left| \log\left(\frac{\lambda}{\beta x_0^\beta}\right) - (\beta+1)\log(\lambda) - \gamma(\beta+1) - 1 \right|, \tag{12}$$

where the Euler-gamma has been introduced and its value is $\gamma \approx 0.577216$. The modulus is included in (12) to enforce the fact that the divergence is greater or equal to zero. The idea now is to work out for what values of β in (12) the divergence approaches zero. That is, what values of β make the Pareto density $q(x)$ be approximate to or become equal to the Exponential $p(x)$ respectively? Let the parameter space of all parameters be written as a vector $\vec{\xi} \equiv (\xi_1, \xi_2, \xi_3) = (\lambda, x_0, \beta)$. Consider the derivative as an operator $\hat{L}_i = \partial/\partial\xi_i$. Taking the index $i = 3$ gives the operator in terms of the parameter β, i.e., $\hat{L}_3$. Using the operator on the left and right of (12) gives, (ignoring the modulus):

$$\hat{L}_3 D(p(x)||q(x)) = \hat{L}_3 \log\left(\frac{\lambda}{\beta x_0^\beta}\right) - \hat{L}_3(\beta+1)\log(\lambda) - \hat{L}_3\gamma(\beta+1) - \hat{L}_3, \tag{13}$$

where $\hat{L}_3 = \partial/\partial\xi_3 = \partial/\partial\beta$. We enforce the need for the left-hand side to be equal to zero as required, i.e., $\hat{L}_3 D(p(x)||q(x)) = 0$ so that

$$-\frac{1}{\beta} - (\gamma + \log(\lambda x_0)) = 0, \tag{14}$$

Solving for β:

$$\beta = -\frac{1}{\gamma + \log(\lambda x_0)}. \tag{15}$$

This means that the density $q(x)$ has a divergence that is zero or close to zero with respect to $p(x)$, the Exponential, whenever β is given by (15). Then, the Pareto density is modified to

$$q(x) = -\frac{x_0^{-\frac{1}{\gamma+\log(\lambda x_0)}}}{\gamma + \log(\lambda x_0)} x^{\frac{1}{\gamma+\log(\lambda x_0)}-1}. \tag{16}$$

The Pareto density (16) is now expressed in terms of the Exponential-density parameter λ. This indicates where the divergence of $p(x)$ from $q(x)$ is approaching zero as a function of λ. When the divergence is acceptably small or even zero, the Pareto model can be adequately described by the simpler one-parameter Exponential model. Thus, substituting (15) into (12) means that the divergence can be written as:

$$D(p(x)||q(x)) = \left| \log\left(-\frac{\lambda x_0^\omega}{\omega}\right) + (\omega-1)\log(\lambda) + \gamma(\omega-1) - 1 \right|, \tag{17}$$

where $\omega = (\gamma + \log(\lambda x_0))^{-1}$. Equation (17) determines the value of the minimum-divergence between the two densities. Figure 1 shows a plot of the divergence (17) as a function of the parameter λ

at $x_0 = 0.01$. The conventional divergence (17) is zero for the unique value of $\lambda \approx 9.458$ in the range considered. Multiple solutions that approach zero are not generally possible. In this case, the divergence between the densities $p(x)$ and $q(x)$ is zero only for the particular value $\lambda \approx 9.458$ and close to zero for small values of λ on either side.

In fact, for the general case, the conventional divergence given by the expression (12) can be plotted as a function of the parameters $(\lambda; x_0, \beta)$. Figure 2a shows the divergence between the two densities as a function of the two parameters λ and β at a fixed Pareto scale parameter value of $x_0 = 0.01$. For the range of λ and β values shown, the divergence is never zero or very close to zero. In terms of Figure 1, the divergence is zero for $\lambda \approx 9.458$ and this occurs when $\beta = 0.561$, which is outside the range of values for β shown in Figure 2a. An exact divergence of zero at only one unique point is not very useful or practical in the general sense anyway. What is required is an extension of the solutions so that zero divergence (or very close to zero) is achieved over a wider parameter range (see Figure 2b). This will require the use of fractional-order calculus and will be discussed in the next section.

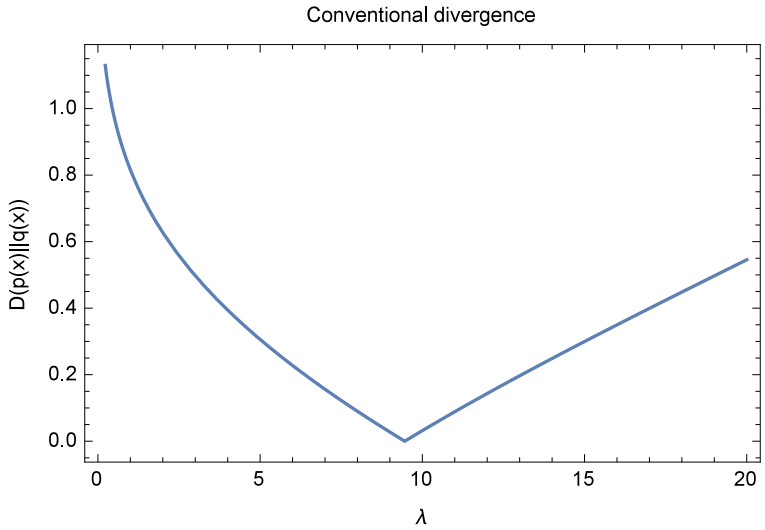
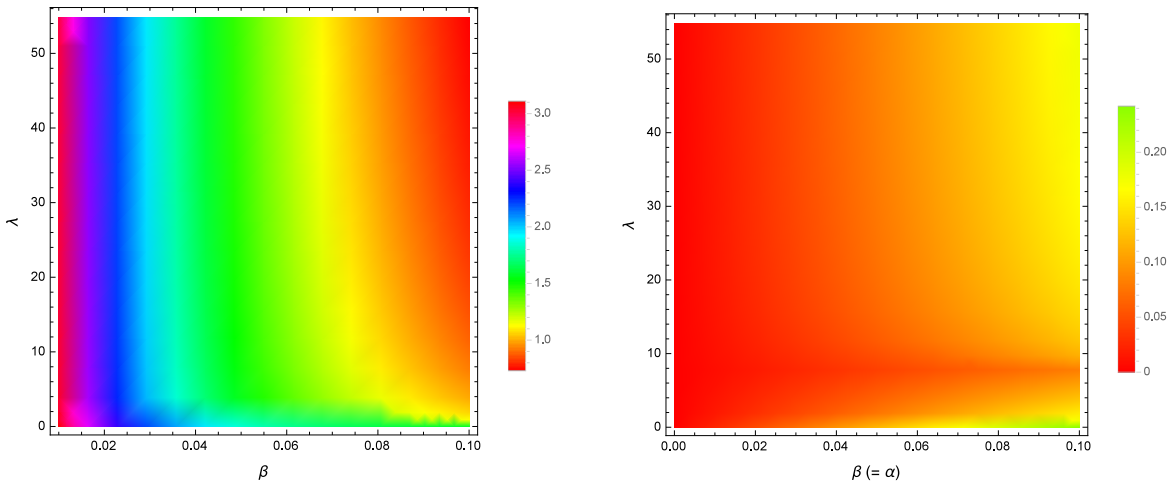


Figure 1. The divergence between the Exponential-density and the Pareto-density for a fixed Pareto scale parameter, $x_0 = 0.01$.



(**a**) Conventional divergence (**b**) Fractional divergence

Figure 2. For $x_0 = 0.01$, the (**a**) conventional and (**b**) fractional divergence is shown, respectively.

