International Journal of Engineering and Information Systems (IJEAIS)
ISSN: 2643-640X
Vol. 9 Issue 5 May - 2025, Pages: 13-16

A Review of Kullback-Leibler Divergence Across Different Probability Distributions

Nada Mohammed Abbas1, *Ahmed Hadi Hussain*[2]

1Education College for Pure Sciences, Department of Mathematics , *University of Babylon, Iraq*
pure.nada.moh@uobabylon.edu.iq
2Department of Air conditioning and Refrigeration Engineering, College of Engineering Al-musayab, University of Babylon, Iraq
met.ahmed.hadi@uobabylon.edu.iq

Abstract— *This paper examines recent advancements and discoveries concerning Kullback-Leibler divergence.*

1- Introduction

In this review, we highlight the Kullback-Leibler divergence as a fundamental concept used in numerous research studies. To provide a deeper understanding of this concept, this section presents the key principles and foundations explored in previous research, organized in a way that clarifies their role and significance in various fields

.**Definition 1.1[1,2,3,4,6,7,8]:** The Kullback-Leibler (KL) distance, measures the difference between two probability distributions p(x) and q(x), given by:$D(p||q) = \int p(x) log \frac{p(x)}{q(x)} dx = E_p[log \frac{p(x)}{q(x)}]$.

Definition 1.2[1]: The J-divergence distance is a symmetric version of the KL distance, defined as $D_s(p||q) = D(p||q) + D(q||p)$.

Definition 1.3[1]: A Reproducing Kernel Hilbert Space (RKHS) is a functional space associated with a positive kernel function $k(x, y)$.

Definition 1.4[3]: The Hyperanalytic Wavelet Transform is defined using the hypercomplex mother wavelet $w_a(x, y)$ associated with a real mother wavelet $w(x, y)$ as follows:

$w_a(x, y) = w(x, y) + iH_x\{w(x, y)\} + jH_y\{w(x, y)\} + kH_x\{H_y\{w(x, y)\}\}$

where H_x and H_y are the Hilbert transforms across rows and columns, respectively.

Definition 1.5[3]: The Complex Generalized Gaussian Distribution CGGD for a bivariate probability density function z_b is defined in the general form:

$$p(x, y) = \frac{\beta(c)}{\sigma^2} exp\left\{-\left(\frac{\Gamma\left(\frac{2}{c}\right)}{\Gamma\left(\frac{1}{c}\right)}\right)^c \left(\frac{x^2 + y^2}{\sigma^2}\right)^c\right\}$$

where $\beta(c) = \frac{c\Gamma\left(\frac{2}{c}\right)}{\pi\Gamma^2\left(\frac{1}{c}\right)}$, and Γ is the gamma function.

Definition 1.6[4]: The probability density function of the Weibull distribution is given by

$$f(x) = \frac{\kappa}{l} (\frac{x}{l})^{\kappa-1} \exp\left|-(\frac{x}{l})^{\kappa}\right|$$

Definition 1.7[4]: The Euler-Mascheroni constant is a mathematical constant that appears frequently in special functions, particularly in expressions involving the Gamma function and logarithmic integrals. Its approximate value is: $\gamma \approx 0.5772$.

Definition 1.8[5]: If X and Y are random variables with cumulative distribution functions $F(x)$ and $G(x)$, respectively, the cumulative Kullback-Leibler information is defined as:

$$G_{KL}(X, Y) = \int_l^{\max\{r_X, r_Y\}} F(t) \log(\frac{F(t)}{G(t)}) dt + E(X) - E(Y)$$

Where $l = \inf\{t \in R: F(t) > 0\}$ and $r_X = \sup\{t \in R: F(t) < 1\}$

Definition 1.9[5]: A measure of the difference between two probability distributions based on their cumulative distribution which is called the cumulative inaccuracy is defined by the functions:

$$K(X, Y) = -\int_l^{max\{r_X, r_Y\}} F(t) \log\big(G(t)\big)\, dt$$

Definition 1.10[6]: A random variable x and a positive random variable y follow a normal-gamma distribution if their joint probability density function is given by:

$$p(x, y) = N(x; \mu, (y \wedge)^{-1}. Gam(y; a, b)$$

where $N(x;\mu,\Sigma)$ is the multivariate normal distribution, and $\mathrm{Gam}(y;a,b)$ is the gamma distribution: $Gam(y;a,b) = \frac{b^a}{\Gamma(a)} y^{a-1}e^{-by}, y > 0$

Definition 1.11[6]: When applying a general linear model (GLM) with normal-gamma conjugate priors:

$$y = X\beta + \epsilon, \epsilon \sim N(0, \sigma^2 V)$$

The model complexity is computed using the KL divergence between the posterior and prior:

$$Com(m) = KL[p(\beta,\tau \mid y) \mid\mid p(\beta,\tau)]$$

Definition 1.12[7]: The discrete normal distribution $N_Z(\mu,\Sigma)$ is the unique discrete distribution defined on the integer lattice support Z^d with mean μ and covariance matrix Σ that maximizes Shannon entropy. Its probability mass function (pmf) is expressed as an exponential family .

