Jestr

Journal of Engineering Science and Technology Review 16 (2) (2023) 18 - 21

Letter to the Editor

JOURNAL OF
Engineering Science and Technology Review

www.jestr.org

A Fractional Cross-Entropy Based on Caputo Fractional-Order Derivative

Slimane Benmahmoud[1,*] and Nora Ouagueni[2]

[1]*Faculty of Technology. Department of Electronics Engineering, Signals and Systems' Laboratory (LASS). M'sila University. M'sila. Algeria*
[2]*Laboratory for Pure and Applied Mathematics. M'sila University. M'sila. Algeria*

Received 13 November 2022; Accepted 17 March 2023

Abstract

Cross-entropy (CE) is a measure of how different two probability distributions are. It is commonly used in machine learning and information theory to compare the predicted probability distribution with the actual probability distribution. On the other hand, fractional calculus is a branch of mathematical analysis which studies different possible approaches of defining fractional-order integrals and derivatives. In this paper, we have derived a novel generalized fractional CE (FCE). To do so, we have differentiated the CE's generating function (i.e., $h(t) := \int_{S_X} \frac{f_X(x)}{f_Y(y)} f_Y^{-t}(x)dx$) using a α -order Caputo fractional-order derivative ${}_{a^+}^{\alpha}CD f$. When the order of differentiation $\alpha \to 1$, we recover the ordinary Shannon's CE, which corresponds to the results from a first-order ordinary differentiation. This allows to calculate the FCE for various values of $\alpha \neq 1$ and compare them to the conventional CE ($\alpha = 1$) to get further insights on its behavior. Some examples and illustrations of the proposed FCE are also presented.

Keywords: Cross-entropy (CE), Riemann-Liouville/Caputo fractional integral/derivative, fractional calculus, entropy's generating function, Tsallis/Rényi entropy, information measure.

1. Introduction

In 1948, Shannon proposed the concept of entropy in the context of communication theory [1]. It consists of a measure of surprise or uncertainty associated with the probability distribution of a random variable (RV). For a discrete RV X taking values in $\mathcal{X} = \{x_1, x_2, \ldots, x_q\}$ and having a probability mass function $p_i = P(X = x_i)$ with $\sum_{i=1}^{q} p_i = 1$ and $p_i \geq 0$ for $i = 1, \ldots, q$, it is given by:

$$H(X) = -\sum_{i=1}^{q} p_i \log p_i. \tag{1}$$

This suggested measure of uncertainty (i.e., Eq (1)) with its properties has shown an agreement with the intuitive notions of randomness and justified its usefulness with respect to statistical problems in communication theory.

Known as the differential entropy, the continuous analogue of the discrete entropy defined in Eq (1), for a continuous RV is given by [2]:

$$h(X) = -\int_{S_X} f_X(x) \log f_X(x) \tag{2}$$

where S_X and $f_X(x)$ are the support and the probability density function (PDF) of the RV X, respectively.

Another very important information measure, in the mathematical statistics, is the Kullback-Leibler divergence (also known as the relative entropy and the I-divergence) [3]. For discrete probability distributions P and Q defined in the same probability space $\mathcal{X}$, it is defined to be [4]:

$$D(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)}\right) \tag{3}$$

It is a type of statistical distance. It measures how the probability distribution P is different from Q.

For two continuous RVs X and Y with marginal PDFs $f_X(x)$ and $f_Y(y)$ and supports $S_X \subseteq S_Y \subseteq \mathbb{R}$, it is given by [5], [6]:

$$\mathcal{D}(P \parallel Q) = \int_{S_X} f_X(x) \log \left(\frac{f_X(x)}{f_Y(x)}\right) dx \tag{4}$$

Eq (4) can be re-written as follows:

$$\mathcal{D}(P \parallel Q) = \overbrace{\int_{S_X} f_X(x) \log f_X(x) dx}^{=-h(X)} \underbrace{- \int_{S_X} f_X(x) \log f_Y(x) dx}_{=h(X,Y)}. \tag{5}$$

The second term in Eq (5), denoted $h(X,Y)$ is referred to as the continuous cross-entropy (CE) of X relative to Y over S_X. Cross-entropy is a measure of how different two probability distributions are. It is commonly used in machine learning and information theory to compare the predicted probability distribution with the actual probability distribution. In machine learning, cross-entropy is often used as a loss function to measure the difference between the predicted probability distribution and the actual probability distribution of a classification problem [7].