4. Fractional Divergence of Exponential and Pareto Densities

Fractional calculus has been around since the time of integer order calculus, which was developed by Newton and Leibniz. The name "fractional" is a misnomer that has endured since around 1695 when l' Hopital queried Leibniz on the meaning of a fractional order of one-half for the derivative operator. It is to be understood that fractional really means "generalised". Fractional-order derivatives and integrals of functions have been studied for a very long time with various definitions appearing in the literature. Among the well known are due to Caputo, Grunwald-Letnikov and Riemann–Liouville. For a comprehensive review of the many versions that have been derived, see [31] and the references therein. Research into fractional order mathematics has been prevalent in recent times in many fields of science, mathematics and engineering [32–41]. In this paper, the interest is in the fractional derivative of functions only and the Riemann–Liouville formulation for the fractional derivative will be considered:

$$ {}_aD_t^{\alpha} f(t) = \frac{1}{\Gamma(\nu-\alpha)} \frac{d^{\nu}}{dt^{\nu}} \int_a^t (t-x)^{\nu-\alpha-1} f(x) dx. \tag{18} $$

The terminal a takes two values. The case $a = -\infty$ is due to Liouville while the case $a = 0$ is due to Riemann. The parameter ν represents values that are integer order, i.e., $\nu \in Z^+$. The parameter α is the fractional order that can be real or complex and is bounded by $\lfloor \alpha \rfloor < \alpha \leqslant \lceil \alpha \rceil$. Here, $\lfloor \cdot \rfloor$ is the floor function and $\lceil \cdot \rceil$ is the ceiling function, respectively. Consider the Riemann–Liouville fractional derivative for $\nu = 1$ and terminal $a = 0$. The following fractional operator can then be defined:

$$ \hat{\Lambda}_i(x \mapsto \xi_i) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x d\xi_i (x-\xi_i)^{-\alpha}. \tag{19} $$

Applying the operator on the conventional divergence, i.e., $\hat{\Lambda}_i(x \mapsto \xi_i) D(p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2))$, gives the fractional divergence $\mathcal{D}(x \mapsto \xi_i, p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2))$ such that the following holds:

Definition 1. *The fractional divergence, which is a generalisation of the conventional divergence, is defined as*

$$ \begin{aligned} \mathcal{D}(x \mapsto \xi_i, p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) &= \left| \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x (x-\xi_i)^{-\alpha} \int_{\Omega} p(x;\vec{\xi}_1) \log\left(\frac{p(x;\vec{\xi}_1)}{q(x;\vec{\xi}_2)} \right) dx d\xi_i \right| \\ &= \left| \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x (x-\xi_i)^{-\alpha} \left\langle \log\left(\frac{p(x;\vec{\xi}_1)}{q(x;\vec{\xi}_2)} \right) \right\rangle_{p(x;\vec{\xi}_1)} d\xi_i \right|, \end{aligned} \tag{20} $$

where $< \cdot >$ is the expectation with respect to the density $p(x;\vec{\xi}_1)$ and the three axioms of probability theory hold for both densities. The modulus $|\cdot|$ is required because $\alpha \in \mathcal{R}$ and $\alpha \in \mathcal{C}$. In addition, the definition: $p(x)\log(p(x)/q(x)) = 0$ whenever $p(x) = 0$ is applicable.

Theorem 1. *If the fractional divergence is a generalised form for the divergence between two densities, it must produce the same solutions as the conventional divergence as a special limit. The latter is true when the fractional order approaches $\alpha = 1$ in (20). Thus,*

$$ \lim_{\alpha \to 1} \mathcal{D}(x \mapsto \xi_i, p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) = L_i(\xi_i) D(p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) \tag{21} $$

or in operator form:

$$ \lim_{\alpha \to 1} \hat{\mathcal{D}}(x \mapsto \xi_i) = \hat{L}_i(\xi_i). \tag{22} $$

Proof. The proof involves showing that the fractional operator of order, α, reduces to the ν-th integer order derivative in the limit $\alpha \to \nu$. The final stage requires setting $\nu = 1$ to complete the proof. Let

$\nu \in \mathcal{N}$ be an arbitrary integer order of the conventional derivative. Let the fractional order operator be written in terms of the integer order derivative ν,

$$\lim_{\alpha\to\nu} \hat{\mathcal{D}}(x \mapsto \xi_i) = \lim_{\alpha\to\nu} \frac{1}{\Gamma(\nu-\alpha)} \frac{d^\nu}{dx^\nu} \int_0^x (x-\xi_i)^{\nu-\alpha-1} d\xi_i \to \lim_{\alpha\to\nu} \frac{1}{\Gamma(\nu-\alpha)} \frac{d^\nu}{dx^\nu} \int_0^x y_i^{\nu-\alpha-1} dy_i, \tag{23}$$

where the expression on the right is obtained by using the transformation $y_i = x - \xi_i$. Then,

$$\begin{aligned} \lim_{\alpha\to\nu} \hat{\mathcal{D}}(x \mapsto \xi_i) &= \lim_{\alpha\to\nu} \frac{1}{\Gamma(\nu-\alpha)} \frac{d^\nu}{dx^\nu} \int_0^x y_i^{\nu-\alpha-1} dy_i \\ &= \lim_{\alpha\to\nu} \frac{1}{\Gamma(\nu-\alpha)} \frac{d^\nu}{dx^\nu} \left[\frac{y_i^{\nu-\alpha}}{\nu-\alpha} \right]_0^x \\ &= \lim_{\alpha\to\nu} \frac{d^\nu}{dx^\nu} \left[\frac{x_i^{\nu-\alpha}}{\Gamma(\nu+1-\alpha)} \right] \\ &= \frac{d^\nu}{dx^\nu}. \end{aligned} \tag{24}$$

The conventional divergence corresponds to the integer order $\nu = 1$, hence

$$\begin{aligned} \lim_{\alpha\to 1} \hat{\mathcal{D}}(\xi_i) &= \frac{d}{d\xi_i} \\ &= \hat{L}_i(\xi_i) \end{aligned} \tag{25}$$

as required. Note that the mapping $(x \mapsto \xi_i)$ has been applied in (25). □

The divergence integral appearing in the integrand of (20), i.e., the expectation, has already been calculated before (see (12)). The parameter vector space for both densities is $\vec{\xi} = (\vec{\xi}_1, \vec{\xi}_2) = (\lambda, x_0, \beta)$. Re-arranging the divergence expression (12), the following form is obtained, (neglecting the modulus until the end):

$$D(p(x)||q(x)) = -\log(\beta) - \omega^{-1}\beta - (\gamma+1), \tag{26}$$

where $\omega^{-1} = \gamma + \log(\lambda x_0)$ and $(\vec{\xi}_1; \vec{\xi}_2)$ have been omitted for brevity. The requirement now is to use the operator and calculate the fractional divergence as follows:

$$\mathcal{D}(x \mapsto \xi_i, p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) = -\hat{\Lambda}_i(x \mapsto \xi_i)\log(\beta) - \omega^{-1}\hat{\Lambda}_i(x \mapsto \xi_i)\beta - (\gamma+1)\hat{\Lambda}_i(x \mapsto \xi_i). \tag{27}$$