Definition 1.13[7]: The Renyi divergence $D_\alpha[r:s]$ between two probability mass functions $r(x)$ and $s(x)$ on the support $\mathrm{X} = Z^d$ is defined for any positive real number $\alpha \neq 1$ as:

$D_\alpha[r:s] = \frac{1}{\alpha-1} log\left(\sum_{x\in \mathrm{X}} r(x)^\alpha s(x)^{1-\alpha}\right)$.

Definition 1.14[7]: The Sharma-Mittal divergence

$D_{\alpha,\beta}[p:q]$ between two probability mass functions $p(x)$ and $q(x)$ on the support X is defined as:

$$\boldsymbol{D_{\alpha,\beta}[p:q] = \frac{1}{\beta-1}\left(\left(\sum_{x\in \mathrm{X}} p(x)^\alpha q(x)^{1-\alpha}\right)^{\frac{1-\beta}{1-\alpha}} - 1\right)}$$

Definition 1.15[7]: The squared Hellinger divergence Hellinger $D^2_{Hellinger}[r,s]$ between two probability mass functions $r(x)$ and $s(x)$ is defined as**:**

$$\boldsymbol{D^2_{Hellinger}[r,s] = \frac{1}{2}\sum_{x\in \mathrm{X}}\left(\sqrt{r(x)} - \sqrt{s(x)}\right)^2}$$

2- Review

Shaohua Kevin Zhou and Rama Chellappa[1] proposed a novel approach for computing the Kullback-Leibler (KL) distance between two Gaussian distributions in Reproducing Kernel Hilbert Space (RKHS). Their study addresses the limitations of traditional KL distance computation, particularly in handling nonlinear data structures. By embedding data into RKHS using kernel methods, they derived new analytical expressions for KL and J-divergence distances between two Gaussian distributions .To overcome the computational challenges posed by the infinite-dimensional nature of RKHS ,they introduced a low-rank approximation of the covariance matrix , ensuring that the dominat eigenpairs are preserved . Additionally , they analyzed the limiting behavior of the KL distance as the regularization parameter ρ approaches zero, demonstrating that the proposed method maintains key statistical properties while improving the computational efficiency . The proposed method was evaluated on both synthetic and real-world datasets. In synthetic experiments, it effectively distinguished non-Gaussian distributions (such as "O", "D", and "X"-shaped uniform distributions) that share the same mean and covariance matrix—cases where traditional Gaussian models fail due to their reliance on second-order statistics only . In a face recognition application, the method achieved higher recognition accuracy (13/15 cases correctly classified) compared to conventional KL-based techniques showcasing its robustness in real-world pattern recognition tasks . The study demonstrated that RKHS enhances pattern separability by enabling a more expressive representation of data , making it a powerful tool for probabilistic modeling, machine learning, and pattern recognition. This approach provides an efficient framework for analyzing distributions in high-dimensional spaces, with potential applications in computer vision, statistical learning and kernel-based classification methods.

Researcher Sergiu C. Dragomir, in his study [2], arrived at several important findings regarding the properties of the exponential function of the Kullback-Leibler divergence. He proved that the function $exp[-D(p||q)]$ exhibits the superadditivity property, meaning that combining two probability distributions does not reduce its value. Additionally, he demonstrated that this function maintains upper and lower bounds under specific likelihood ratio conditions, providing constraints on its possible values. Furthermore, he established that $exp[-D(p||q)]$ is concave over the convex cone of probability distributions, implying that a mixture of distributions results in a value greater than or equal to the weighted average of the original values. The researcher also introduced lower bounds for the exponential function of the Kullback-Leibler divergence when using the harmonic mean of probability distributions, proving that $exp[D(p||H(q,r))]$ is always greater than or equal to the arithmetic mean of the original values. Finally, he conducted numerical experiments and analyzed the behavior of the relationships using graphical representations, confirming the theoretical results and reinforcing their applications in various fields such as Bayesian statistics, data analysis, and machine learning models.

Corina Nafornita and her colleagues (Yannick Berthoumieu, Ioan Nafornita, Alexandru Isar)[3] studied the Kullback-Leibler Distance (KL Distance) between Complex Generalized Gaussian Distributions (CGGD). The research focused on feature extraction in complex domain transforms, such as the Hyperanalytic Wavelet Transform (HWT), which exhibits a circularly symmetric distribution for subband coefficients. The CGGD model was used to model these coefficients, and a closed-form expression for the Kullback-Leibler distance between two CGGD distributions was derived.The results showed that the Kullback-Leibler distance is zero when the shape parameters of the two distributions are equal, and the sensitivity of this distance increases when the shape parameters differ. However, it was observed that the distance varies only slightly in certain intervals, which may limit its effectiveness in texture classification. The sensitivity of the distance to the shape parameters was analyzed, and it was found to be more responsive to shape parameter values closer to 0.3 compared to the Gaussian case (where the shape parameter is 1).The research concluded that the Kullback-Leibler distance can be useful for measuring the similarity between probability density functions of subbands, but it may not be sufficient in all cases, calling for the study of additional measures for texture classification in the future.