The cross-entropy loss penalizes incorrect predictions more heavily than correct predictions, which makes it a useful metric for training models [8].

The cross-entropy has several variants, including binary cross-entropy, categorical cross-entropy, and sparse categorical cross-entropy. Binary cross-entropy is used for binary classification problems, while categorical cross-entropy and sparse categorical cross-entropy are used for multi-class classification problems [9], [10].

*E-mail address: slimane34a@gmail.com
ISSN: 1791-2377
doi:10.25103/jestr.162.03

Although CE is widely used as an information measure in various fields such as machine learning, information theory, and statistics, it has some limitations that are worth considering, including the following [11]- [15]:

- It assumes that the random variables X and Y are independent.
- It is not symmetric (i.e., $h(X,Y) \neq h(Y,X)$), which makes it challenging to interpret the results and compare different models.
- It can be biased towards certain types of distributions, depending on the choice of the base distribution.

In this paper, we introduce a generalized version of it (named fractional-order cross-entropy (FCE)) by re-writing the previous ordinary CE in the form of a differ-integral equation, then we deform the ordinary derivative to a Caputo fractional-order one. The use of this later allows to get a broader idea in a generalized metric space concerning the CE and other related information measures.

2. A Review on Fractional Integrals and Derivatives

Fractional calculus (a mathematical analysis branch which studies different possible approaches of defining fractional-order integrals and derivatives) can be traced back to a letter written to l'Hopital by Leibniz in 1695 [16]. In 1832, Liouville carried out a heavy-handed investigation on FC [17]. After that, the Riemann-Liouville (RL) fractional integro-differential operator was introduced by Riemann in [18] along with a comprehensive theory of FC.

The left-sided Riemann-Liouville (RL) fractional integral $\mathrm{RL}I_{a^+}^{\alpha} f$ of order $\alpha \in \mathbb{R}\,(\alpha > 0)$ of an integrable function $f: [a,b] \to \mathbb{R}, (0 \leq a < b \leq \infty)$ is defined as [10]:

$$\left(\mathrm{RL}I_{a^+}^{\alpha} f\right)(t) = {}^{\mathrm{RL}}\, I_{a^+}^{\alpha}[f(x)](t) = \frac{1}{\Gamma(\alpha)} \int_a^t (t-x)^{\alpha-1} f(x)dx \tag{6}$$

with $a \in \mathbb{R}, t > a, \alpha > 0$, where $\Gamma(.)$ is Euler's gamma function defined as $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1} e^{-x} dx, (\alpha \in \mathbb{R}\,(\alpha > 0))$.

The left-sided Caputo fractional derivative $CD_{a^+}^{\alpha} f$ of order $\alpha \in \mathbb{R}(\alpha > 0)$ of an integrable and differentiable function $f: [a,b] \to \mathbb{R}, (0 \leq a < b \leq \infty)$is defined as [19]:

$$\left(CD_{a^+}^{\alpha} f\right)(t) = {}^{C}\, D_{a^+}^{\alpha}[f(x)](t) = \left(\mathrm{RL}I_{a^+}^{n-\alpha} \left(\frac{d}{dt}\right)^n f\right)(t) \tag{7}$$

with $a \in \mathbb{R}, t > a, \alpha > 0, n = [\alpha] + 1$.