The argument $(x \mapsto \xi_i)$ implies that the variable x maps on to the variable ξ_i. This will be elucidated further in what follows below. Recall that the parameter vector is given by $\vec{\xi} = (\lambda, x_0, \beta)$ and as before, in Section 3, the interest is in the parameter β, i.e., $i = 3$ so that $\xi_3 = \beta$. In addition, the condition $\mathcal{D}(x \mapsto \xi_i, p(x;\vec{\xi}_1)||q(x;\vec{\xi}_2)) = 0$ is enforced so that (27) becomes

$$\hat{\Lambda}_3(x \mapsto \beta)\log(\beta) + \omega^{-1}\hat{\Lambda}_3(x \mapsto \beta)\beta + (\gamma+1)\hat{\Lambda}_3(x \mapsto \beta) = 0. \tag{28}$$

Each term appearing in (28) will now be calculated. Before proceeding, it is important to re-visit the meaning of the mapping $(x \mapsto \xi_i)$. Once the operator $\hat{\Lambda}_i$ is used, the final result is a function of the variable x, which must then be replaced by the variable ξ_i, i.e, $\hat{\Lambda}_i(x \mapsto \xi_i) \to \hat{\Lambda}_i(\xi_i)$. The first term in (28) will be calculated last as it is more involved than the other two. In addition, the function $\log(z)$, for some argument z, always appears in these kinds of problems involving divergence or parameter estimation, and, for this reason, it will be treated in full. The other two terms contain monomials β^1 and $\beta^0 = 1$. It can be shown, by using the Riemann–Liouville fractional formulation, that the

fractional derivative of monomials with power n results in a form that is the exact version of Euler's generalisation of the integer derivatives of monomials:

$$\frac{d^{\alpha}}{d\beta^{\alpha}}\beta^{n} = \frac{\Gamma(n+1)}{\Gamma(n+1-\alpha)}\beta^{n-\alpha} \tag{29}$$

for monomial powers n. To verify this, the second term is (leaving out the coefficient):

$$\hat{\Lambda}_3(x \mapsto \beta)\beta = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^x (x-\beta)^{-\alpha}\beta d\beta. \tag{30}$$

Let the above integral be transformed to the form

$$\begin{aligned}\hat{\Lambda}_3(x \mapsto \beta)\beta &= \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^x y^{-\alpha}(x-y)dy \\ &= \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\left[\frac{x^{2-\alpha}}{(1-\alpha)} - \frac{x^{2-\alpha}}{(2-\alpha)}\right] \\ &= \frac{x^{1-\alpha}}{(1-\alpha)\Gamma(1-\alpha)}\end{aligned} \tag{31}$$

using the transformation $y = x - \beta$ and $dy = -d\beta$. The requirement now is to map the variable x such that $\hat{\Lambda}_3(x \mapsto \beta)\beta \to \hat{\Lambda}_3(\beta)\beta$ in (31) to obtain the final result

$$\hat{\Lambda}_3(\beta)\beta = \frac{\beta^{1-\alpha}}{\Gamma(2-\alpha)} \tag{32}$$

since $(1-\alpha)\Gamma(1-\alpha) \equiv \Gamma(2-\alpha)$. As stated above, this result is equivalent to that obtained by using Euler's form (29) for $n = 1$. In a similar way, the final term in (28) can be obtained as follows (leaving out the coefficient again),

$$\begin{aligned}\hat{\Lambda}_3(x \mapsto \beta) &= \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^x (x-\beta)^{-\alpha}d\beta \\ &= \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^x y^{-\alpha}dy \\ &= \frac{x^{-\alpha}}{\Gamma(1-\alpha)},\end{aligned} \tag{33}$$

where the transformation $y = x - \beta$ and $dy = -d\beta$ have been applied. The final result then becomes:

$$\hat{\Lambda}_3(\beta) = \frac{\beta^{-\alpha}}{\Gamma(1-\alpha)}. \tag{34}$$

Once again, this result can be obtained directly from the Euler Equation (29) for $n = 0$. The first term of (28) is now evaluated as follows:

$$\hat{\Lambda}_3(x \mapsto \beta)\log(\beta) = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^x (x-\beta)^{-\alpha}\log(\beta)d\beta. \tag{35}$$

To perform the integration in (35), let $y = x - \beta$ so that $dy = -d\beta$ and this gives

$$\int_0^x y^{-\alpha}\log(x-y)dy = \int_0^x y^{-\alpha}\log(x)dy + \int_0^x y^{-\alpha}\log(1-y/x)dy, \tag{36}$$

where $\log(x-y) \equiv \log(x(1-y/x)) = \log(x) + \log(1-y/x)$ has been used in (36) to expand the integrand. The first integral on the right of (36) is only dependent on the variable y so that it is trivial to show that

$$\int_0^x y^{-\alpha} \log(x) dy = \frac{\log(x)}{1-\alpha} x^{1-\alpha}. \tag{37}$$

The second integral in (36) can be solved if $z = y/x$ so that $dz = dy/x$ and the integral becomes,

$$\begin{aligned} \int_0^x y^{-\alpha} \log(1-y/x) dy &= x^{1-\alpha} \int_0^1 z^{-\alpha} \log(1-z) dz \\ &= \frac{x^{1-\alpha}}{\alpha - 1} H_{1-\alpha}. \end{aligned} \tag{38}$$

Here, $H_{1-\alpha}$ is the harmonic-function that is related to the polygamma-function of the zeroth order or digamma-function $\psi_0(\cdot)$ via $H_{1-\alpha} = \gamma + \psi_0(2-\alpha)$, where $\gamma \approx 0.577216$ is the Euler gamma constant. The digamma-function $\psi_0(2-\alpha)$ can be simplified further by using the identity:

$$\psi_n(z+1) = \psi_n(z) + (-1)^n \frac{n!}{z^{n+1}}. \tag{39}$$

Setting $z = 1-\alpha$ and $n = 0$ in the identity, one obtains $\psi_0(2-\alpha) = \psi_0(1-\alpha) + \frac{1}{1-\alpha}$. Hence, (38) can be re-written as:

$$\int_0^x y^{-\alpha} \log(1-y/x) dy = \frac{x^{1-\alpha}}{\alpha - 1} \left[\gamma + \psi_0(1-\alpha) + \frac{1}{1-\alpha} \right]. \tag{40}$$

Substituting (40) and (37) into (36), (35) becomes:

$$\hat{\Lambda}_3(x \mapsto \beta) \log(\beta) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \left[\frac{x^{1-\alpha} \log(x)}{1-\alpha} + \frac{x^{1-\alpha}}{\alpha - 1} \left(\gamma + \psi_0(1-\alpha) + \frac{1}{1-\alpha} \right) \right]. \tag{41}$$

After performing the simple differentiation in (41) and noting that $\hat{\Lambda}_3(x \mapsto \beta) \log(\beta) \to \hat{\Lambda}_3(\beta) \log(\beta)$, we have:

$$\hat{\Lambda}_3(\beta) \log(\beta) = \frac{\beta^{-\alpha}}{\Gamma(1-\alpha)} \left[\log(\beta) - \psi_0(1-\alpha) - \gamma \right]. \tag{42}$$

It is now a matter of substituting (42), (34) and (32) into (28) to obtain the final result:

$$\beta^{-\alpha} \left[\log(\beta) - \psi_0(1-\alpha) - \gamma \right] + \frac{\omega^{-1} \beta^{1-\alpha}}{(1-\alpha)} + (\gamma + 1) \beta^{-\alpha} = 0. \tag{43}$$

The problem now requires the solution of (43) in terms of the parameter β, which will be the fractional analogue of the conventional version as discussed in Section 3. Unfortunately, due to the fact that (43) is a transcendental equation in β, it means that solutions can only be obtained numerically. However it is possible to rewrite (43) in such a way as to obtain closed form analytic solutions. Equation (43) can be re-arranged to:

$$\beta = \omega(\alpha - 1) \log(\beta) + \omega(1-\alpha) \left[\psi_0(1-\alpha) - 1 \right]. \tag{44}$$

