Relationships between tail entropies and local intrinsic dimensionality and their use for estimation and feature representation

James Bailey[a], Michael E. Houle[b,a], Xingjun Ma[c]

[a]*The University of Melbourne, Melbourne, Victoria, 3010, Australia*
[b]*New Jersey Institute of Technology, Newark, NJ, 07102, USA*
[c]*Fudan University, Shanghai, 200438, China*

Abstract

The local intrinsic dimensionality (LID) model assesses the complexity of data within the vicinity of a query point, through the growth rate of the probability measure within an expanding neighborhood. In this paper, we show how LID is asymptotically related to the entropy of the lower tail of the distribution of distances from the query. We establish relationships for cumulative Shannon entropy, entropy power, Bregman formulation of cumulative Kullback-Leibler divergence, and generalized Tsallis entropy variants. Leveraging these relationships, we propose four new estimators of LID, one of them expressible in an intuitive analytic form. We investigate the effectiveness of these new estimators, as well as the effectiveness of entropy power as the basis for feature representations in classification.

Keywords: entropy, tail entropy, cumulative entropy, entropy power, intrinsic dimensionality, local intrinsic dimension, cumulative divergence, Bregman divergence

1. Introduction

Assessing the complexity of high dimensional data is a fundamental task that underpins many activities in machine learning and data mining. One well-known measure of data complexity is the intrinsic dimensionality, a unitless quantity that can be interpreted as the minimum number of latent variables needed to describe the data.

The many extant formulations of intrinsic dimensionality can be divided into two broad groups, global and local. Global intrinsic dimensionality, which takes contributions from the full dataset to measure its complexity as a whole, has been more widely investigated. By contrast, local variants of intrinsic dimensionality assess the complexity of the data in the vicinity of a designated query location, most notably in terms of the growth rate in the probability measure captured by an expanding neighborhood. Local variants can therefore associate different intrinsic dimensional values to different locations in the data domain.

Our focus in this paper is on the local intrinsic dimension (LID) as formulated in [1, 2], and in particular, establishing how it relates to entropy, perhaps the most fundamental and widely-used model of data complexity. In its essence, entropy can be regarded as a measure of the uncertainty of a distribution. Our study of entropy considers the distribution of distances to a query location, where the distances are induced by a global data distribution. In particular, we consider the entropy of the lower tail of the neighbor distance distribution (the *tail entropy*), and consider its asymptotic tendency as the neighborhood radius approaches zero.

Our analysis of the relationship between the tail entropy and local intrinsic dimensionality has further implications due to an established relationship between the latter and the statistical theory of extreme values (EVT) [3]. For any distribution of distances satisfying appropriate smoothness assumptions in the lower tail, as the neighborhood radius approaches zero, the tail distribution takes the form of a power law. Asymptotically, power law distributions can be said to arise naturally in the lower tail, with the exponent of the power law corresponding to the LID value.

We formulate asymptotic results that relate local intrinsic dimensionality with multiple variants of tail entropy. In particular, we relate LID to:

- **The cumulative tail entropy**. Cumulative entropy [4, 5] is an information-theoretic measure popular in reliability theory, where it is used to model uncertainty over time intervals. It corresponds to the expected value of the mean inactivity time. Compared to ordinary Shannon differential entropy, cumulative entropy has certain attractive properties, such as non-negativity and ease of estimation.

- **The cumulative tail entropy divergence**: The cumulative KL divergence has been proposed for comparing the cumulative distribution functions (CDFs) of two distributions [6, 7], and is closely related to the cumulative entropy. In this paper, we will be concerned with the Bregman variant of the cumulative KL divergence [8].

- **The tail entropy power**. The entropy power is the exponential of the entropy, and is also known as *perplexity* in the natural language processing community.

- **Generalized tail entropies** (tail cumulative q-entropy and tail q-entropy power). Generalized Tsallis entropies [9, 10] are a family of entropies characterized via an exponent parameter q applied to the probabilities, in which the traditional (Shannon) entropy variants are obtained as the special case $q \to 1$. The use of such a parameter can often facilitate more accurate fitting of data characteristics and robustness to outliers.

We believe our theoretical results are interesting in that they capture fundamental properties of local neighborhood geometry, and since they hold asymptotically for essentially all smooth data distributions.

These results also have two interesting applications which we explore in this paper:

- **Estimation:** Our theoretical results connect LID to a range of entropic quantities. These relationships immediately open the door to formulation of new estimators for LID, based on estimation of the entropic quantities. Our theory allows the development of several new estimators for the LID of a query point, by applying existing estimators for entropy [11], cumulative entropy [4] and cumulative q-entropy [10] to samples of a sufficiently-small neighborhood of the query. We are also formulate a new LID estimator with an appealing analytic form related to Bregman divergences, based on minimization of the cumulative KL divergence between the empirical distribution and an estimate of the true distribution.