When $0 < \alpha \leq 1$, we get:

$$\left(CD_{a^+}^{\alpha} f\right)(t) = \left(\mathrm{RL}I_{a^+}^{1-\alpha} \frac{d}{dt} f\right)(t), \quad (a \in \mathbb{R}, t > a) \tag{8}$$

3. A Caputo Fractional Derivative-Based FCE

3.1. Definition 1.

Let X and Y be two non-negative continuous random variables (RVs) with probability density functions (PDFs) $f_X(x)$, $f_Y(y)$ and supports S_X, S_Y respectively. The CE, $h(X,Y)$, of X relative to Y over S_X is defined as:

$$h(X,Y) := -\int_{S_X} f_X(x) \log f_Y(x) dx = E_X[-\log f_Y(X)] \tag{9}$$

when the integral exists. $E[X]$ is the expected value of the RV X. CE is commonly used in the fields of information theory, statistics, and machine learning for tasks such as density estimation, generative modeling, and anomaly detection. It is often used as a loss function for training generative models such as variational autoencoders and generative adversarial networks.

3.2. Remark

Through the whole paper, the base of the logarithm will be set to Euler's number $e = \sum_{n=0}^{\infty} \frac{1}{n!}$.

Our basic idea consists of re-writing Eq (4) as follows:

$$h(X,Y) := -\lim_{t \to -1} \frac{d}{dt} \int_{S_X} \frac{f_X(x)}{f_Y(x)} f_Y^{-t}(x) dx \tag{10}$$

Then, we deform the ordinary differential operator $\frac{d}{dt}$ in Eq (5) to the Caputo fractional differential operator $CD_{a^+}^{\alpha}$ defined in Eq (3) (which reduces to $\frac{d}{dt}$ in the limit $\alpha \to 1$). Based on these ideas, we derive in the following theorem a new class of FCE.

3.3. Theorem 1

Let X and Y be two non-negative continuous random variables (RVs) with probability density functions (PDFs) $f_X(x)$ and $f_Y(y)$ and supports $S_X$and S_Y respectively. The FCE, $h^{\alpha}(X,Y)$ of order α, of X relative to Y over S_X is defined as:

$$h^{\alpha}(X,Y) := \int_{S_X} f_X(x)(-\log f_Y(x))^{\alpha} dx = E_X[(-\log f_Y(X))^{\alpha}] \tag{11}$$

with $0 < \alpha \leq 1$.

3.4. Proof

Using the operator defined in Eq (3) (where the lower limit of the RL-integral is taken to zero, i.e., $a = 0$, without loss of generality), Eq (5) can be re-written as follows:

$$h^{\alpha}(X,Y): = -\lim_{t \to -1} \frac{d}{dt} \left(\mathrm{RL}I_0^{1-\alpha} \left[\int_{S_X} \frac{f_X(x)}{f_Y(y)} e^{-t \log f_Y(x)} dx \right](t) \right),$$

with $0 < \alpha \leq 1$.

Therefore, we need to solve the following integral:

$$h^{\alpha}(X,Y) := -\lim_{t \to -1} \frac{d}{dt} \int_{S_X} \left(\int_0^t (t - y)^{-\alpha} \frac{f_X(x)}{f_Y(y)} e^{-y \log f_Y(x)} dy \right) dx$$

By letting $w = t - y$, using the definition of the $\Gamma(.)$ function, taking the ordinary derivative and setting $t = 1$, we get Eq (4).

The expression in (11) allows us to calculate the FCE ($h^{\alpha}(X,Y)$) for different values of α and compare them to each other and to the conventional CE ($\alpha = 1$) at the same time to get further insights on its behavior.

In the following, we give few examples of the FCE for some common continuous probability distributions.

3.5. Example 1

Let X and Ybe two non-negative continuous uniformally distributed RVs on $[0, a_X]$, $a_X > 0$, and $[0, a_Y]$, $a_Y > 0$, respectively .i.e., $X \sim Uniform\ (0,\ a_X)$ and $Y \sim Uniform(0, a_Y)$. Then, the FCE of Xrelative to Y is given by:

$$h^{\alpha}(X,Y) = (log\ a_Y)^{\alpha} \tag{12}$$

with $0 < \alpha \leq 1$

To get further insights, in Fig. 1, we have plotted FCE in Eq(12) as a function of α.