Define A and B as follows:

$$A = \omega(\alpha - 1) \quad \text{and} \quad B = \omega(1-\alpha) \left[\psi_0(1-\alpha) - 1 \right] \tag{45}$$

so that (44) becomes:

$$\beta = A\log(\beta) + B, \tag{46}$$

which allows the solution in terms of β to be in closed form if it can be transformed to resemble the Lambert W-function or product-log function. The W-function has the form

$$ye^{y} = f(x). \tag{47}$$

That is, if any equation can be written so that the left-hand side resembles the left-hand side of (47), then for any function on the right side, $f(x)$, the solution for y is given by: $y = W_n(f(x))$, where $n = 0, -1$ are the two branch cuts of the Lambert W-function. Equation (46) can now be solved via the W-function if it is transposed as follows:

$$\beta = \exp\left(\frac{\beta}{A} - \frac{B}{A}\right) \iff -\frac{\beta}{A}e^{-\frac{\beta}{A}} = -\frac{1}{A}e^{-\frac{B}{A}}. \tag{48}$$

Then, by (47), the solution for fractional β is obtained from the W-function as:

$$\beta = -AW_n\left(-\frac{1}{A}\exp\left(-\frac{B}{A}\right)\right). \tag{49}$$

Substituting both A and B while noting that $\omega = 1/(\gamma + \log(\lambda x_0))$ gives the fractional β as:

$$\beta = \frac{(1-\alpha)}{\gamma + \log(\lambda x_0)} W_0(\chi), \tag{50}$$

where the argument of the W-function, χ is

$$\chi = \frac{\gamma + \log(\lambda x_0)}{(1-\alpha)} \exp\left(\psi_0(1-\alpha) - 1\right), \tag{51}$$

and the $n = 0$ branch cut is considered for the W-function. The fractional form for the Pareto shape parameter, (50), can now be substituted into the conventional Pareto to obtain the fractional Pareto density (PDF) that minimizes the divergence with respect to the Exponential-density:

$$q(x) = \frac{(1-\alpha)}{\gamma + \log(\lambda x_0)} W_0(\chi)\, x_0^{\frac{(1-\alpha)}{\gamma+\log(\lambda x_0)}W_0(\chi)}\, x^{-\left(1+\frac{(1-\alpha)}{\gamma+\log(\lambda x_0)}W_0(\chi)\right)}. \tag{52}$$

This is the fractional analogue of (16). Equation (50) can be substituted into the divergence Equation (12) as was done for the conventional solution for $\beta = -\omega$ (see (15)). Thus, the fractional divergence becomes:

$$\mathcal{D}(p(x)||q(x)) = \left|\log\left(\frac{\gamma + \log(\lambda x_0)}{(1-\alpha)W_0(\chi)}\right) + (\alpha - 1)W_0(\chi) - (\gamma + 1)\right|. \tag{53}$$

The modulus $|\cdot|$ in (53) has been reinstated not only to ensure a divergence greater or equal to zero but also because the fractional order can take, not just real, but also complex values. The interesting aspect of the fractional order α appearing in (51) and (53) is that the fractional β now depends on α (see (50)). There is no reason why the fractional order α cannot be replaced by the variable β. This means of course that β takes on the same domain or range of values that α does so defining the correct

range is critical. In this instance, using (51) and (53) is essentially the same as using the following forms. Set $\alpha = \beta$ to obtain:

$$\chi = \frac{\gamma + \log(\lambda x_0)}{(1-\beta)} \exp\left(\psi_0(1-\beta) - 1\right) \tag{54}$$

and

$$\mathcal{D}(p(x)||q(x)) = \left| \log\left(\frac{\gamma + \log(\lambda x_0)}{(1-\beta)W_0(\chi)}\right) + (\beta - 1)W_0(\chi) - (\gamma + 1) \right|. \tag{55}$$

Thus, in keeping with the conventional divergence plot shown in Figure 2a, Figure 2b shows a plot of the fractional divergence (55) (or (53)) for the parameters λ and β. As can be seen from the color bars, the divergence is large for the conventional divergence. However, the fractional version shows not only much smaller divergence separations for various values of λ and β, but a large region where the divergence is everywhere equal to zero. It is worth noting that the *minimum* divergence achieved by the conventional divergence is $D \approx 0.75$, which is still much greater than the *maximum* fractional divergence of $\mathcal{D} \approx 0.16$.

5. Manipulation of the Divergence between Two Exponential Densities via the Fractional Orders

The fractional divergence between two Exponential-densities will be investigated in this section with the aim of showing that it gives non-trivial solutions and that it is possible to manipulate the divergence via the fractional order(s). There is a good reason for analysing two Exponential-densities as opposed to any other densities. Unlike the divergence solutions obtained for arbitrary densities, which are not entirely known, there is absolute certainty as to the expected divergence profile for the Exponential-densities. This is because, according to the conventional divergence, there is zero divergence whenever their parameters are equal. There are no other solutions that minimise the divergence for two Exponential-densities. Let

$$p(y;u) = ue^{-uy} \quad \text{and} \quad q(y;v) = ve^{-vy} \tag{56}$$

be two Exponential-densities. The two Exponential-densities (56) have one parameter each so that $\vec{\xi}_1 = \xi_1 = u$ and $\vec{\xi}_2 = \xi_2 = v$. This corresponds to $i = 1, 2$ respectively. Omitting the modulus for now, the expression for the fractional divergence becomes,

$$\mathcal{D}(x \mapsto \xi_i, p(y;\vec{\xi}_1)||q(y;\vec{\xi}_2)) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x \int_\Omega (x - \xi_i)^{-\alpha} p(y;\vec{\xi}_1) \log\left(\frac{p(y;\vec{\xi}_1)}{q(y;\vec{\xi}_2)}\right) dy d\xi_i \tag{57}$$

The following two equations are obtained from (57):

$$\mathcal{D}(x \mapsto u, p(y;u)||q(y;v)) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x \int_0^\infty (x - u)^{-\alpha} p(y;u) \log\left(\frac{p(y;u)}{q(y;v)}\right) dy du \tag{58}$$

when $i = 1$ and

$$\mathcal{D}(x \mapsto v, p(y;u)||q(y;v)) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x \int_0^\infty (x - v)^{-\alpha} p(y;u) \log\left(\frac{p(y;u)}{q(y;v)}\right) dy dv \tag{59}$$

when $i = 2$. The domain of integration for the two densities is $\Omega \in [0, \infty)$. The conventional divergence $D(p(y;u)||q(y;v))$ which is embedded in (58) and (59), is evaluated as follows:

$$\begin{aligned} D(p(y;u)||q(y;v)) &= \int_0^\infty p(y;u) \log\left(\frac{p(y;u)}{q(y;v)}\right) dy \\ &= \int_0^\infty u e^{-uy} \left[\log\left(\frac{u}{v}\right) + \log\left(e^{(v-u)y}\right)\right] dy \end{aligned} \tag{60}$$

The first terms in (60) is straightforward since the second axiom of probability applies, while the second term requires integration by parts. The conventional divergence between two Exponential-densities takes the form:

$$D(p(y;u)||q(y;v)) = \log\left(\frac{u}{v}\right) + \frac{v}{u} - 1 \tag{61}$$

Substituting (61) into (58) gives the following result:

$$\mathcal{D}(x \mapsto u, p(y;u)||q(y;v)) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dx} \int_0^x (x-u)^{-\alpha} \left[\log\left(\frac{u}{v}\right) + \frac{v}{u} - 1\right] du \tag{62}$$