- **Feature representation**: LID estimates can be used as features or as characterizations within machine learning models, such as for the detection of adversarial examples [12] or overfitting during learning [13]. However, the nonlinearity of LID may degrade its utility as a feature when used in linear classifiers, resulting in lower accuracy of trained models. However, we may instead use the tail entropy power as a feature. As we show in this paper, it tends toward a nonlinear transformation of LID as the tail size tends to zero, and thus has potential for use as a feature in linear models. In this paper, we provide experimental evidence of its effectiveness, by comparing its performance against raw LID features for an adversarial classification scenario.

Overall, our key contributions are the development of new theory that asymptotically relates tail entropy quantities and LID, with applications of this theory for estimation and feature representation.

2. Related Work

Our work relates to intrinsic dimensionality and its estimation, as well as tail entropy and its varieties such as generalized tail entropy and cumulative tail entropy, which we will define formally in Section 4. We briefly review each of these in turn.

Intrinsic dimensionality can be assessed either globally (for all data points) or locally (with respect to a chosen query point). Surveys of the field provide more detail [14, 15, 16]. In the global case, considerable work has focused on topological models, with accompanying estimation methods [17, 18, 19]. Examples here include PCA and its variants [20], graph based methods [21] and fractal models [14, 22]. Other types of techniques such as IDEA [23, 24] and DANCo [25] estimate the dimension based on concentration of norms and angles, or 2-nearest neighbors [26].

For local intrinsic dimensionality, a popular estimator is the maximum likelihood estimator, studied in the Euclidean setting by Levina and Bickel [27] and later formulated under more general assumptions in the context of EVT by Amsaleg et al. [2, 3], who showed it to be equivalent to the classic Hill estimator [28]. Other local estimators include expected simplex skewness [29], the tight locality estimator [30], the MiND framework [24] and the manifold adaptive dimension [31]. A recent estimator, LIDL [32], leverages neural density estimation methods to achieve strong performance for scenarios where the LID is very high.

Local intrinsic dimensionality has been used in a range of applications. These include modeling deformation in complex materials [33, 34], dimension reduction via local PCA [35], interpreting basketball player tracking data [36], climate science [37], assessing the complexity of COVID-19 data [38], similarity search [39], outlier detection [40], adversarial example detection [12], adversarial nearest neighbor [41, 42] and deep learning understanding [13, 43], clustering [44] and statistical manifold learning [45]. For example, in deep learning, it has been shown that adversarial examples are located in high-LID subspaces, and such a characteristic can be leveraged to build accurate adversarial example detectors [12]. It has also been found that the LID of deep representations [43] or input data [46] is an indicator of the generalization performance of deep neural networks (DNNs). A manifold *dimensionality expansion* phenomenon has been observed when DNNs overfit to noisy labels [13].

Cumulative entropy was formulated in [4] and is a variant of cumulative residual entropy [5]. Outside of reliability theory analysis, it has been used in such data mining tasks as dependency analysis [47] and subspace cluster analysis [48], where it has proved effective due to the existence of good estimators. Such investigation has been at a global level (over the entire data domain), rather than at the local level as in our study. Generalized variants based on Tsallis q-statistics have been developed for both entropy [9] and cumulative entropy [10].

Cumulative entropy has also been used as a building block in the cumulative Kullback-Leibler (KL) divergence [49] for comparing the CDFs of two distributions [6, 7]. This form of KL divergence is similar to the cumulative residual KL information [50]. Another alternative formulation, which we will consider in this paper, is the Bregman variant of the KL divergence. The broad family of Bregman divergences can be regarded as a generalization of the notion of distance, one that does not satisfy the triangle inequality. Many types of Bregman divergence exist [8], with a wide range of applications [51, 52].

The concept of tail entropy has been used in financial applications for assessing the expected shortfall [53] in the upper tail using quantization. This is different from our context, where we analyze lower tails and develop exact results for an asymptotic regime.

This paper is an extended version of a preliminary conference paper [54]. It extends that work by i) establishing a theoretical connection to LID for the cumulative Bregman KL divergence (requiring extension of technical lemmas from [54] to be able to handle divergences as well as entropies); ii) proposing four new estimators based on variants of tail entropy and tail divergence; and iii) providing an experimental investigation of both the effectiveness of these new estimators and the effectiveness of using tail entropies (instead of raw LID) for feature representation.

