RESEARCH ARTICLE

Adaptive Learning Density Estimators for Tsallis Entropy and Kapur Entropy with Applications in System Training

Vijay Kumar[1,*], **Arti Saxena**[1] **and Leena Chawla**[2]

[1] Manav Rachna International Institute of Research and Studies, Faridabad, Haryana 121010, India
[2] DPG Institute of Technology and Management, Gurugram, Haryana, India

Abstract

Adaptation learning is a data-driven technique that gives instructions based on the experiences made during data analysis. It plays an integral role in providing engineering solutions based on specific needs. Researchers have used the second-order statistics criterion for decades to conceptualize the optimality criteria using Shannon and Renyis information-theoretic measures. Some gaps have been identified in this research work, and useful findings have been proved with generalized information-theoretic measures of Renyis as Tsallis entropy of order α and Kapur entropy of order α and type β using the Parzen-Rosenblatt window. This work explored the problem of constructing kernel density estimators and their application in adaptive systems training.

Keywords: generalized entropies, adaptive learning, machine learning, kernel density estimation, information theoretic measures.

Submitted: 14 November 2025
Accepted: 23 November 2025
Published: 06 January 2026

Vol. 2, **No.** 1, 2026.
10.62762/TMI.2025.317970

***Corresponding author:**
✉ Vijay Kumar
drvijaykumarsudan@gmail.com

1 Introduction

In today's world, information means knowledge, ideas, opinions, emotions, feelings, errors, experiences, etc., which is either a content of direct or indirect observation. Information is indeed an asset, and it grows at an exponential pace. Year by year data has been grown in abundance but the distillation of information from the data is the main concern and a significant problem in information processing. This allows us to design optimal data processing system to extract information from data. The increasingly complex situations will increase the number of uncertain phenomena and the uncertainties about each phenomenon always tend to increase. To decrease uncertainty, information is collected, but that is uncertain in this world, too. Uncertainty is modeled through PDF or PMF, but working with functions takes a lot of work. Therefore, PDF can be synthesized through statistical descriptors, such as statistical moments. When the information is available underlying distribution, it is easy to construct learning systems that use the PDF(or joint PDF, marginal PDF, or conditional PDF) of the data obtained from the joint distributions but main challenges are data smoothing to process vast volumes of data. Kernel smoothing technique is the prominent non-parametric technique in which a PDF can be estimated to explore useful patterns in the data, while ignoring immaterial information. These problems have applications in training adaptive systems to find optimality criterion

Citation
Kumar, V., Saxena, A., & Chawla, L. (2026). Adaptive Learning Density Estimators for Tsallis Entropy and Kapur Entropy with Applications in System Training. *ICCK Transactions on Machine Intelligence*, 2(1), 12–27.

while designing fast and accurate algorithms that speed up data processing. To solve classification problems for feature reduction, these algorithms have been used. Therefore, information-theoretic criterion measures have been required that lower down the uncertainty in the system. Wiener [1] used second-order statistics to take up training in adaptive systems, but second-order statistics is not sufficient to deal with the problems of optimality criterion. In adaption and learning, the objective is to explore entropy estimators that estimates quantity and optimize parameters. In engineering applications, entropy estimation from real data is a non-trivial problem. Information theory provides a theoretical framework that helps design and optimizes learning algorithms. Many researchers used generalized Shannon [2] entropy and Renyi [3] entropy to derive estimators with applications in training adaptation systems and learning. Therefore, new non-parametric estimators for Tsallis [4] entropy and Kapur [5] entropy have been proposed, which have applications in designing learning algorithms to train adaptive systems.

2 Literature Review

Entropy as a learning criterion was proposed by Barlow et al. [6] for various feature-learning algorithms proposed by Atick [7], Intrator [8], Olshausen et al. [9]. Univariate and multivariate PDFs-based Shannon entropy expressions were discussed by Lazo et al. [10] and Ahmed et al. [11]. Shannon [2] proposed a measure of uncertainty, which is additive in nature. In 1961, Renyi [3] generalized the idea of Shannon [2] and proposed a non-additive measure of entropy to calculate the mean using the expectation of a recursive estimator. Other forms of information-theoretic measures appeared from the works of entropy [12, 13] that give strength to the field of information theory to grow. Over the last two decades, information theoretic measures have been used in learning problem to improve performance and determine the information extracted from the data [14]. Researchers [15–19] further used Renyi entropy and proposed various measures. A group of Researchers [14, 20, 21], used the applications of Renyis entropy to solve the problem of machine learning, such as, dimensionality reduction, feature extraction, blind source separation, etc. Principle [20] introduced information theoretic learning into adaptive systems. k-NN technique has limited scope due to its poor performance, whereas, Parzen window method performs slightly better in comparison, but it is quite challenging to implement due to discontinuities. These limitations can be addressed by using a smooth kernel function. Researchers [22–30] described the classical kernel density and described different facets of KDE and its practical importance. Non-parametric power system security risk assessment model was proposed by Ul Hassan et al. [31]. They used Parzen window density estimation and obtained PDFs for power systems to estimate PDF using KDE. Kernel density estimators are widely used to estimate entropy [32] because of easy implementation, computationally faster, and simple to understand [33].

Shannon and Renyi entropy-based entropy measures occupied a lot of research, but many entropy estimators of generalized non-additive entropy measures need to be addressed. These entropy estimators may be simple to implement and have applications in the optimization of parameters of adaptive systems. The main concern is to test the consistency properties of the proposed estimator, which is the primary concern discussed in this paper and establishes a solid theory behind ITL using Tsallis and Kapur entropy of order α and type β.

3 Information Theoretic Learning and Machine Learning

Entropy is a scalar descriptor that quantifies PDF to identify optimal design goals. To estimate entropy and mutual information, non-parametric estimators of entropy and mutual information have been predicted using information measures. In this research work, Tsallis [4] and Kapur [5] contributions have been used to estimate kernel density function that have applications in training adaptive systems. An adaptive algorithm changes its behavior during the run time based on the available information in an a priori defined reward criterion.

Let t_i be an output produced from the input data points x_i. A learning machine having a set of free parameters χ, constructed to get input from the data sources to produce output. Comparing that how similar t_i is to Δ_i so that their difference is minimum by changing the parameters χ with some systematic approach according to some criterion. Finally, a system is activated that approximates the unknown z with y, when new data point is obtained from the data source x. In this way, a model/system has been build that established relationship between x and z. The system is called a *adaptive system* or *learning machine* that finds parameters from data, learning or adaption. This is the idea behind training adaptive systems as shown in Figure 1, which is known as supervised learning. The