Using the operator form and enforcing the condition $\mathcal{D}(x \mapsto u, p(y;u)||q(y;v)) = 0$ means that the fractional divergence with respect to parameter u becomes

$$\hat{\Lambda}_u(x \mapsto u) \log(u) - \hat{\Lambda}_u(x \mapsto u) \log(v) + \hat{\Lambda}_u(x \mapsto u) \left(\frac{v}{u}\right) - \hat{\Lambda}_u(x \mapsto u) = 0. \tag{63}$$

Applying the fractional operator on the function $\log(u)$ has been addressed in the previous section. The result here follows a similar process that gives:

$$\begin{aligned} \hat{\Lambda}_u(x \mapsto u) \log(u) &= \frac{x^{-\alpha}}{\Gamma(1-\alpha)} \left[\log(x) - \psi_0(1-\alpha) - \gamma\right] \rightarrow \\ \hat{\Lambda}_u(u) \log(u) &= \frac{u^{-\alpha}}{\Gamma(1-\alpha)} \left[\log(u) - \psi_0(1-\alpha) - \gamma\right]. \end{aligned} \tag{64}$$

Once again, $\psi_0(1-\alpha)$ is the digamma function and γ is the Euler constant. The next term is evaluated to give the result:

$$\begin{aligned} \hat{\Lambda}_u(x \mapsto u) \left[\log(v) + 1\right] &= \frac{x^{-\alpha}}{\Gamma(1-\alpha)} \left[\log(v) + 1\right] \rightarrow \\ \hat{\Lambda}_u(u) \left[\log(v) + 1\right] &= \frac{u^{-\alpha}}{\Gamma(1-\alpha)} \left[\log(v) + 1\right]. \end{aligned} \tag{65}$$

The final requirement is to evaluate the ratio v/u. Application of the fractional operator on this ratio gives the result:

$$\begin{aligned} \hat{\Lambda}_u(x \mapsto u) \left(\frac{v}{u}\right) &= (-1)^\alpha v \Gamma(\alpha+1) x^{-(\alpha+1)} \rightarrow \\ \hat{\Lambda}_u(u) \left(\frac{v}{u}\right) &= v e^{i\alpha\pi} \Gamma(\alpha+1) u^{-(\alpha+1)}. \end{aligned} \tag{66}$$

Substitution of the expressions (64)–(66) into (63) and rearranging results in the following:

$$u = -\frac{e^{-i\alpha\pi}}{v\Gamma(\alpha+1)\Gamma(1-\alpha)} \log(u) + \frac{\psi_0(1-\alpha) + \gamma + \log(v) + 1}{v\Gamma(\alpha+1)\Gamma(1-\alpha)e^{i\alpha\pi}}. \tag{67}$$

Equation (67) can only be solved numerically for u in its present form. However, as shown in the previous section, it can be transformed so that its solutions can be obtained analytically by using the Lambert W-function. Setting

$$\begin{aligned} A &= \frac{e^{-i\alpha\pi}}{v\Gamma(\alpha+1)\Gamma(1-\alpha)} \\ B &= \frac{\psi_0(1-\alpha)+\gamma+\log(v)+1}{v\Gamma(\alpha+1)\Gamma(1-\alpha)e^{i\alpha\pi}} \end{aligned} \tag{68}$$

requires the solution of u using the form

$$u = -A\log(u) + B. \tag{69}$$

Transforming this expression to a form that allows solution using the W-function finally gives (see previous section):

$$u = AW_0\left(\frac{\exp\left(\frac{B}{A}\right)}{A}\right). \tag{70}$$

The solution (70) is a function of the fractional order α as well as other parameters. The fractional order belonging to u will be distinguished from now on and will be defined as $\alpha = \alpha_1$. The same will be done later for the solution v, which will be a function of its own fractional order $\alpha = \alpha_2$. Hence, substituting (68) into (70), the final result becomes:

$$u = \frac{e^{-i\alpha_1\pi}}{v\Gamma(\alpha_1+1)\Gamma(1-\alpha_1)}W_0(\chi_1), \tag{71}$$

where the argument χ_1 in the W-function is given by,

$$\chi_1 = v\Gamma(\alpha_1+1)\Gamma(1-\alpha_1)\exp\left(i\alpha_1\pi + \psi_0(1-\alpha_1) + \gamma + \log(v) + 1\right). \tag{72}$$

The next step is to complete a similar process for the parameter v. Substitution of the conventional divergence (61) into (59) requires the solution of

$$\mathcal{D}(x\mapsto v, p(y;u)||q(y;v)) = \frac{1}{\Gamma(1-\alpha)}\frac{d}{dx}\int_0^x (x-v)^{-\alpha}\left[\log\left(\frac{u}{v}\right)+\frac{v}{u}-1\right]dv. \tag{73}$$

Using the operator formulation, and noting that $\mathcal{D}(x\mapsto v, p(y;u)||q(y;v)) = 0$, gives the expression:

$$\hat{\Lambda}_v(x\mapsto v)\left[\log(u)-1\right] - \hat{\Lambda}_v(x\mapsto v)\log(v) + \hat{\Lambda}_v(x\mapsto v)\left(\frac{v}{u}\right) = 0. \tag{74}$$

Each term is now evaluated beginning with the first term:

$$\begin{aligned} \hat{\Lambda}_v(x\mapsto v)\left[\log(u)-1\right] &= \frac{x^{-\alpha}}{\Gamma(1-\alpha)}\left[\log(u)-1\right] \\ \hat{\Lambda}_v(v)\left[\log(u)-1\right] &= \frac{v^{-\alpha}}{\Gamma(1-\alpha)}\left[\log(u)-1\right]. \end{aligned} \tag{75}$$

The next term involves the log-function, which has been treated before in detail. Following the same process gives:

$$\begin{aligned}\hat{\Lambda}_v(x \mapsto v)\log(v) &= \frac{x^{-\alpha}}{\Gamma(1-\alpha)}\left[\log(v) - \psi_0(1-\alpha) - \gamma\right] \\ \hat{\Lambda}_v(v)\log(v) &= \frac{v^{-\alpha}}{\Gamma(1-\alpha)}\left[\log(v) - \psi_0(1-\alpha) - \gamma\right],\end{aligned} \tag{76}$$

where once again $\psi_0(1-\alpha)$ is the digamma-function and γ is the Euler constant. The final term is evaluated to be:

$$\begin{aligned}\hat{\Lambda}_v(x \mapsto v)\left(\frac{v}{u}\right) &= \frac{x^{1-\alpha}}{u\Gamma(2-\alpha)} \\ \hat{\Lambda}_v(v)\left(\frac{v}{u}\right) &= \frac{v^{1-\alpha}}{u\Gamma(2-\alpha)}.\end{aligned} \tag{77}$$

It is now a matter of substituting (75)–(77) into (74). Rearranging the expression gives the following form:

$$v = u(1-\alpha)\log(v) - u(1-\alpha)\left[\psi_0(1-\alpha) + \gamma + \log(u) - 1\right]. \tag{78}$$

In order to solve this equation using the W-function, set

$$\begin{aligned}A &= u(1-\alpha) \\ B &= u(1-\alpha)\left[\psi_0(1-\alpha) + \gamma + \log(u) - 1\right].\end{aligned} \tag{79}$$