Another related paper by Bailey et al. [55] extends the results of [54] to a wide range of statistical divergences and distances, as well as formulating extensions of the theory to a multivariate context. The analysis differs from that of [54] (and this paper), in that [55] provides a only theoretical toolkit by which relationships to LID can be derived through many steps — no fully-worked general formulations are given. Also, it should be emphasized that [54, 55] do not deal with Bregman divergences or the cumulative KL divergence, nor do they deal with the estimation and feature representation issues that are the main focus of this paper (Sections 5 and 6).

3. Local Intrinsic Dimensionality

In this section, we summarize the LID model using the formulation of [2].

LID can be regarded as a continuous extension of the expansion dimension due to Karger and Ruhl [56, 57]. Like earlier expansion-based models of intrinsic dimension, it draws its motivation from the relationship between volume and radius in an expanding ball, where (as originally stated in [1]) the volume of the ball is taken to be the probability measure associated with the region it encloses. The probability as a function of radius — denoted by $F(r)$ — has the form of a univariate cumulative distribution function (CDF). The model formulation (as stated in [2]) generalizes this notion to real-valued functions F for which $F(0) = 0$, under appropriate assumptions of smoothness.

Definition 1 ([2]). *Let F be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. The* intrinsic dimensionality of F at r *is defined as follows whenever the limit exists:*

$$\mathrm{IntrDim}_F(r) \;\triangleq\; \lim_{\epsilon\to 0} \frac{\ln\left(F((1+\epsilon)r)/F(r)\right)}{\ln(1+\epsilon)}\,.$$

When F satisfies certain smoothness conditions in the vicinity of r, its intrinsic dimensionality has a convenient known form:

Theorem 1 ([2]). *Let F be a real-valued function that is non-zero over some open interval containing $r \in \mathbb{R}$, $r \neq 0$. If F is continuously differentiable at r, then*

$$\mathrm{ID}_F(r) \;\triangleq\; \frac{r \cdot F'(r)}{F(r)} \;=\; \mathrm{IntrDim}_F(r)\,.$$

Let $\mathbf{x}$ be a location of interest within a data domain $\mathcal{S}$ for which the distance measure d has been defined. To any generated sample $\mathbf{y} \in \mathcal{D}$ we can associate the distance $r = d(\mathbf{x}, \mathbf{y})$; in this way, the global distribution that produces samples $\mathbf{y}$ can be said to induce a local distance distribution with CDF F with respect to $\mathbf{x}$. In characterizing the local intrinsic dimensionality in the vicinity of location $\mathbf{x}$, we are interested in the limit of $\mathrm{ID}_F(r)$ as the distance r tends to 0, which we denote by

$$\mathrm{ID}^*_F \triangleq \lim_{r \to 0} \mathrm{ID}_F(r) \, .$$

Henceforth, when we refer to the local intrinsic dimensionality (LID) of a function F, or of a point $\mathbf{x}$ whose induced distance distribution has F as its CDF, we will take 'LID' to mean the quantity ID^*_F.

To gain a better intuitive understanding of LID and how it can be interpreted, consider the ideal case in which points in the neighborhood of $\mathbf{x}$ are distributed uniformly within a submanifold in $\mathcal{D}$. Here, in this ideal setting, the dimension of the submanifold would equal ID^*_F. In general, however, data distributions are not ideal, the manifold model of data does not perfectly apply, and ID^*_F is not necessarily an integer. In practice, estimation of the LID at $\mathbf{x}$ would give an indication of the dimension of the submanifold containing $\mathbf{x}$ that best fits the distribution.

The function ID_F can be seen to fully characterize its associated function F. This result is analogous to a foundational result from the statistical theory of extreme values (EVT), in that it corresponds under an inversion transformation to the Karamata representation theorem [58] for the upper tails of regularly varying functions. For more information on EVT and how the LID model relates to it, we refer the reader to [59, 2, 60]. This is captured in the following theorem, whose proof can be found in [2].

Theorem 2 (LID Representation Theorem [2]). *Let $F : \mathbb{R} \to \mathbb{R}$ be a real-valued function, and assume that ID^*_F exists. Let x and w be values for which x/w and $F(x)/F(w)$ are both positive. If F is non-zero and continuously differentiable everywhere in the interval $[\min\{x, w\}, \max\{x, w\}]$, then*

$$\frac{F(x)}{F(w)} = \left(\frac{x}{w}\right)^{\mathrm{ID}^*_F} \cdot A_F(x, w), \quad \textit{where} \quad A_F(x, w) \triangleq \exp\left(\int_x^w \frac{\mathrm{ID}^*_F - \mathrm{ID}_F(t)}{t} \, \mathrm{d}t\right),$$

whenever the integral exists.

In [2], conditions on x and w are provided for which the factor $A_F(x, w)$ can be seen to tend to 1 as $x, w \to 0$. The convergence characteristics of F to its asymptotic form are expressed by the factor $A_F(x, w)$, which is related to the slowly-varying component of functions as studied in EVT [59]. In the next section, we make use of the LID Representation Theorem in our analysis of the limits of tail entropy variants under a form of normalization.