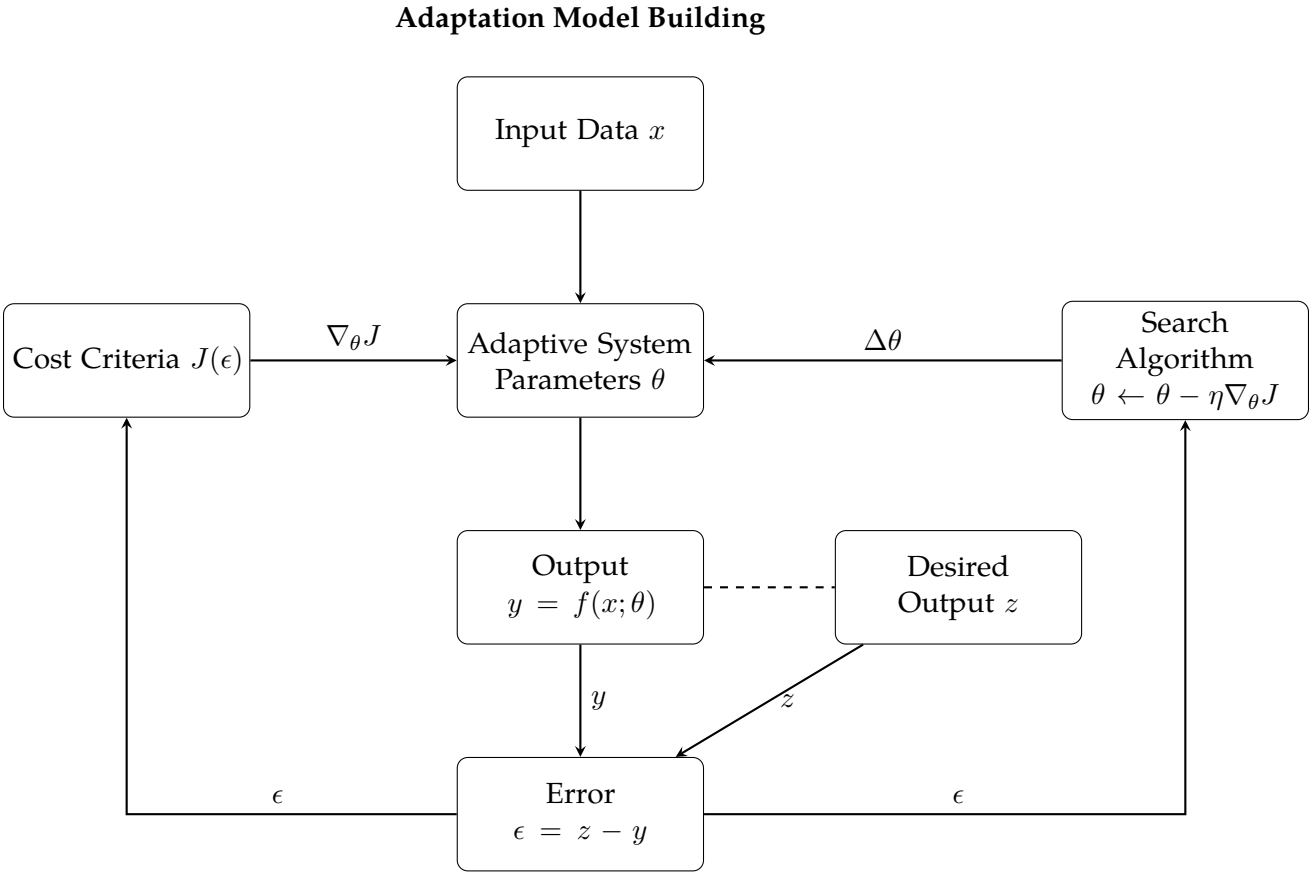

Figure 1. Adaptation model building.

learning in which data source x is available but z is not available is called unsupervised learning.

4 Preliminaries

4.1 Shannon Measure of Entropy

The logarithmic measure of information for an experiment X denoted by $H^S(X)$ as given by Shannon [2] is defined as:

$$H^S(X) = -\sum_{i=1}^{n} f_i \log(f_i)$$

In case of continuous random variable X, the Shannon entropy is defined as:

$$H^S(X) = -\int_{-\infty}^{\infty} w(x) \log(w(x))\, dx$$

where, w is a PDF of random variable X.

Shannon entropy is additive in nature, arises from statistical concepts and fundamental from application point of view.

Shannon entropy has generalizations such as: Renyi entropy [3], Kapur entropy [5], Tsallis entropy [4], Arimoto entropy [34], Havrda entropy [12], etc.

4.2 Renyi Entropy

Renyi [3] proposed a measure of entropy and is defined as:

$$H_\alpha^R(X) = \frac{1}{1-\alpha} \log\left(\sum_{j=1}^{n} w_j^\alpha\right); \quad \alpha \neq 1, \alpha > 0,$$

is called the Renyis entropy of order α.

In case of continuous random variable X, Renyi entropy is defined as:

$$H_\alpha^R(X) = \frac{1}{1-\alpha} \log\left(\int_{-\infty}^{\infty} w^\alpha(x)\, dx\right),$$

where, $w(x)$ is a PDF of random variable X.

As $\alpha \to 1$, $H_\alpha^R(X) \to H^S(X)$.

In other words, Renyi entropy is the Shannon entropy of order 1.

The information potential of Renyi entropy is denoted as $IP_\alpha^R(X)$ and is defined as:

$$IP_\alpha^R(X) = \int_{-\infty}^{\infty} w^\alpha(x)\, dx.$$

or equivalently:

$$H_\alpha^R(X) = \frac{1}{1-\alpha} \log\left(IP_\alpha^R(X)\right).$$

4.3 Tsallis Entropy

Tsallis [4] introduced an entropic expression identical to Havrda & Charvat [12] in the generalization of standard statistical mechanics and was of non-additive nature. Let $w : \mathbb{R}^n \to \mathbb{R}$ be the PMF of a discrete random variable X, the Tsallis entropy is defined as:

$$H_\alpha^T(X) = \frac{1}{1-\alpha}\left(\sum_{i=1}^{n} w_i^\alpha - 1\right)$$

where, $\alpha \in \mathbb{R}^+$ is referred to non-extensive index or entropy index and characterizes the degree of non-linearity. Tsallis entropy is referred as α-statistic.

As $\alpha \to 1$, $H_\alpha^T(X) \to H_\alpha^R(X)$.

In case of continuous random variable X, the Tsallis entropy is defined as:

$$H_\alpha^T(X) = \frac{1}{1-\alpha}\left(\int_{-\infty}^{\infty} w^\alpha(x)\, dx - 1\right)$$

where, $w(x)$ is a PDF of random variable X.

The information potential of Tsallis entropy is denoted as $IP_\alpha^T(X)$ and is defined as:

$$IP_\alpha^T(X) = \int_{-\infty}^{\infty} w^\alpha(x)\, dx - 1$$

$$H_\alpha^T(X) = \frac{1}{1-\alpha} IP_\alpha^T(X)$$

Tsallis entropy with expectation operator is written as:

$$\begin{aligned} H_\alpha^T(Y) &= \frac{1}{1-\alpha} \log\left(\int_{-\infty}^{\infty} w_Y^\alpha(y)\, dy\right) \\ &= \frac{1}{1-\alpha} \log\left(E_Y\left[w_Y^{\alpha-1}(y)\right]\right) \end{aligned}$$

Using Parzen window estimator, the non-parametric estimator of Tsallis entropy is given as:

$$\widehat{H}_\alpha^T(Y) = \frac{1}{1-\alpha} \log\left(\frac{1}{N^\alpha}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} \Phi_j(y_i - y_j)\right)^{\alpha-1}\right)$$

4.4 Kapur Measure of Entropy

Kapur [5] generalized Renyis entropy for an incomplete probability distribution as:

$$H_{\alpha,\beta}^K(X) = \frac{1}{\beta-\alpha} \log\left[\frac{\sum_{j=1}^n w_j^\alpha}{\sum_{j=1}^n w_j^\beta}\right]$$

where, $w_i \geq 0, \sum_{j=1}^n w_i \leq 1, \alpha \neq \beta, \alpha, \beta > 0$, is called Kapur entropy of order α and type β.

As $\alpha \to 1, H_{\alpha,\beta}^K(X) \to H_\alpha^R(X)$ and as $\alpha \to 1, \beta = 1, H_{\alpha,\beta}^K(X) \to H^S(X)$.

In case of continuous random variable X, the Kapur entropy of order α and type β is defined as:

$$H_{\alpha,\beta}^K(X) = \frac{1}{\beta-\alpha} \log\left[\frac{\int_{-\infty}^{\infty} w^\alpha(x)\, dx}{\int_{-\infty}^{\infty} w^\beta(x)\, dx}\right]$$

where, $w(x)$ is a PDF of random variable X.

The information potential of Kapur entropy is denoted as $IP_\alpha^K(X)$ and $IP_\beta^K(X)$ and is defined as:

$$IP_\alpha^K(X) = \int_{-\infty}^{\infty} w^\alpha(x)\, dx$$

$$\text{and} \quad IP_\beta^K(X) = \int_{-\infty}^{\infty} w^\beta(x)\, dx.$$

$$H_{\alpha,\beta}^K(X) = \frac{1}{\beta-\alpha}\left[\log\left(IP_\alpha^K(X)\right) - \log\left(IP_\beta^K(X)\right)\right]$$

4.5 Window Function

For a hypercube of unit length centered at origin, the window function is defined as:

$$\Theta(y) = \begin{cases} 1, & |u_i| \leq \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}; \quad (i = 1, 2, \ldots, d)$$

The generalization of the window function is known as the Parzen window, a technique to estimate density function. This is a non-parametric density estimation technique, defined as:

$$P_m(y) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{h^d} f\left(\frac{y_j - y}{h^m}\right)$$
$$\text{provided} \quad f\left(\frac{y - y_j}{h^m}\right) = k$$

where m, h, f and $p(y)$ are the number of elements, dimension, window function and probability density of y, respectively.

Window width and kernel are the two critical parameters of Parzen window. Let $\{y_1, \ldots, y_N\}$ be the N samples drawn from the random variable. These samples are independent and identically distributed (i.i.d). The kernel function (arbitrary) $K_\sigma(\cdot)$ estimate of the PDF is given by Parzen [35], and is defined as:

$$\hat{f}_r(y) = \frac{1}{N}\sum_{j=1}^{N}\Phi_\sigma(y - y_j)$$

Parzen window function is used to propose Tsallis entropy estimator of order α and Kapur entropy estimator of order α and type β.