The required equation takes the form

$$v = A\log(v) - B. \tag{80}$$

Rearranging this equation into the form that allows a solution by the W-function finally gives

$$v = -AW_0\left(-\frac{\exp(B/A)}{A}\right). \tag{81}$$

As was done for the u-solution, the fractional order of v will be set to $\alpha = \alpha_2$ to distinguish it from α_1 belonging to the parameter u. With this in mind and substituting the definitions for A and B, namely (79), gives

$$v = u(\alpha_2 - 1)W_0(\chi_2), \tag{82}$$

where

$$\chi_2 = \frac{1}{u(\alpha_2-1)}\exp\left(\psi_0(1-\alpha_2) + \gamma + \log(u) - 1\right). \tag{83}$$

The conventional divergence can now be transformed to the fractional divergence between two Exponential-densities by substituting the fractional solutions (71)–(72) for u and (82)–(83) for v into (61) to obtain the final form:

$$\begin{aligned}&\mathcal{D}(p(y;u)||q(y;v)) = \\ &\left|\log\left(\frac{W_0(\chi_1)e^{-i\alpha_1\pi}}{uv(\alpha_2-1)\Gamma(\alpha_1+1)\Gamma(1-\alpha_1)W_0(\chi_2)}\right) + uv(\alpha_2-1)\Gamma(\alpha_1+1)\Gamma(1-\alpha_1)e^{i\alpha_1\pi}\frac{W_0(\chi_2)}{W_0(\chi_1)} - 1\right|,\end{aligned} \tag{84}$$

where the arguments χ_1 and χ_2 are given by:

$$\begin{aligned}\chi_1 &= v\Gamma(\alpha_1+1)\Gamma(1-\alpha_1)\exp\left(i\alpha_1\pi+\psi_0(1-\alpha_1)+\gamma+\log(v)+1\right),\\ \chi_2 &= \frac{1}{u(\alpha_2-1)}\exp\left(\psi_0(1-\alpha_2)+\gamma+\log(u)-1\right).\end{aligned} \tag{85}$$

The modulus is used because $\alpha_{1,2} \in \mathcal{R}$ as well as $\alpha_{1,2} \in \mathcal{C}$ as can be seen from (84) and (85). In Figure 3, the conventional divergence, which is exact with the fractional divergence when $\alpha = 1$ in the latter, is shown as a divergence manifold (top-left) with the line $u = v$ running down the middle where the divergence is zero.

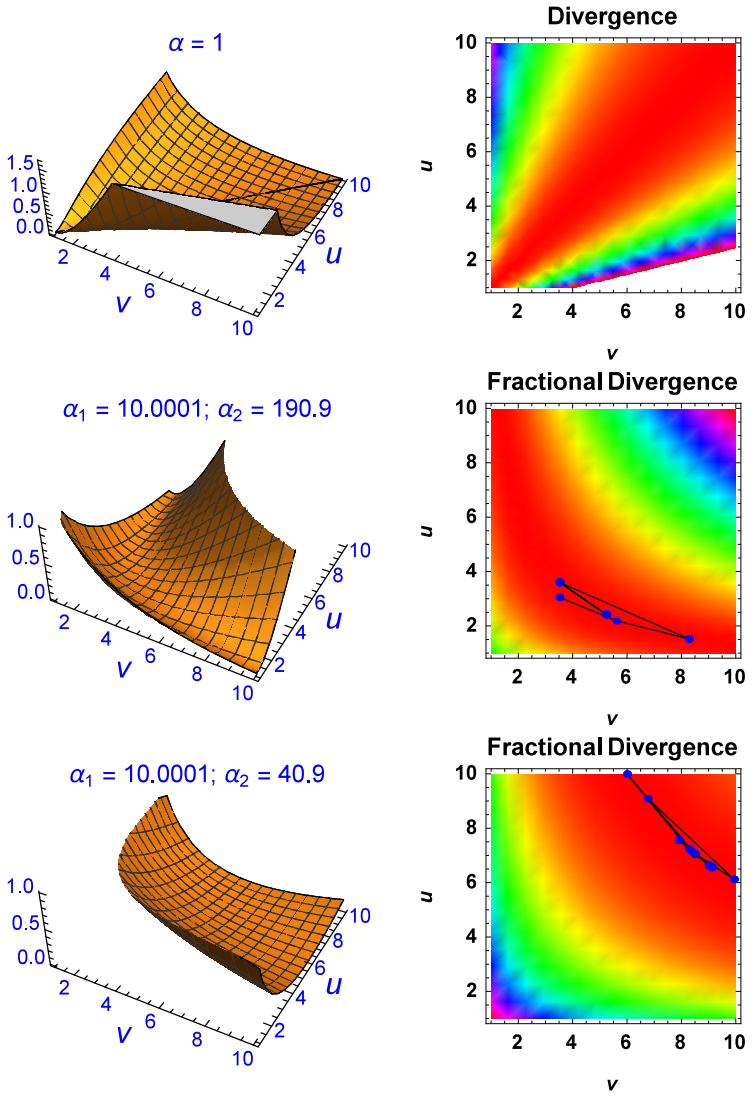


Figure 3. Variation of the (fractional) divergence manifold between two Exponential-densities in terms of the fractional orders α_1 and α_2. The case $\alpha = 1$ corresponds to the conventional divergence.

The conventional divergence is also shown on the right as an image map where the red region indicates small divergence on either side of the $u = v$ line (not shown). According to the conventional divergence between two Exponential-densities, the only solutions which give zero are those where $u = v$. However, as the middle two and last two plots indicate, the fractional divergence can make the divergence between them zero or close to zero for regions (solutions) where the conventional version fails. The middle two figures show manipulation of the divergence manifold for $\alpha_1 = 10.0001$ and $\alpha_2 = 190.9$

in which the divergence manifold has been minimised perpendicular to the conventional version ($\alpha = 1$). The image map on the right also contains iteration lines with each point being an iteration step in the process of finding the global minimum of the divergence using a differential-evolution numerical algorithm. This minimum occurs when $u = 2.32623$ and $v = 5.45556$ and at those parametric coordinates the fractional divergence is $\mathcal{D} = 10^{-8}$. The bottom two plots show further manipulation of the divergence manifold for $\alpha_2 = 40.9$, giving a fractional divergence of 10^{-8} for a global minimum in this case given by $u = 8.88506$ and $v = 6.73169$. The last four plots confirm that the fractional divergence approach can give essentially zero divergence for parameter values (u, v), which are not equal, unlike the expected results from the conventional divergence approach.

Further evidence of this can be seen in Figure 4. The manipulation of the divergence manifold is not only possible via $\alpha_{1,2} \in \mathcal{R}$ but also when α_1 and α_2 are complex (bottom-right plot). The fractional divergence has a global minimum for the complex solution of $\mathcal{D} = 10^{-14}$ at $(u, v) = (5.77781, 1.01829)$. There are numerous other non-trivial solutions with divergence of the order of 10^{-22} or less which have been omitted for brevity reasons. The results shown in Figures 3 and 4 indicate that the fractional divergence formulation makes it possible to find parameter values (u, v) that achieve zero divergence even when the conventional approach does not. When the fractional order is $\alpha = 1$, the fractional divergence recovers the same 'trivial' solutions as the conventional method, hence the former is a generalisation of the latter. Note that one can set $\alpha_1 \neq \alpha_2$ or $\alpha_1 = \alpha_2 = \alpha$ or any combination, where $\alpha_i \in \mathcal{R}$ and $\alpha_i \in \mathcal{C}$.