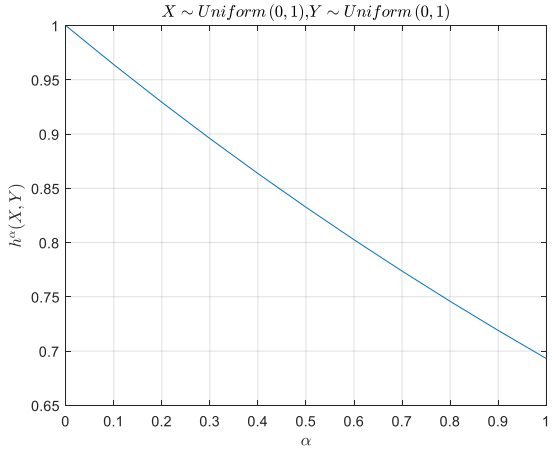

Fig.1. The FCE $h^{\alpha}(X,Y)$ in Eq (12) as a function of α.

3.6. Example 2

Let X and Y be two non-negative continuous exponentially distributed RVs rate parameters $\lambda_X$and λ_Y, i.e., $X \sim Exp(\lambda_X)$ and $Y \sim Exp(\lambda_Y)$. Then, the FCE of X relative to $Y$is given by:

$$h^{\alpha}(X,Y) = \frac{\lambda_Y^{\alpha-\frac{\lambda_X}{\lambda_Y}}}{\lambda_X^{-\alpha}} \Gamma\left(\alpha + 1, -\frac{\lambda_X\, log\, \lambda_Y}{\lambda_Y}\right) \tag{13}$$

with $0 < \alpha \leq 1 \& \lambda_Y \leq 1$, where $\Gamma(.,.)$ is the upper incomplete gamma function (See Eq. (8.2.2) in [20]).

To get further insights, in Fig. 2, we have plotted FCE in Eq(13) as a function of α.

3.7. Example 3

Let $X$and Y be two non-negative continuous normally distributed RVs with mean and standard deviation parameters (μ_X, σ_X) and (μ_Y, σ_Y), respectively. i.e., $X \sim \mathcal{N}(\mu_X, \sigma_X)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$. Then, the FCE of Xrelative to $Y$is given by:

$$h^{\alpha}(X,Y) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_0^{+\infty} e^{\frac{-(x-\mu_X)^2}{2\sigma_X^2}} \left(\frac{1}{2} log(2\pi\sigma_Y^2) + \frac{(x-\mu_Y)^2}{2\sigma_Y^2}\right)^{\alpha} dx \tag{14}$$

with $0 < \alpha \leq 1$.

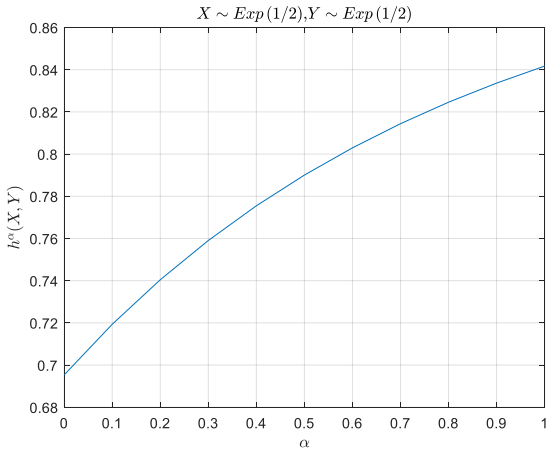

Fig.2. The FCE $h^{\alpha}(X,Y)$ in Eq (13) as a function of α.

4. Conclusion

In this paper, we have introduced a new CE functional, $h^{\alpha}(X,Y)$, which generalizes Shannon's conventional CE, $h(X,Y)$, by applying Caputo fractional-order derivative to the CE's generating function. The proposed FCE has inherited the merits of the conventional CE, since the former is a generalization of the latter. As a future perspective, the possibility of estimating the FCE from empirical random samples is worth being explored. Testing its reliability through simulations and exploring its application in computer vision are also worthy to be considered. Cross-entropy also has many applications in machine learning, including natural language processing, computer vision, and recommendation systems. It is a widely used metric for evaluating the performance of models in these fields.