4.6 Kernel

A window function is fitted on each data point to determine the fraction of the data points used for the density estimation within the window. Choosing fixed kernel bandwidth may cover very little observation in low-density regions, while extensive observations may cover high-density areas. To deal with such situations, adaptive or variable kernel bandwidth approaches that vary the bandwidth from one to another has been applicable. To approximate the probability distribution of the given data without defined distribution, kernel density estimation algorithm is used in this work.

4.7 Kernel Density Estimation (KDE)

To estimate the density function from the underlying data, KDE can be used to understand the topology of the density. The first multivariate non-parametric density estimator is defined as:

$$\hat{f}(y) = \frac{1}{nh^d}\sum_{j=1}^{N}\Phi\left(\frac{y_j - y}{h}\right);$$
$$x_j \in \mathbb{R}^d \text{ and } \Phi : \mathbb{R}^d \to \mathbb{R}^1,$$

where $\{y_j\}$ is an i.i.d random sample of size n and h is the smoothing parameter.

As $n \to \infty$, $h \to 0$.

In many cases, h is treated as constant and to improve the result of the density estimate, h is treated as a variable and the following kernel estimators are given as:

For uniform density, the kernel estimator [36] is defined as:

$$\hat{f}(y) = \frac{1}{nh_k(y)^d}\sum_{j=1}^{N}\Phi\left(\frac{y_j - y}{h_k(y)}\right).$$

For non-parametric density, kernel estimator [37] is defined as:

$$\hat{f}_h(y) = \frac{1}{n}\sum_{j=1}^{n}\Phi_h(y_j - y) = \frac{1}{n}\sum_{j=1}^{n}\Phi\left(\frac{y_j - y}{h}\right),$$

where $\Phi > 0$ is the kernel function with h as smoothing parameter.

However, the estimator independent of the observations is called generalized kernel estimator.

4.8 Consistency of the Kernel Density Estimator

Rosenblatt [38] proposed kernel-based estimator whose idea was extended by researchers [39–41]. Convergence rates were studied by Stute [42, 43] and observed that they depends on the sample size, density, and kernel and obtained some valuable results on the convergence rates of the estimator as: The generalization of empirical processes has been proposed by Einmahl et al. [44], in which the authors proved the convergence through mathematical techniques by considering bandwidth as a variable within a small interval.

Weak Consistency.

According to Wied et al. [45], let us consider that

$$\lim_{n\to\infty} nh_n = \infty; \quad \forall\, n \in \mathbb{N}$$

At the point y, f is continuous, the estimator $f_n(y)$ is weakly consistent in probability space P, i.e.,

$$\lim_{n\to\infty} P\left(|f_n(y) - f(y)| > \epsilon\right) = 0; \quad \forall\, \epsilon > 0$$

Strong consistency.

Nadaraja [46] formulated the theorem for kernel density estimator of almost sure uniform convergence, which is the extension of the Parzen [35].

4.9 Entropy Estimation for Learning

From the literature it has been revealed that Shannon entropy measure attracts the attention of the researchers to develop algorithms for learning systems. Estimating Shannon entropy has been applied to the generalized Shannon entropy such as Renyi entropy. This work proposes to expedite the generalized Renyi's entropy to propose simple entropy estimator for training learning systems using Tsallis entropy of order α and Kapur entropy of order α and type β. The main objective of the work is to understand the mathematical insights of the concept.

4.10 Proposed Kernel Estimator for Tsallis Entropy and Kapur Entropy

$$\widehat{H}_{\alpha}^{T}(y) = \frac{1}{1-\alpha} \cdot \frac{1}{N^{\alpha}} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \Phi_{\lambda}(y_i - y_j) \right) - \frac{1}{1-\alpha}.$$

$$\widehat{H}_{\alpha,\beta}^{K}(Y) = \frac{1}{\beta-\alpha} \log \left(\frac{\frac{1}{N^{\alpha}} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \Phi_{\lambda}(y_i - y_j) \right)^{\alpha-1}}{\frac{1}{N^{\beta}} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \Phi_{\lambda}(y_i - y_j) \right)^{\beta-1}} \right).$$

5 Terminologies

i. Component of Y: $Y = \{y_1, \ldots, y_N\}$ with $Y^0 \to 0^n$.

ii. Single dimensional kernel: $\Phi_2^0(\cdot)$.

iii. Multidimensional Kernel for joint PDF: $\Phi_{\Sigma}^0(\cdot)$.

iv. Parzen estimate for joint PDF:

$$f_Y^n(y) = \frac{1}{N} \sum_{j=1}^{N} \Phi_{\Sigma}(y - y_j).$$

v. Parzen estimate for marginal density of Y^0:

$$f_{Y^0}^n(y^0) = \frac{1}{N} \sum_{j=1}^{N} \Phi_{\lambda_0}^0(y^0 - y_j^0).$$

6 Main Results

Information-theoretic measures play a vital role in understanding the uncertainty of the system. This work uses the Tsallis entropy of order α and Kapur entropy of order α and type β with the Parzen window function to introduce kernel estimators. Various theorems (I-IV) and the properties (I-III) have been proposed for the entropy estimators (4.10) and are discussed as follows.

Theorem 6.1 (I). ***Statement:*** *Consider that Parzen windowing is consistent with sample mean, the proposed entropy estimators (4.10) are consistent underlying PDF of linearly independent samples.*

Proof:

Let $Z_1, Z_2, \ldots, Z_N$ be the N samples drawn from independent and identically distributed (i.i.d) with sample means

$$\overline{Z_1}, \overline{Z_2}, \ldots, \overline{Z_N}$$

drawn from independent density function.

According to Parzen [35], the consistency of the estimate in the estimation of PDF:

As $N \to \infty$,

$$\overline{Z_k} \to E[Z]$$

where N, $\overline{Z_k}$ and $E[Z]$ represent the number of samples, sample mean and expected value respectively.

Theorem 6.1 discusses the non-asymptotic nature of the Tsallis entropy of order α and Kapur entropy of order α and type β estimator.

In learning and adaptive system problems, finite number of samples are given that provide consistent estimates of the entropy until the global optimum is received in the desired solution.

Theorem 6.2 (II). ***Statement:*** *For equal samples: $z_i = z_j$, the proposed entropy estimators (4.10) achieve their minimum and maximum value when the kernel function is evaluated at $\Phi_{\lambda}(0)$.*

Proof. **For Tsallis entropy:**

The Tsallis entropy estimator is given by:

$$\hat{H}_{\alpha}^{T}(z) = \frac{1}{1-\alpha} \cdot \frac{1}{N^{\alpha}} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \Phi_{\lambda}(z_i - z_j) \right)^{\alpha-1} - \frac{1}{1-\alpha}$$

For equal samples: $z_i = z_j$, we have:

$$\hat{H}_{\alpha}^{T}(z) = \frac{1}{1-\alpha} \cdot \frac{1}{N^{\alpha}} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \Phi_{\lambda}(0) \right)^{\alpha-1} - \frac{1}{1-\alpha}$$

Simplifying:

$$\hat{H}_{\alpha}^{T}(z) = \frac{1}{1-\alpha} \cdot \Phi_{\lambda}(0) - \frac{1}{1-\alpha}$$

To prove that the proposed Tsallis entropy estimator attains its minimum, we need to show that:

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \geq N^\alpha(-\Phi_\lambda(0)+1)(1-\alpha)+N^\alpha$$

Case I: For $\alpha > 1$

We have:

$$\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \leq \Phi_\lambda(0)$$

Since:

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \leq N\max_i\left[\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right]$$

Applying this upper bound:

$$\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \leq \frac{1}{N^{\alpha-1}}\max_i\left[\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right]$$

For equal samples $z_j = z_i$:

$$\leq \max_{i,j}\Phi_\lambda(z_i-z_j)^{\alpha-1} = \Phi_\lambda(0)^{\alpha-1}$$

Case II: For $\alpha < 1$,

$$\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \geq N\min_i\left[\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right]$$

Dividing both sides by N^α:

$$\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \geq \frac{1}{N^{\alpha-1}}\min_i\left[\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right]$$

For identical samples $z_j = z_i$, this simplifies to:

$$\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1} \geq \min_{i,j}\left[\Phi_\lambda^{\alpha-1}(z_i-z_j)\right]$$

□

For Kapur Entropy:

$$\hat{H}^K_{\alpha,\beta}(z) = \frac{1}{\beta-\alpha}\left[\log\left(\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right) - \log\left(\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}\right)\right]$$