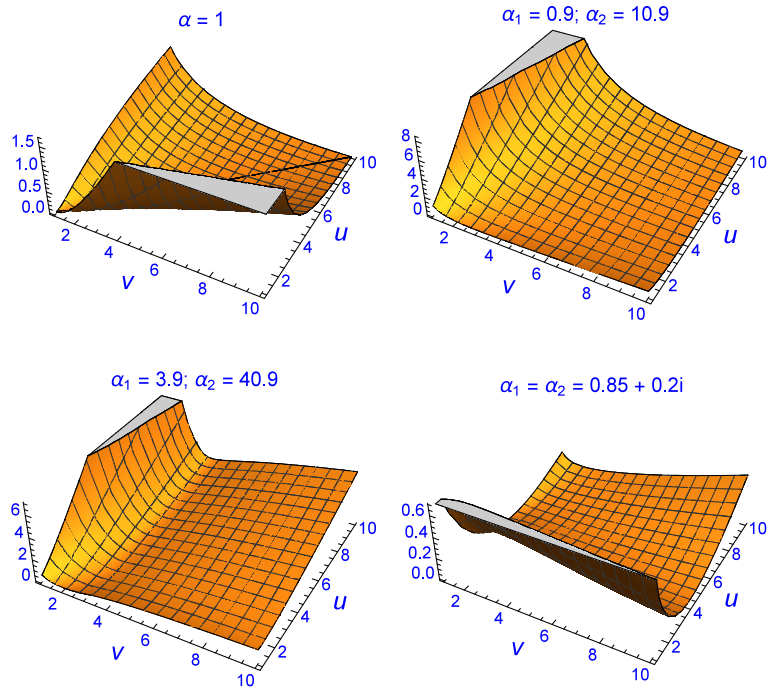


Figure 4. Further manipulation of the (fractional) divergence manifold between two Exponential-densities via the fractional orders α_1 and α_2.

Finally, it is worth discussing the $\alpha = 1$ or conventional divergence image map on the right of Figure 3. At first glance it appears that the divergence is also very small on either side of the $u = v$ solutions which would indicate that there must be other solutions apart from those given by $u = v$. However this is misleading. As the (u, v) parameters of the Exponential-densities increase in value, ($u \to \infty$ and $v \to \infty$), the Exponential-densities decay very quickly to zero. As this happens to both of

them simultaneously, the densities tend to have the same asymptotic behaviour whenever (u, v) are large, giving the impression that the divergence is zero between them. In other words,

$$D(p(y;u)||q(y;v)) = \lim_{u,v\to\infty} \int_\Omega p(y;u) \log\left(\frac{p(y;u)}{q(y;v)}\right) dy = \int_\Omega (\sim 0) \log\left(\frac{\sim 0}{\sim 0}\right) dy = 0, \tag{86}$$

where the last term on the right is valid by the *Definition* found in the previous section and ~ 0 means that the densities asymptotically approach zero (rapidly) for large (u, v). Caution must be used when interpreting the divergence solutions for the conventional case on either side of the $u = v$ line. These solutions are trivial and are due to the decay process of the densities and not because there are alternative solutions in addition to the ones given by $u = v$. This explains the $''V''$-shape that is diagonal to the $u - v$ axes.

6. An Application of the Fractional Divergence to Detection Theory

In this section, it will be shown how the fractional divergence can be used to solve an important problem in the field of signal processing. The problem consists of detecting signals embedded in background noise or clutter. Suppose that a hypothesis test is constructed. Set H_0 to be the null hypothesis which describes only the noise/clutter. Let H_1 be the alternative hypothesis that there is a signal of interest that has to be detected in the noise/clutter. That is,

$$\begin{aligned} H_0 &: \textit{noise/clutter}. \\ H_1 &: \textit{signal} + \textit{noise/clutter}. \end{aligned} \tag{87}$$

It is usually the case where the density that describes the noise/clutter is known, e.g., Gaussian or Normal. Let $q_0(x)$ be a density that represents this situation. Let the alternative hypothesis be represented by the density $q_1(x)$, i.e., that there is a signal of interest embedded inside the noise/clutter. It is possible to construct a detector that can discriminate in some optimal fashion whether there is a signal present or not when sampling observed data. Let $p(x)$ be a density that is constructed by observing/measuring i.i.d. random variables. What is required is a metric which determines how close the observed data $p(x)$ is to either $q_0(x)$ and $q_1(x)$. If $p(x)$ is closer to $q_0(x)$, then it is more likely that it is not a signal of interest but rather what is being detected is merely noise/clutter. If the separation of $p(x)$ is closer to $q_1(x)$ instead, then it is highly probable that a signal is present, so a detection is declared. It should be clear that a minimum divergence detector can be constructed, which can differentiate if there is a signal present or not by calculating the divergence between the observed density and that of the the null and the alternative densities.

According to the Neyman–Pearson theorem that optimises the detection probability for a given false alarm rate, the log-likelihood ratio for the hypothesis test is:

$$\theta' = \prod_{i=1}^{N} \frac{q_1(x_i)}{q_0(x_i)}, \tag{88}$$

where the total number of samples observed is N. Taking the log-likelihood of (88) and normalising by N gives:

$$\theta \equiv \frac{1}{N} \log(\theta') = \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{q_1(x_i)}{q_0(x_i)}\right). \tag{89}$$

The log-likelihood θ is essentially a random variable. It is an average of N i.i.d. random variables $\theta_i = \log(q_1(x_i)/q_0(x_i))$. Accordingly, from the law of large numbers, for large N,

$$\theta \to \langle \theta_i \rangle, \tag{90}$$

where $< \cdot >$ is the expectation and $i = 1, 2, ..., N$. By the expectation (90) for the continuous case, one has

$$\begin{aligned}\langle\theta\rangle &= \int p(x)\log\left(\frac{q_1(x)}{q_0(x)}\right)dx \\ &= \int p(x)\log\left(\frac{q_1(x)p(x)}{q_0(x)p(x)}\right)dx \\ &= \int p(x)\log\left(\frac{p(x)}{q_0(x)}\right)dx - \int p(x)\log\left(\frac{p(x)}{q_1(x)}\right)dx \\ &= D(p(x)||q_0(x)) - D(p(x)||q_1(x)).\end{aligned} \tag{91}$$

Hence, the divergence is related to the expectation of the log-likelihood ratio. For large N and by the Neyman–Pearson theorem:

$$\langle\theta\rangle \underset{H_0}{\overset{H_1}{\gtrless}} \tau', \tag{92}$$

where τ' is the un-normalised threshold. The minimum distance detector based on the divergence is given by:

$$D(p(x)||q_0(x)) - D(p(x)||q_1(x)) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{N}\tau' \equiv \tau, \tag{93}$$

with τ being the normalised by N threshold. For a threshold $\tau = 0$, the detection scheme becomes

$$D(p(x)||q_0(x)) \underset{H_0}{\overset{H_1}{\gtrless}} D(p(x)||q_1(x)). \tag{94}$$

If the divergence indicates that the distance of $p(x)$ to the null hypothesis $q_0(x)$ is greater than the distance to the alternative hypothesis $q_1(x)$, then H_1 is true, which means that a signal of interest is detected and vice versa. The main problem is that the detection scheme (93) or (94) requires the estimation of parameters for each density, i.e., $p(x;\vec{\xi}_1)$, $p(x;\vec{\xi}_2)$ and $p(x;\vec{\xi}_3)$. The critical issue that arises is that the parameters $(\vec{\xi}_1, \vec{\xi}_2, \vec{\xi}_3)$ are estimated from the observed data. Unfortunately, in order to obtain accurate estimates for these parameters, the number of samples N must be very large. In reality, however, this is never the case. There are only a small number of samples n that can be used for estimation purposes, i.e., $n \in N : n << N$. This introduces error in the estimation of $(\vec{\xi}_1, \vec{\xi}_2, \vec{\xi}_3)$ and, as a consequence, the divergence detector does not perform optimally.