Researcher Christian Bauckhage[4], in this study, derived a closed-form solution for computing the Kullback-Leibler (KL) Divergence between two Weibull distributions, which is considered a significant scientific contribution since this divergence has often been mentioned in the literature but rarely presented explicitly and clearly. The researcher demonstrated that this divergence depends on the shape and scale parameters of the two distributions and involves logarithmic functions, the Euler-Mascheroni constant, and the Gamma function. The validity of the derived formula was confirmed by applying it to a special case where both distributions are exponential, simplifying the formula to the well-known expression for KL divergence between exponential distributions, thereby verifying the correctness of the mathematical approach. The researcher also highlighted the importance of this result for practical applications, particularly in fields such as data analysis, machine learning, and statistical model comparison, noting its potential use as a basis for constructing kernel functions that can be employed for classification using Support Vector Machines (SVM). Additionally, the study emphasized that this result has broad applications in text analysis, information retrieval, and image and text processing, as Weibull distributions are widely used to describe data characteristics in these domains.

The authors Antonio Di Crescenzo and Maria Longobardi[5] studied some properties and applications of cumulative Kullback-Leibler information, which serves as an extension of the traditional Kullback-Leibler information to the cumulative distribution function. The study included developing lower and upper bounds for this measure, as well as proposing a dynamic version for its application to past lifetimes, linking it to new concepts of relative aging. Additionally, two main applications were presented: the first involves analyzing the failure of nanocomponents by assessing the impact of stress levels on load duration, while the second focuses on digital image analysis using the empirical version of this measure to evaluate differences in gray levels between images.

Joram Soch and Carsten Allefeld[6] derived the Kullback-Leibler (KL) divergence for the normal-gamma distribution and demonstrated its equivalence to the Bayesian complexity penalty in the univariate general linear model (GLM) with conjugate priors. They applied this finding to two case studies: one using simulated data to analyze polynomial basis function selection and another using empirical neuroimaging data to compare different GLM formulations for functional MRI (fMRI) analysis. Their results highlight that the KL divergence effectively quantifies model complexity and helps detect differences between models that cannot be captured by accuracy alone, offering a more nuanced approach to model selection compared to traditional criteria like AIC or BIC.

Frank Nielsen[7] provided a comprehensive analysis of statistical divergences between discrete normal distributions, with a particular focus on the Kullback-Leibler divergence. The researcher demonstrated that discrete normal distributions form an exponential family with a cumulant function related to the Riemann theta function. The study presented several formulas for calculating common statistical divergences between these distributions, including the Renyi divergence and the Sharma-Mittal divergence. Additionally, the researcher introduced an efficient approximation technique for computing the Kullback-Leibler divergence using Renyi divergences or projective γ-divergences. Numerical examples were provided to illustrate these formulas and approximation techniques, highlighting their practical applications in fields such as machine learning and lattice-based cryptography.

Yufeng Zhang et al.[8] presented several key findings. The researchers proved that the reverse KL divergence can be bounded by an upper limit based on the **Lambert W** function when the forward KL divergence is constrained, indicating an **approximate symmetry** between the two directions for small values. Additionally, the study showed that KL divergence between three Gaussian distributions satisfies a **relaxed triangle inequality**, meaning that KL(N1||N3) can be expressed in terms of KL(N1||N2) and KL(N2||N3). These results are **dimension-independent**, making them highly applicable in high-dimensional fields such as **deep learning, reinforcement learning, and sample complexity research**, where they have been used in **out-of-distribution (OOD) detection** analysis and to provide **theoretical guarantees** for Gaussian policies in reinforcement learning.

REFERENCES

[1] Zhou, Shaohua Kevin, and Rama Chellappa. "Kullback-Leibler distance between two Gaussian densities in reproducing kernel Hilbert space." *International Symposium on Information Theory*. 2004.

[2] Sergiu C. Dragomir " Some properties for the exponential of the Kullback-Leibler divergence". Tamsui Oxford Journal of Mathematical Sciences, 2008, 24.2: 141-151.

[3] Nafornita, Corina, et al. "Kullback-Leibler distance between complex generalized Gaussian distributions." 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO). IEEE, 2012.

[4] Bauckhage, Christian. "Computing the kullback-leibler divergence between two weibull distributions." arXiv preprint arXiv:1310.3713 (2013).

[5] Di Crescenzo, Antonio, and Maria Longobardi. "Some properties and applications of cumulative Kullback-Leibler information." Applied stochastic models in business and industry 31.6 (2015): 875-891.

[6] Soch, Joram, and Carsten Allefeld. "Kullback-leibler divergence for the normal-gamma distribution." arXiv preprint arXiv:1611.01437 (2016).

[7] Frank Nielsen ," On the Kullback-Leibler divergence between discrete normal distributions "Sony Computer Science Laboratories Inc.Tokyo, Japan (2021).

[8] Zhang, Yufeng, et al. "On the properties of kullback-leibler divergence between multivariate gaussian distributions." Advances in Neural Information Processing Systems 36 (2023): 58152-58165.