For equal samples: $z_i = z_j$,

$$\hat{H}^K_{\alpha,\beta}(z) = \frac{1}{\beta-\alpha}\left[\log\left(\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(0)\right)^{\alpha-1}\right) - \log\left(\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(0)\right)^{\beta-1}\right)\right]$$

To prove that the proposed Kapur entropy estimator is minimum, we shall show that:

$$\frac{1}{\beta-\alpha}\left[\log\left(\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right) - \log\left(\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}\right)\right] \geq -\log\Phi_\lambda(0)$$

Case I: For $\alpha > \beta$,

$$\Rightarrow \frac{\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}} \leq \Phi_\lambda^{\alpha-\beta}(0)$$

$$\Rightarrow \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}} \leq N^{\alpha-\beta}\Phi_\lambda^{\alpha-\beta}(0)$$

$$\Rightarrow \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}} \leq \frac{N\max_i\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{N\max_i\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}}$$

Since LHS is replaced with the upper bound.

$$\leq \frac{\max_j\left(N^{\alpha-1}\max_i\left(\Phi_\lambda^{\alpha-1}(z_i-z_j)\right)\right)}{\max_j\left(N^{\beta-1}\max_i\left(\Phi_\lambda^{\beta-1}(z_i-z_j)\right)\right)}$$
$$= N^{\alpha-\beta}\max_{i,j}\left(\Phi_\lambda^{\alpha-\beta}(z_i-z_j)\right)$$
$$= N^{\alpha-\beta}\Phi_\lambda^{\alpha-\beta}(0).$$

Case II: For $\alpha < \beta$,

$$\Rightarrow \frac{\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}} \geq \Phi_\lambda^{\alpha-\beta}(0)$$
$$\Rightarrow \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}} \geq N^{\alpha-\beta}\Phi_\lambda^{\alpha-\beta}(0).$$

$$\Rightarrow \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}} \geq \frac{N\max_i\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{N\max_i\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}}.$$

Since LHS is replaced with the upper bound:

$$\geq \frac{\min_j\left(N^{\alpha-1}\min_i\Phi_\lambda^{\alpha-1}(z_i-z_j)\right)}{\min_j\left(N^{\beta-1}\min_i\Phi_\lambda^{\beta-1}(z_i-z_j)\right)}$$
$$\geq N^{\alpha-\beta}\min_{i,j}\Phi_\lambda^{\alpha-\beta}(z_i-z_j)$$
$$\geq N^{\alpha-\beta}\Phi_\lambda^{\alpha-\beta}(0).$$

Theorem 6.2 is applicable in training supervised learning models when all the samples are equal or the error in samples is zero, the cost function attains its global minimum.

Based on the Theorems 6.1 and 6.2, following properties are discussed as:

Property 6.1 (I). ***Statement:*** *Consider the density function of the samples, the proposed entropy estimators (4.10) are invariant to the mean of the given density function underlying actual entropy* [47, 48].

Proof. Let us consider that Z and $\widehat{Z}$ be two random variables in which $\widehat{Z} = Z + m$ with $m \in \mathbb{R}$.

For Tsallis Entropy:

$$H_\alpha^T(\widehat{Z}) = \frac{1}{1-\alpha}\left[\int f_Z^\alpha(z)\,dz - 1\right]$$
$$= \frac{1}{1-\alpha}\left[\int f_Z^\alpha(z+m)\,dz - 1\right]$$
$$= \frac{1}{1-\alpha}\left[\int f_Z^\alpha(z)\,dz - 1\right]$$
$$= H_\alpha^T(Z).$$

For Kapur Entropy:

$$H_{\alpha,\beta}^K(\widehat{Z}) = \frac{1}{\beta-\alpha}\left[\log\left(\int f_Z^\alpha(z)\,dz\right) - \log\left(\int f_Z^\beta(z)\,dz\right)\right]$$
$$= \frac{1}{\beta-\alpha}\left[\log\left(\int f_Z^\alpha(z+m)\,dz\right) - \log\left(\int f_Z^\beta(z+m)\,dz\right)\right]$$
$$= \frac{1}{\beta-\alpha}\left[\log\left(\int f_Z^\alpha(z)\,dz\right) - \log\left(\int f_Z^\beta(z)\,dz\right)\right]$$
$$= H_{\alpha,\beta}^K(Z).$$

Let $\{z_1, z_2, \ldots, z_N\}$ be the samples of random variable Z and $\{z_1+m, z_2+m, \ldots, z_N+m\}$ are the samples of random variable $\widehat{Z}$.

(1) Tsallis entropy.

$$\hat{H}_\alpha^T(\bar{Z}) = \frac{1}{1-\alpha}\left[\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(\hat{z}_i - z_j)\right)^{\alpha-1} - 1\right]$$
$$= \frac{1}{1-\alpha}\left[\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i - z_j)\right)^{\alpha-1} - 1\right]$$
$$= \hat{H}_\alpha^T(Z)$$

(2) Kapur Entropy:

$$\begin{aligned}\hat{H}^K_{\alpha,\beta}(\bar{Z}) &= \frac{1}{\beta-\alpha}\left[\log\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(\bar{z}_i-z_j)\right)^{\alpha-1}\right.\\ &\qquad\left.-\log\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(\bar{z}_i-z_j)\right)^{\beta-1}\right]\\ &= \frac{1}{\beta-\alpha}\left[\log\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}\right.\\ &\qquad\left.-\log\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}\right]\\ &= \hat{H}^K_{\alpha,\beta}(Z)\end{aligned}$$

□

Property 6.2 (II). *With λ as the kernel size, if entropy can be estimated for samples $\{z_1,\ldots,z_N\}$ of a random variable Z, then to estimate the samples $\{cz_1,\ldots,cz_N\}$ of the scaled random variable cZ, a kernel size $|c|\lambda$ should be used.*

Proof. Let $cZ=\{cz_1,\ldots,cz_N\}$ be the scaled random variable.

(1) Tsallis entropy. The Tsallis entropy of Z is defined as

$$H^T_\alpha(Z)=\frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty}f^\alpha_Z(z)\,dz-1\right].$$

For the scaled variable cZ, we have

$$\begin{aligned}H^T_\alpha(cZ) &= \frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty}\frac{1}{|c|^\alpha}f^\alpha_Z\Big(\frac{z}{c}\Big)\,dz-1\right]\\ &= \frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty}f^\alpha_Z(t)\,dt-1\right],\qquad \Big(t=\frac{z}{c}\Big).\end{aligned}$$

Thus,

$$H^T_\alpha(cZ)=H^T_\alpha(Z).$$

(2) Kapur entropy. The Kapur entropy is defined as

$$H^K_{\alpha,\beta}(Z)=\frac{1}{\beta-\alpha}\left[\frac{\log\int f^\alpha_Z(z)\,dz}{\log\int f^\beta_Z(z)\,dz}\right].$$

For cZ, it follows that

$$\begin{aligned}H^K_{\alpha,\beta}(cZ) &= \frac{1}{\beta-\alpha}\left[\log\int_{-\infty}^{\infty}\frac{1}{|c|^\alpha}f^\alpha_Z\Big(\frac{z}{c}\Big)\,dz\right.\\ &\qquad\left.-\log\int_{-\infty}^{\infty}\frac{1}{|c|^\beta}f^\beta_Z\Big(\frac{z}{c}\Big)\,dz\right]\\ &= \frac{1}{\beta-\alpha}\Big[(1-\alpha)H^R_\alpha(Z)\\ &\qquad-(1-\beta)H^R_\beta(Z)\Big]+\log|c|,\end{aligned}$$

where H^R_α denotes the Rényi entropy and the substitution $t=z/c$ is used.

Therefore,

$$H^K_{\alpha,\beta}(cZ)=\begin{cases}H^R_\alpha(Z)+\log|c|, & \beta=1,\\ H^R_\beta(Z)+\log|c|, & \alpha=1.\end{cases}$$

From the above analysis, the kernel size must be scaled by a factor of $|c|$ to preserve the entropy estimation, which completes the proof. □

Theorem 6.3 (III). *The global minimum of the entropy estimators in (4.10) is smooth for a continuous, differentiable, symmetric, and unimodal kernel function $\Phi_\lambda(\cdot)$.*

Proof. To show that the proposed entropy estimators in (4.10) attain a global minimum, it is sufficient to show that the underlying Hessian matrix is semi-definite.