Using the fractional divergence approach means that the parameters depend on the fractional order, $(\vec{\xi}_1(\alpha), \vec{\xi}_2(\alpha), \vec{\xi}_3(\alpha))$. Thus, even if the parameters are estimated using only a small sample n in each case, the fractional order can be changed in order to compensate for this by varying the divergences to obtain the optimal solution as if the sampling was very large to begin with. The fractional-order(s) 'fine-tunes' the performance of the detector by acting as a correction factor to the loss experienced in the estimation process for the parameters because of poor or small sampling.

7. Conclusions

It has been shown that the divergence between different probability densities can be studied using the Kullback–Leibler approach. It is possible to find solutions that indicate where two competing density models approach each other asymptotically, but the solutions are generally unique or trivial in nature. The fractional divergence employs fractional calculus to improve on the conventional divergence results beyond the trivial or unique cases. Apart from the improved overall performance, fractional solutions open up the possibility of giving further insights into problems requiring this type of analysis.

Acknowledgments: The author would like to thank the reviewers for their suggestions on how to improve the paper.

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

References

1. Jeffrey, H. *Theory of Probability*, 2nd ed.; Clarendon Press: Oxford, UK, 1948.
2. Flemming, T. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory* **2000**, *46*, 1602–1609.
3. Renyi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematics, Statistics and Probability, Berkeley, CA, USA, 20 June–30 July 1960; Volume 1, pp. 547–561.
4. Borland, L.; Plastino, A.R.; Tsallis, C. Information gain within nonextensive thermostatistics. *J. Math. Phys.* **1998**, *39*, 6490–6501.
5. Ubriaco, M.R. Entropies based on fractional calculus. *Phys. Lett. A* **2009**, *373*, 2516–2519.
6. Machado, J.T. Fractional order generalized information. *Entropy* **2014**, *16*, 2350–2361.
7. Lin, J. Divergence measures based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
8. Machado, J.T. A probabilistic interpretation of the fractional-order differentiation. *Fract. Calc. Appl. Anal.* **2003**, *6*, 73–80.
9. Nguyen, V.T. Fractional calculus in probability. *Probab. Math. Stat.* **1984**, *3*, 173–189.
10. Machado, J.T. Fractional coins and fractional derivatives. *Abstr. Appl. Anal.* **2013**, 5, doi:10.1155/2013/205097.
11. Jumarie, G. Probability calculus of fractional order and fractional Taylor's series application to Fokker-Planck equation and information of non-random functions. *Chaos Solitons Fractals* **2009**, *40*, 1428–1448, doi:10.1016/j.chaos.2007.09.028.
12. Resnik, S.I. *A Probability Path*; Birkhauser: Boston, MA, USA, 1998.
13. Mostafaei, M.; Ahmadi Ghotbi, P. Fractional probability measure and its properties. *J. Sci.* **2010**, *21*, 259–264.
14. El-Shehawy, S.A. On properties of fractional probability measure. *Int. Math. Forum* **2016**, *11*, 1175–1184 .
15. Swerling, P. Probability of detection for fluctuating targets. *IRE Trans. Inf. Theory* **1960**, *IT-6*, 269–308.
16. Gandhi, P.; Kassam, S. Analysis of CFAR processors in nonhomogeneous background. *IEEE Trans. Aerosp. Electron. Syst.* **1988**, *24*, 427–445.
17. Rohling, H. Radar CFAR thresholding in clutter and multiple target situations. *IEEE Trans. Aerosp. Electron. Syst.* **1983**, *19*, 608–621.
18. Tuzlukov, V.P. *Signal Detection Theory*; Springer: Boston, MA, USA, 2001.
19. Levanon, N. *Radar Principles*; Wiley: New York, NY, USA, 1988.
20. Amari, S.; Nagaoka, H. Methods of information geometry. In *Translations of Mathematical Monographs*; American Mathematical Society: Provindence, RI, USA, 2000; Volume v191, ISBN 978-0821805312.
21. Alexopoulos, A. One-parameter Weibull-type distribution and its relative entropy. *Digit. Signal Process.* **2017**, under review.
22. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
23. Goutis, C.; Robert, C.P. Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika* **1998**, *85*, 29–37.
24. Van Erven, T.; Harremoes, P. Renyi divergence and Kullback–Leibler divergence. *IEEE Trans. Inf. Theory* **2014**, *60*, 3797–3820.
25. Do, M.N.; Vetterli, M. Wavelet-based texture retrieval using generalized Gaussian density and Kullback–Leibler distance. *IEEE Trans. Image Process.* **2002**, *11*, 146–158.
26. Perez-Cruz, F. Kullback–Leibler divergence estimation of continuous distributions. In Proceedings of the IEEE International Symposium on Information Theory (ISIT), Toronto, ON, Canada, 6–11 July 2008.
27. Lee, Y.K.; Park, B.U. Estimation of Kullback–Leibler divergence by local likelihood. *Ann. Inst. Stat. Math.* **2006**, *58*, 327–340.
28. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley-Interscience: New York, NY, USA, 1991.
29. Wang, C.P.; Ghosh, M. A Kullback–Leibler divergence for Bayesian model diagnostics. *Open J. Stat.* **2011**, *1*, 172–184.
30. Alexopoulos, A.; Weinberg, G.V. Fractional-order Pareto distributions with application to X-band maritime radar clutter. *IET Radar Sonar Navig.* **2015**, *9*, 817–826.

31. De Oliveira, E.C.; Machado, J.A.T. A review of definitions for fractional derivatives and integral. *Math. Probl. Eng.* **2014**, 6, doi:10.1155/2014/238459.
32. Alexopoulos, A.; Weinberg, G.V. Fractional-order formulation of power-law and exponential distributions. *Phys. Lett. A* **2014**, *378*, 2478–2481.
33. Kulish, V.V.; Lage, J.L. Application of fractional calculus to fluid mechanics. *Fluids Eng.* **2002**, *124*, 803–806.
34. Douglas, J.F. Some applications of fractional calculus to polymer science. *Adv. Chem. Phys.* **1997**, *102*, 121–192.
35. Fellah, Z.E.A.; Depollier, C. Application of fractional calculus to the sound waves propagation in rigid porous materials: Validation via ultrasonic measurement. *Acta Acust.* **2002**, *88*, 34–39.
36. Assaleh, K.; Ahmad, W.M. Modeling of speech signals using fractional calculus. In Proceedings of the 9th International Symposium on Signal Processing and Its Applications (ISSPA), Sharjah, UAE, 12–15 February 2007; pp. 1–4.
37. Mathieu, B.; Melchior, P.; Oustaloup, A.; Ceyral, C. Fractional differentiation for edge detection. *Fract. Signal Process. Appl.* **2003**, *83*, 2285–2480.
38. Soczkiewicz, E. Application of fractional calculus in the theory of viscoelasticity. *Mol. Quantum Acoust.* **2002**, *23*, 397–404.
39. Machado, J.A.T.; Jesus, I.S.; Cunha, J.B.; Tar, J.K. Fractional dynamics and control of distributed parameter systems. *Intell. Syst. Serv. Mank.* **2006**, *2*, 295–305.
40. Hilfer, R. *Applications of Fractional Calculus in Physics*; World Scientific Publishing: Singapore, 2000.
41. Podlubny, I. *Fractional Differential Equations*; Academic Press: Cambridge, MA, USA, 1999; Volume 198.