Notations: Here we give, for quick reference, the common notations used in this paper. $f_X(x)$, $F_X(x)$, $\mathbb{E}[X]$, S_X, $h(X)$ denote, respectively, the probability density function (PDF), the cumulative distribution function (CDF), the expectation, the support the random variable X. $h(X,Y)$ and $h^{\alpha}(X,Y)$ are the CE and FCE of Xrelative to Yover S_X, respectively. $RLI_{a^+}^{\alpha} f$ and $CD_{a^+}^{\alpha} f$are the left-sided Riemann-Liouville (RL) fractional integral and the left-sided Caputo fractional derivative of order $\alpha \in \mathbb{R}(\alpha > 0)$. $\Gamma(.,.)$ is the upper incomplete gamma function.

References

1. Shannon, Claude Elwood. "A mathematical theory of communication." *The Bell system technical journal* 27.3, 1948, pp. 379-423.
2. Thomas, M. T. C. A. J., and A. Thomas Joy. *Elements of information theory*. Wiley-Inter-science, 2006.

3. Csiszár, Imre. "I-divergence geometry of probability distributions and minimization problems." *The annals of probability*, 3, 1975, pp. 146-158.
4. MacKay, David JC, and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.
5. Kullback, Solomon, and Richard A. Leibler. "On information and sufficiency." *The annals of mathematical statistics* 22.1, 1951, pp.79-86.
6. Kullback, Solomon. *Information theory and statistics*. Courier Corporation, 1997.
7. Boer, P. T. de, Dirk P. Kroese, Shie Mannor and Reuven Y. Rubinstein. "A Tutorial on the Cross-Entropy Method." Annals of Operations Research 134, 2005, pp. 19-67.
8. Ho, Yaoshiang and Samuel Wookey. "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling." *IEEE Access 8*, 2020, pp. 4806-4813.
9. Saleem, Muhammad Asif, Norhalina Senan, Fazli Wahid, Muhammad Aamir, Ali Samad and Mukhtaj Khan. "Comparative Analysis of Recent Architecture of Convolutional Neural Network." *Mathematical Problems in Engineering,* 2022, 2022.
10. Kim, Da Ye, and Chul Min Song. "Developing a Discharge Estimation Model for Ungauged Watershed Using CNN and Hydrological Image" *Water 12*, (12), 2020, p. 3534.
11. Cover, T. M., & Thomas, J. A. *Elements of information theory*. John Wiley & Sons. 2012.
12. Goodfellow, I., Bengio, Y., & Courville, A. *Deep learning*. MIT press. 2016
13. Kullback, S., & Leibler, R. A. "On information and sufficiency". *The Annals of Mathematical Statistics*, 22(1), 1951, pp.79-86.
14. Li, M., Vitanyi, P. M. B., Li, M., & Vitányi, P. M. B. *An introduction to Kolmogorov complexity and its applications* (3rd ed.). Springer. 2008.
15. Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press. 2012
16. Leibniz, Gottfried Wilhelm. "Letter from Hanover, Germany to GFA L'Hospital, September 30, 1695." *Mathematische Schriften* 2, 1849, pp. 301-302.
17. Liouville, Joseph. "*Mémoire sur l'usage que l'on peut faire de la formule de Fourier, dans le calcul des différentielles à indices quelconques*.", 1835, pp. 219-232.
18. Riemann, Bernhard. "Versuch einer allgemeinen Auffassung der Integration und Differentiation." *Gesammelte Werke* 62. 1876
19. Kilbas, Anatoli Aleksandrovich, Hari M. Srivastava, and Juan J. Trujillo. *Theory and applications of fractional differential equations*. Vol. 204. Elsevier, 2006.
20. Olver, Frank WJ, et al., eds. *NIST handbook of mathematical functions hardback and CD-ROM*. Cambridge university press, 2010.