(1) Tsallis entropy. The Tsallis entropy estimator is given by

$$\hat{H}^T_\alpha(z)=\frac{1}{1-\alpha}\left[\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}-1\right].$$

Define the auxiliary variable

$$\hat{P}_\alpha=\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}.$$

Then the first- and second-order partial derivatives satisfy

$$\frac{\partial\hat{H}_\alpha}{\partial z_k}=\frac{1}{1-\alpha}\frac{\partial\hat{P}_\alpha}{\partial z_k},\qquad \frac{\partial^2\hat{H}_\alpha}{\partial z_\ell\partial z_k}=\frac{1}{1-\alpha}\frac{\partial^2\hat{P}_\alpha}{\partial z_\ell\partial z_k}.$$

At $\bar{z}=0$, the parameters reduce to

$$\hat{P}_\alpha\big|_{\bar{z}=0} = \Phi_\lambda^{\alpha-1}(0).$$

The first derivative is

$$\begin{aligned}\frac{\partial \hat{P}_\alpha}{\partial z_k}\bigg|_{\bar{z}=0} &= \frac{\alpha-1}{N^\alpha}\Big[N^{\alpha-1}\Phi_\lambda^{\alpha-2}(0)\Phi'(0) \\ &\quad - N^{\alpha-1}\Phi_\lambda^{\alpha-2}(0)\Phi'(0)\Big] \\ &= 0.\end{aligned}$$

The second derivatives are given by

$$\begin{aligned}\frac{\partial^2 \hat{P}_\alpha}{\partial z_k^2}\bigg|_{\bar{z}=0} &= \frac{(\alpha-1)(N-1)\Phi_\lambda^{\alpha-3}(0)}{N^2} \\ &\quad \times \left[(\alpha-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\right].\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \hat{P}_\alpha}{\partial z_\ell \partial z_k}\bigg|_{\bar{z}=0} &= -\frac{(\alpha-1)\Phi_\lambda^{\alpha-3}(0)}{N^2} \\ &\quad \times \left[(\alpha-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\right], \ell \neq k.\end{aligned}$$

Therefore, the Hessian matrix of $\hat{H}_\alpha$ at $\bar{z}=0$ satisfies

$$\frac{\partial^2 \hat{H}_\alpha}{\partial z_\ell \partial z_k}\Big|_{\bar{z}=0} = \begin{cases} -\frac{(N-1)\Phi_\lambda^{\alpha-3}(0)}{N^2}\Big[(\alpha-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\Big], & \ell = k, \\ \frac{\Phi_\lambda^{\alpha-3}(0)}{N^2}\Big[(\alpha-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\Big], & \ell \neq k. \end{cases}$$

Since $\Phi_\lambda(\cdot)$ is continuous, symmetric, unimodal, and twice differentiable, the Hessian matrix is semi-definite, which completes the proof.

(2) Kapur entropy.

The Kapur entropy estimator is given by

$$\hat{H}^K_{\alpha,\beta}(z) = \tfrac{1}{\beta-\alpha}\log\left[\frac{\frac{1}{N^\alpha}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\frac{1}{N^\beta}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\lambda(z_i-z_j)\right)^{\beta-1}}\right].$$

Define the auxiliary variables

$$\hat{IP}^K_\alpha = \frac{1}{N^\alpha}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\lambda(z_i-z_j)\right)^{\alpha-1},$$

$$\hat{IP}^K_\beta = \frac{1}{N^\beta}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\lambda(z_i-z_j)\right)^{\beta-1}.$$

Then the first- and second-order partial derivatives are

$$\frac{\partial \hat{H}^K_{\alpha,\beta}(z)}{\partial z_k} = \frac{1}{\beta-\alpha}\left(\frac{1}{\hat{IP}^K_\alpha}\frac{\partial \hat{IP}^K_\alpha}{\partial z_k} - \frac{1}{\hat{IP}^K_\beta}\frac{\partial \hat{IP}^K_\beta}{\partial z_k}\right),$$

$$\frac{\partial^2 \hat{H}^K_{\alpha,\beta}(z)}{\partial z_\ell \partial z_k} = \frac{1}{\beta-\alpha}\frac{\partial}{\partial z_\ell}\left(\frac{1}{\hat{IP}^K_\alpha}\frac{\partial \hat{IP}^K_\alpha}{\partial z_k} - \frac{1}{\hat{IP}^K_\beta}\frac{\partial \hat{IP}^K_\beta}{\partial z_k}\right).$$

At $\bar{z}=0$, the parameters reduce to

$$\hat{IP}^K_\alpha\big|_{\bar{z}=0} = \Phi_\lambda^{\alpha-1}(0), \qquad \hat{IP}^K_\beta\big|_{\bar{z}=0} = \Phi_\lambda^{\beta-1}(0).$$

The first-order derivatives vanish:

$$\begin{aligned}\frac{\partial \hat{IP}^K_\alpha}{\partial z_k}\bigg|_{\bar{z}=0} &= \frac{\alpha-1}{N^\alpha}\Big[N^{\alpha-1}\Phi_\lambda^{\alpha-2}(0)\Phi'(0) - N^{\alpha-1}\Phi_\lambda^{\alpha-2}(0)\Phi'(0)\Big] \\ &= 0.\end{aligned}$$

$$\begin{aligned}\frac{\partial \hat{IP}^K_\beta}{\partial z_k}\bigg|_{\bar{z}=0} &= \frac{\beta-1}{N^\beta}\Big[N^{\beta-1}\Phi_\lambda^{\beta-2}(0)\Phi'(0) - N^{\beta-1}\Phi_\lambda^{\beta-2}(0)\Phi'(0)\Big] \\ &= 0.\end{aligned}$$

The second-order derivatives are given by

$$\begin{aligned}\frac{\partial^2 \hat{IP}^K_\alpha}{\partial z_k^2}\bigg|_{\bar{z}=0} &= \frac{(\alpha-1)(N-1)\Phi_\lambda^{\alpha-3}(0)}{N^2} \\ &\quad \times \left[(\alpha-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\right].\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \hat{IP}^K_\beta}{\partial z_k^2}\bigg|_{\bar{z}=0} &= \frac{(\beta-1)(N-1)\Phi_\lambda^{\beta-3}(0)}{N^2} \\ &\quad \times \left[(\beta-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\right].\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \hat{IP}^K_\alpha}{\partial z_\ell \partial z_k}\bigg|_{\bar{z}=0} &= -\frac{(\alpha-1)\Phi_\lambda^{\alpha-3}(0)}{N^2} \\ &\quad \times \left[(\alpha-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\right], \ell \neq k.\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \hat{IP}^K_\beta}{\partial z_\ell \partial z_k}\bigg|_{\bar{z}=0} &= -\frac{(\beta-1)\Phi_\lambda^{\beta-3}(0)}{N^2} \\ &\quad \times \left[(\beta-2)\Phi'^2(0) + 2\Phi(0)\Phi''(0)\right], \ell \neq k.\end{aligned}$$

which shows that

$$\frac{\partial^2 \hat{H}^K_{\alpha,\beta}(z)}{\partial z_\ell \partial z_k}\Big|_{\bar{z}=0} = \begin{cases} \frac{1}{N^2}\begin{bmatrix} -(N-1)\Phi_\lambda^{\alpha-3}(0)\Big((\alpha-2)\Phi'^2(0)+2\Phi(0)\Phi''(0)\Big) \\ +(N-1)\Phi_\lambda^{\beta-3}(0)\Big((\beta-2)\Phi'^2(0)+2\Phi(0)\Phi''(0)\Big)\end{bmatrix}, & \ell = k, \\ \frac{1}{N^2}\begin{bmatrix}\Phi_\lambda^{\alpha-3}(0)\Big((\alpha-2)\Phi'^2(0)+2\Phi(0)\Phi''(0)\Big) \\ -\Phi_\lambda^{\beta-3}(0)\Big((\beta-2)\Phi'^2(0)+2\Phi(0)\Phi''(0)\Big)\end{bmatrix}, & \ell \neq k. \end{cases}$$

Equivalently,

$$\frac{\partial^2 \hat{H}^K_{\alpha,\beta}(z)}{\partial z_\ell \partial z_k}\Big|_{\bar{z}=0} = \begin{cases} \frac{1}{N^2}\begin{bmatrix}(N-1)\Phi'^2(0)\Big((\beta-2)\Phi_\lambda^{\beta-3}(0)-(\alpha-2)\Phi_\lambda^{\alpha-3}(0)\Big) \\ +(N-1)2\Phi(0)\Phi''(0)\Big(\Phi_\lambda^{\beta-3}(0)-\Phi_\lambda^{\alpha-3}(0)\Big)\end{bmatrix}, & \ell = k, \\ \frac{1}{N^2}\begin{bmatrix}\Phi'^2(0)\Big((\alpha-2)\Phi_\lambda^{\alpha-3}(0)-(\beta-2)\Phi_\lambda^{\beta-3}(0)\Big) \\ +2\Phi(0)\Phi''(0)\Big(\Phi_\lambda^{\alpha-3}(0)-\Phi_\lambda^{\beta-3}(0)\Big)\end{bmatrix}, & \ell \neq k. \end{cases}$$

The eigen-pairs of the Hessian matrix of both estimators in (4.10) are

$$\Big\{0,\ [1,\dots,1]^T\Big\}, \qquad \Big\{\frac{kN}{N-1},\ [1,-1,0,\dots,0]^T\Big\},$$
$$\Big\{\frac{kN}{N-1},\ [1,0,-1,0,\dots,0]^T\Big\}.$$

where k and ℓ denote the diagonal and off-diagonal entries of the Hessian matrix, respectively. □

Since every eigenvector corresponding to the unique nonzero eigenvalue changes the mean of the data, the Hessian matrix is positive semi-definite. From the results, it is concluded that the Hessian matrix is positive semi-definite provided that

$$\begin{cases} \Phi_\lambda(\cdot) > 0, \text{the eigen value is positive for } N > 1, \\ \Phi_\lambda(\cdot), \text{the nonzero eigenvalue has multiplicity } N > 1, \\ \Phi'_\lambda(\cdot) = 0 \quad \text{and} \quad \Phi''_\lambda(0) > 0. \end{cases}$$

In adaptive systems, the global optimum is characterized by a finite-eigenvalue Hessian matrix in the weight space with zero gradients. Therefore, the proposed entropy estimators in (4.10) attain a global minimum and are suitable for adaptive entropy minimization systems.

Property 6.3 (III)**.** *In the case of joint entropy estimation, let the multi-dimensional kernel function $\Phi_\Sigma(\cdot)$ and an orthonormal matrix R satisfy*

$$\Phi_\Sigma(\vartheta) = \Phi_\Sigma(R^{-1}\vartheta).$$

Then, the proposed entropy estimators in (4.10) are invariant under rotation.

Proof. Let the random vectors Z^n and $\bar{Z}^n$ be related by

$$\bar{Z} = RZ,$$

where $R \in \mathbb{R}^{n\times n}$ is a real orthonormal matrix satisfying $R^T R = I$.

Discrete case.

(1) Tsallis entropy. The Tsallis entropy estimator for the rotated samples $\bar{Z}$ is given by

$$\hat{H}^T_\alpha(\bar{Z}) = \frac{1}{1-\alpha}\left[\frac{1}{N^\alpha}\sum_{i=1}^N\Big(\sum_{j=1}^N \Phi_\Sigma(Rz_i - Rz_j)\Big)^{\alpha-1} - 1\right].$$

Using the kernel invariance property $\Phi_\Sigma(Rz_i - Rz_j) = \Phi_\Sigma(z_i - z_j)$, we obtain

$$\begin{aligned}\hat{H}^T_\alpha(\bar{Z}) &= \frac{1}{1-\alpha}\left[|R|^{1-\alpha}\frac{1}{N^\alpha}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\Sigma(z_i - z_j)\right)^{\alpha-1} - 1\right] \\ &= |R|^{1-\alpha}\big[\hat{H}^T_\alpha(Z)+1\big] - \frac{1}{1-\alpha}.\end{aligned}$$

Since R is orthonormal, $|R| = 1$, which yields

$$\hat{H}^T_\alpha(\bar{Z}) = \hat{H}^T_\alpha(Z).$$

(2) Kapur entropy. The Kapur entropy estimator for $\bar{Z}$ is given by

$$\hat{H}^K_{\alpha,\beta}(\bar{Z}) = \frac{1}{\beta-\alpha}\log\left[\frac{\frac{1}{N^\alpha}\sum_{i=1}^N\Big(\sum_{j=1}^N \Phi_\Sigma(Rz_i - Rz_j)\Big)^{\alpha-1}}{\frac{1}{N^\beta}\sum_{i=1}^N\Big(\sum_{j=1}^N \Phi_\Sigma(Rz_i - Rz_j)\Big)^{\beta-1}}\right].$$

Applying the kernel invariance again, we have

$$\begin{aligned}\hat{H}^K_{\alpha,\beta}(\bar{Z}) &= \frac{1}{\beta-\alpha}\left[\log\left(|R|^{1-\alpha}\frac{1}{N^\alpha}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\Sigma(z_i - z_j)\right)^{\alpha-1}\right)\right. \\ &\qquad \left. -\log\left(|R|^{1-\beta}\frac{1}{N^\beta}\sum_{i=1}^N\left(\sum_{j=1}^N \Phi_\Sigma(z_i - z_j)\right)^{\beta-1}\right)\right] \\ &= \frac{1}{\beta-\alpha}\Big[(1-\alpha)H^R_\alpha(Z) - (1-\beta)H^R_\beta(Z)\Big].\end{aligned}$$

Thus,

$$\hat{H}^K_{\alpha,\beta}(\bar{Z}) = \begin{cases} H^R_\alpha(Z), & \beta = 1, \\ H^R_\beta(Z), & \alpha = 1. \end{cases}$$

Therefore, the proposed entropy estimators are invariant under rotation, which completes the proof. □

Continuous case.

(1) Tsallis entropy. The Tsallis entropy of a continuous random variable Z is defined as

$$\hat{H}_\alpha^T(\bar{Z}) = \frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty} f_{\bar{Z}}^\alpha(\bar{z})\,d\bar{z} - 1\right].$$

Using the change of variables $\bar{z} = Rz$ with R being an orthonormal matrix, we have

$$\begin{aligned}\hat{H}_\alpha^T(\bar{Z}) &= \frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty} \frac{1}{|R|^\alpha} f_Z^\alpha(z)\,|R|\,dz - 1\right] \\ &= \frac{1}{1-\alpha}\left[|R|^{1-\alpha}\int_{-\infty}^{\infty} f_Z^\alpha(z)\,dz - |R|^{1-\alpha} \right. \\ &\quad \left. + |R|^{1-\alpha} - 1\right].\end{aligned}$$

Rearranging the terms yields

$$\begin{aligned}\hat{H}_\alpha^T(\bar{Z}) &= |R|^{1-\alpha}\frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty} f_Z^\alpha(z)\,dz - 1\right] \\ &\quad + |R|^{1-\alpha}\frac{1}{1-\alpha} - \frac{1}{1-\alpha} \\ &= |R|^{1-\alpha} H_\alpha^T(Z) + |R|^{1-\alpha}\frac{1}{1-\alpha} - \frac{1}{1-\alpha}.\end{aligned}$$

Hence,

$$\hat{H}_\alpha^T(\bar{Z}) = |R|^{1-\alpha}\left[H_\alpha^T(Z) + 1\right] - \frac{1}{1-\alpha}.$$

Since R is orthonormal, $|R| = 1$, which implies

$$\hat{H}_\alpha^T(\bar{Z}) = H_\alpha^T(Z).$$

(2) Kapur entropy. The Kapur entropy in the continuous case is given by

$$H_{\alpha,\beta}^K(\bar{Z}) = \frac{1}{\beta-\alpha}\log\left[\frac{\int f_{\bar{Z}}^\alpha(\bar{z})\,d\bar{z}}{\int f_{\bar{Z}}^\beta(\bar{z})\,d\bar{z}}\right].$$

Applying the same change of variables, we obtain

$$\begin{aligned}H_{\alpha,\beta}^K(\bar{Z}) &= \frac{1}{\beta-\alpha}\left[\log\left(|R|^{1-\alpha}\int_{-\infty}^{\infty} f_Z^\alpha(z)\,dz\right)\right. \\ &\quad \left. - \log\left(|R|^{1-\beta}\int_{-\infty}^{\infty} f_Z^\beta(z)\,dz\right)\right] \\ &= \frac{1}{\beta-\alpha}\Big[(1-\alpha)H_\alpha^R(Z) - (1-\beta)H_\beta^R(Z)\Big].\end{aligned}$$

Therefore,

$$H_{\alpha,\beta}^K(\bar{Z}) = \begin{cases} H_\alpha^R(Z) + \log|R|, & \beta = 1, \\ H_\beta^R(Z) + \log|R|, & \alpha = 1. \end{cases}$$

Since R is orthonormal and $|R| = 1$, the Kapur entropy is also invariant under rotation.

Theorem 6.4 (III)**.** *Let $\hat{Z}$ be a random variable with the PDF $f_{\hat{Z}}(\cdot) = f_Z(\cdot) * \Phi_\lambda(\cdot)$. Then:*

(i) *For the Tsallis entropy estimator,*

$$\lim_{N\to\infty} \hat{H}_\alpha^T(Z) = H_\alpha^T(\hat{Z}) \geq H_\alpha^T(Z).$$

Moreover, $\Phi_\lambda(\cdot) = 0$ if and only if

$$\lim_{N\to\infty} \hat{H}_\alpha^T(Z) = H_\alpha^T(\hat{Z}) = H_\alpha^T(Z).$$

(ii) *For the Kapur entropy estimator,*

$$\lim_{N\to\infty} \hat{H}_{\alpha,\beta}^K(Z) = H_{\alpha,\beta}^K(\hat{Z}) \geq H_{\alpha,\beta}^K(Z).$$

Moreover, $\Phi_\lambda(\cdot) = 0$ if and only if

$$\lim_{N\to\infty} \hat{H}_{\alpha,\beta}^K(Z) = H_{\alpha,\beta}^K(\hat{Z}) = H_{\alpha,\beta}^K(Z).$$

Proof. It is known that the Parzen estimate of the PDF of a random variable Z converges to $f_Z(\cdot) * \Phi_\lambda(\cdot)$ as $N \to \infty$.

(i) Tsallis entropy. The Tsallis entropy estimator $\hat{H}_\alpha^T(Z)$ converges to the true entropy of the estimated PDF. Recall that

$$\hat{H}_\alpha^T(Z) = \frac{1}{1-\alpha}\left[\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i - z_j)\right)^{\alpha-1} - 1\right].$$

To prove that $H_\alpha^T(\hat{Z}) \geq H_\alpha^T(Z)$, consider

$$H_\alpha^T(\hat{Z}) = \frac{1}{1-\alpha}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\big[f_Z(z-\tau)\big]^\alpha\,d\tau\,dz - 1\right].$$

Equivalently,

$$(1-\alpha)H_\alpha^T(\hat{Z}) + 1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\big[f_Z(z-\tau)\big]^\alpha\,d\tau\,dz.$$

Using Jensen's inequality, we distinguish two cases.

Case I: $\alpha > 1$.

$$\begin{aligned}(1-\alpha)H_\alpha^T(\hat{Z})+1 &\le \int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty} f_Z^\alpha(z-\tau)\,dz\right]d\tau \\ &\le \int_{-\infty}^{\infty}\Phi_\lambda(\tau)\,IP_\alpha^T(Z)\,d\tau \\ &\le IP_\alpha^T(Z).\end{aligned}$$

Case II: $\alpha < 1$.

$$\begin{aligned}(1-\alpha)H_\alpha^T(\hat{Z})+1 &\ge \int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty} f_Z^\alpha(z-\tau)\,dz\right]d\tau \\ &\ge \int_{-\infty}^{\infty}\Phi_\lambda(\tau)\,IP_\alpha^T(Z)\,d\tau \\ &\ge IP_\alpha^T(Z).\end{aligned}$$

Therefore, for both $\alpha > 1$ and $\alpha < 1$, it follows that

$$H_\alpha^T(\hat{Z}) \ge H_\alpha^T(Z),$$

with equality if and only if $\Phi_\lambda(\cdot) = 0$. □

(ii) Kapur entropy. The Kapur entropy estimator $\hat{H}_{\alpha,\beta}^K(Z)$ converges to the actual entropy of the estimated PDF. Recall that

$$\hat{H}_{\alpha,\beta}^K(Z) = \tfrac{1}{\beta-\alpha}\log\left(\frac{\frac{1}{N^\alpha}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\alpha-1}}{\frac{1}{N^\beta}\sum_{i=1}^{N}\left(\sum_{j=1}^{N}\Phi_\lambda(z_i-z_j)\right)^{\beta-1}}\right).$$

To prove that $H_{\alpha,\beta}^K(\hat{Z}) \ge H_{\alpha,\beta}^K(Z)$, consider the continuous form

$$H_{\alpha,\beta}^K(\hat{Z}) = \tfrac{1}{\beta-\alpha}\log\left(\frac{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\alpha d\tau\right]dz}{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\beta d\tau\right]dz}\right).$$

Equivalently, we may write

$$\exp\big((\beta-\alpha)H_{\alpha,\beta}^K(\hat{Z})\big) = \left(\frac{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\alpha d\tau\right]dz}{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\beta d\tau\right]dz}\right).$$

Using Jensen's inequality and noting that $\Phi_\lambda(\tau)$ is a valid probability kernel, we obtain

$$\begin{aligned}\exp\big((\beta-\alpha)H_{\alpha,\beta}^K(\hat{Z})\big) &\ge \frac{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty} f_Z^\alpha(z-\tau)\,dz\right]d\tau}{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty} f_Z^\beta(z-\tau)\,dz\right]d\tau} \\ &= \frac{IP_\alpha^T(Z)}{IP_\beta^T(Z)}.\end{aligned}$$

Taking the logarithm on both sides yields

$$H_{\alpha,\beta}^K(\hat{Z}) \ge H_{\alpha,\beta}^K(Z).$$

Equality holds if and only if $\Phi_\lambda(\cdot) = 0$, which completes the proof of (ii).

Case I: $\alpha > \beta$.

$$\begin{aligned}&\exp\big[(\beta-\alpha)\hat{H}_{\alpha,\beta}^K(\hat{Z})\big] \\ &\le \frac{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\alpha dz\right]d\tau}{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\beta dz\right]d\tau} \\ &\le \frac{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty} f_Z^\alpha(z-\tau)\,dz\right]d\tau}{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty} f_Z^\beta(z-\tau)\,dz\right]d\tau} \\ &\le \frac{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\,IP_\alpha^K(Z)\,d\tau}{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\,IP_\beta^K(Z)\,d\tau} \\ &\le \frac{IP_\alpha^K(Z)}{IP_\beta^K(Z)}.\end{aligned}$$

Case II: $\alpha < \beta$.

$$\exp\left[(\beta-\alpha)\hat{H}^K_{\alpha,\beta}(\hat{Z})\right] \geq \frac{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\alpha dz\right]d\tau}{\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[f_Z(z-\tau)\right]^\beta dz\right]d\tau}$$
$$\geq \frac{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty}f_Z^\alpha(z-\tau)\,dz\right]d\tau}{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\left[\int_{-\infty}^{\infty}f_Z^\beta(z-\tau)\,dz\right]d\tau}$$
$$\geq \frac{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\,IP^K_\alpha(Z)\,d\tau}{\int_{-\infty}^{\infty}\Phi_\lambda(\tau)\,IP^K_\beta(Z)\,d\tau}$$
$$\geq \frac{IP^K_\alpha(Z)}{IP^K_\beta(Z)}.$$

From the above results, it is concluded that the mean-invariant quantities $IP^T_\alpha(Z)$, $IP^K_\alpha(Z)$, and $IP^K_\beta(Z)$ correspond to the integrals of the α-th and β-th powers of the probability density function of Z.

Regardless of the values of α and β, the direction of the inequality is preserved, yielding

$$H^T_\alpha(\hat{Z}) \geq H^T_\alpha(Z), \qquad H^K_{\alpha,\beta}(\hat{Z}) \geq H^K_{\alpha,\beta}(Z).$$

Since

$$\mathbb{E}\left[\hat{f}_Z(\cdot)\right] = f_Z(\cdot) * \Phi_\lambda(\cdot),$$

the above results remain valid in the finite-sample setting.

Consequently, it is proven that, for a finite sample space, asymptotic noise is effectively rejected during the adaptation process. Therefore, Theorem 6.4 provides a theoretical foundation for information-theoretic, entropy-based adaptation criteria.

7 Results and Discussion

The Parzen window approach is employed to construct kernel density estimators for the Tsallis entropy of order α and the Kapur entropy of order α and type β. The proposed entropy estimators are used to optimize feature parameters, and the theoretical properties are established through Theorems 6.1–6.4 and Properties 6.1–6.3. The main findings are summarized as follows:

- **Non-asymptotic behavior:** The non-asymptotic nature of the proposed entropy estimators is established in Theorem 6.1, demonstrating that consistent entropy estimates are obtained before the global optimum is reached in the desired solution space.
- **Global optimality:** Theorem 6.2 proves that the cost function attains a global minimum, making the proposed entropy estimators applicable to the training of supervised learning models with equal sample representations.
- **Zero-gradient global minima:** In Theorem 6.3, the global minimum of the entropy estimator is shown to occur at zero gradients. This property is essential for achieving the global optimum in the weight space during adaptive learning.
- **Asymptotic noise rejection:** Theorem 6.4 establishes the asymptotic noise rejection capability of the proposed entropy estimators under entropy-based adaptation criteria, thereby demonstrating their effectiveness in statistical noise suppression for adaptive systems.

8 Conclusion

Kernel density estimation has been employed to construct entropy estimates directly from observed data points using kernel functions, providing an effective representation of the underlying data distribution. Each data point contributes to the density estimate through kernel-based weighting, resulting in a smooth and robust probability density approximation.

In this work, kernel density estimators based on the Parzen–Rosenblatt window have been developed for Tsallis entropy of order α and Kapur entropy of order α and type β. Theoretical analysis demonstrates that the proposed estimators possess desirable properties such as global optimality, rotation and scale invariance, and asymptotic noise rejection. These properties make the proposed entropy estimators well suited for training adaptive systems and entropy-based learning frameworks.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*. The MIT press. [CrossRef]

[2] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal, 27*(3), 379-423. [CrossRef]

[3] Rényi, A. (1961, January). On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics* (Vol. 4, pp. 547-562). University of California Press.

[4] Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of statistical physics, 52*(1), 479-487. [CrossRef]

[5] Kapur, J. N. (1969, April). Some properties of entropy of order α and type β. In *Proceedings of the Indian Academy of Sciences-Section A* (Vol. 69, No. 4, pp. 201-211). New Delhi: Springer India. [CrossRef]

[6] Barlow, H. B., Kaushal, T. P., & Mitchison, G. J. (1989). Finding minimum entropy codes. *Neural computation, 1*(3), 412-423. [CrossRef]

[7] Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing?. *Network: Computation in neural systems, 3*(2), 213-251. [CrossRef]

[8] Intrator, N. (1991). Feature extraction using an unsupervised neural network. In *Connectionist Models* (pp. 310-318). Morgan Kaufmann. [CrossRef]

[9] Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems, 7*(2), 333. [CrossRef]

[10] Lazo, A. V., & Rathie, P. (1978). On the entropy of continuous probability distributions (corresp.). *IEEE Transactions on Information Theory, 24*(1), 120-122. [CrossRef]

[11] Ahmed, N. A., & Gokhale, D. V. (1989). Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory, 35*(3), 688-692. [CrossRef]

[12] Havrda, J., & Charvat, F. (1967). Quantification method of classification processes. Concept of structural α-entropy. *Kybernetika, 3*(1), 30-35. [CrossRef]

[13] Kapur, J. N. (1994). *Measures of information and their applications*. Wiley. Retrieved from https://cir.nii.ac.jp/crid/1971430859819604036 (accessed on 05 January 2026).

[14] Principe, J. C., Xu, D., Zhao, Q., & Fisher Iii, J. W. (2000). Learning from examples with information theoretic criteria. *Journal of VLSI signal processing systems for signal, image and video technology, 26*(1), 61-77. [CrossRef]

[15] Sahoo, P., Wilkins, C., & Yeager, J. (1997). Threshold selection using Renyi's entropy. *Pattern recognition, 30*(1), 71-84. [CrossRef]

[16] Mittal, D. P. (1975). On additive and non-additive entropies. *Kybernetika, 11*(4), 271-276. [CrossRef]

[17] Campbell, L. L. (1965). A coding theorem and Rényi's entropy. *Information and control, 8*(4), 423-429. [CrossRef]

[18] Bassat, M. B., & Raviv, J. (1978). Renyi's entropy and the probability of error. *IEEE Transactions on Information Theory, 24*(3), 324-331. [CrossRef]

[19] Cachin, C. (1997). Smooth entropy and Renyi entropy. In *International Conference on the Theory and Applications of Cryptographic Techniques* (pp. 193-208). Springer. [CrossRef]

[20] Principe, J. C. (2010). *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media. [CrossRef]

[21] Xu, D. (1998). *Energy, entropy and information potential for neural computation*. University of Florida.

[22] Bowman, A. W., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations* (Vol. 18). OUP Oxford.

[23] Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons. [CrossRef]

[24] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge. [CrossRef]

[25] Härdle, W. (2004). *Nonparametric and semiparametric models*. Springer Science & Business Media. [CrossRef]

[26] Horová, I., Koláček, J., & Zelinka, J. (2012). *Kernel Smoothing in MATLAB: theory and practice of kernel smoothing*. World scientific. [CrossRef]

[27] Devroye, L., & Krzyzak, A. (1999). On the Hilbert kernel density estimate. *Statistics & Probability Letters, 44*(3), 299-308. [CrossRef]

[28] Duong, T., & Hazelton, M. L. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis, 52*(9), 4225-4242. [CrossRef]

[29] Lake, D. E. (2009). Nonparametric entropy estimation using kernel densities. *Methods in enzymology, 467*, 531-546. [CrossRef]

[30] Gyorfi, L., & Van der Meulen, E. C. (1991). On the nonparametric estimation of the entropy functional. In G. Roussas (Ed.), *Nonparametric Functional Estimation and Related Topics* (pp. 81-95). Springer. [CrossRef]

[31] Ul Hassan, R., Yan, J., & Liu, Y. (2022). Security risk assessment of wind integrated power system using Parzen window density estimation. *Electrical Engineering, 104*(4), 1997-2008. [CrossRef]

[32] Beirlant, J., Dudewicz, E. J., Gyorfi, L., & Van der Meulen, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of*

Mathematical and Statistical Sciences, 6(1), 17-39.

[33] Ahmad, I., & Lin, P. E. (1976). A nonparametric estimation of the entropy for absolutely continuous distributions. *IEEE Transactions on Information Theory, 22*(3), 372-375. [CrossRef]

[34] Arimoto, S. (1971). Information-theoretical considerations on estimation problems. *Information and control, 19*(3), 181-194. [CrossRef]

[35] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065-1076.

[36] Loftsgaarden, D. O., & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics, 36*(3), 1049-1051. [CrossRef]

[37] Breiman, L., Meisel, W., & Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics, 19*(2), 135-144.

[38] Davis, R. A., Lii, K. S., & Politis, D. N. (2011). Remarks on some nonparametric estimates of a density function. In *Selected Works of Murray Rosenblatt* (pp. 95-100). New York, NY: Springer New York. [CrossRef]

[39] Hardle, W. (1991). *Smoothing techniques: With implementation in S*. Springer. [CrossRef]

[40] Hall, P., Hu, T. C., & Marron, J. S. (1995). Improved variable window kernel estimates of probability densities. *The Annals of Statistics, 23*(1), 1-10.

[41] Wand, M. P., & Jones, M. C. (1994). *Kernel smoothing*. CRC Press.

[42] Stute, W. (1982). A law of the logarithm for kernel density estimators. *The Annals of Probability, 10*(2), 414-422.

[43] Stute, W. (1982). The oscillation behavior of empirical processes. *The Annals of Probability, 10*(1), 86-107.

[44] Einmahl, U., & Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics, 33*(3), 1380-1403. [CrossRef]

[45] Wied, D., & Weibbach, R. (2012). Consistency of the kernel density estimator: A survey. *Statistical Papers, 53*(1), 1-21. [CrossRef]

[46] Nadaraja, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications, 10*(1), 186-190. [CrossRef]

[47] Erdogmus, D., & Principe, J. C. (2001, September). Convergence analysis of the information potential criterion in adaline training. In *Neural Networks for Signal Processing XI: Proceedings of the 2001 IEEE Signal Processing Society Workshop* (IEEE Cat. No. 01TH8584) (pp. 123-132). IEEE. [CrossRef]

[48] Erdogmus, D., & Principe, J. C. (2001). An on-line adaptation algorithm for adaptive system training with minimum error entropy: Stochastic information gradient. In *Proceedings of the 2001 International Conference on Independent Component Analysis and Signal Separation* (pp. 7-12).

Dr. Vijay Kumar is working as a Professor of Mathematics, Manav Rachna International Institute of Research & Studies, Faridabad Haryana INDIA. His research interest includes Decision Making and Machine Learning. He has authored more than 60 research publications including research papers, conference papers and chapters in Books. He can be contacted at email: drvijaykumarsudan@gmail.com

Dr. Arti Saxena is working as associate professor in department of Applied Sciences, Faculty of Engineering and Technology, Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India. She has more than 40 research articles in various national and international journals of repute. Her areas of interest are Mathematical modeling, Fixed point Theory and Machine Learning.

Dr. Leena Chawla is working as assistant professor in DPG Institute of Technology and Management, Gurugram, Haryana, India. She has more than 10 years of teaching experience of graduates and postgraduates. She has more than 6 research articles published in journals.