4. Tail Entropy and LID

In this section, we will establish relationships between local intrinsic dimensionality and several forms of entropy *conditioned* on the lower tails of smooth functions on domains bounded from below at zero. The results presented in this section all hold *asymptotically* as the tail boundary tends toward zero, when *normalized* with respect to the length of the tail.

4.1. Definitions of Tail Entropy Variants

We begin with formal definitions of the tail entropies considered in this paper. In each case, we assume that we are given a function F over the non-negative real numbers, whose restriction to the lower tail $[0, w]$ satisfies the following smooth growth properties:

- $F(0) = 0$, and $F(t) > 0$ for $t \in (0, w]$;
- F is strictly monotonically increasing;
- F is continuously differentiable.

The function $F_w(t) \triangleq F(t)/F(w)$ thus satisfies the conditions of a cumulative distribution function over $t \in [0, w]$ (recall that $F(t|t \leq w) = F(t)/F(w)$ over $t \in [0, w]$), with the derivative $F'_w(t) = F'(t)/F(w)$ as its corresponding probability density function (PDF).

The following definitions apply to any functions F and G satisfying the conditions stated above.

We begin by defining the tail entropy. When F corresponds to the CDF of the lower tail of a query's neighbor distance distribution, the tail entropy assesses the uncertainty in the possible distances.

Definition 2 (Tail Entropy). *The entropy of F conditioned on $[0, w]$ is given by*

$$\mathrm{H}(F, w) \triangleq -\int_0^w F'_w(t) \ln F'_w(t) \, \mathrm{d}t \, .$$

The cumulative entropy is a variant of entropy proposed in [4, 5] due to its attractive theoretical properties. When F corresponds to the CDF of a query's neighbor distance distribution, the cumulative entropy can be regarded as an alternative measure of uncertainty for the possible distances.

Tail conditioning on the cumulative entropy has the same general form as that of the tail entropy.

Definition 3 (Cumulative Tail Entropy). *The cumulative entropy of F conditioned on $[0, w]$ is*

$$\mathrm{cH}(F, w) \triangleq -\int_0^w F_w(t) \ln F_w(t) \, \mathrm{d}t \, .$$

There are several standard definitions of entropy power in the research literature. For our purposes, we adopt the simplest — the exponential of Shannon entropy — for our definition conditioned to the tail. Entropy power can be interpreted as a statistical measure of a distribution's dispersion [61]. We will find that using the tail entropy power to be more appropriate than using the entropy, in our formulation for asymptotically small neighborhoods.

Definition 4 (Tail Entropy Power). *The entropy power of F conditioned on $[0, w]$ is defined to be*

$$\mathrm{HP}(F, w) \ \triangleq \ \exp\left(\mathrm{H}(F, w)\right).$$

For each of the tail entropy variants introduced above, we also propose analogous variants based on the q-entropy formulation due to Tsallis [9]. Tsallis q-entropies are a type of generalized entropy characterized by the use of a parameter q; due to their flexibility, they have found a wide range of applications [62]. In general, q-entropy formulations can be shown to be identical to their Shannon entropy analogues in the limit as q tends to 1.

Definition 5 (Tail q-Entropy). *For any $q > 0$ $(q \neq 1)$, the q-entropy of F conditioned on $[0, w]$ is defined to be*

$$\mathrm{H}_q(F, w) \ \triangleq \ \frac{1}{q-1}\left(1 - \int_0^w \left(F'_w(t)\right)^q \mathrm{d}t\right) \ = \ \frac{1}{q-1}\int_0^w F'_w(t) - \left(F'_w(t)\right)^q \mathrm{d}t\,.$$

Definition 6 (Cumulative Tail q-Entropy). *For any $q > 0$ $(q \neq 1)$, the cumulative q-entropy of F conditioned on $[0, w]$ is defined to be*

$$\mathrm{cH}_q(F, w) \triangleq \frac{1}{q-1}\int_0^w F_w(t) - \left(F_w(t)\right)^q \mathrm{d}t\,.$$

We define the tail q-entropy power using the q-exponential function from Tsallis statistics [9], $\exp_q(x) \triangleq [1 + (1-q)x]^{\frac{1}{1-q}}$. Note that L'Hôpital's rule can be used to show that $\exp_q(x) \to e^x$ as $q \to 1$.

Definition 7 (Tail q-Entropy Power). *For any $q > 0$ $(q \neq 1)$, the q-entropy power of F conditioned on $[0, w]$ is defined to be*

$$\mathrm{HP}_q(F, w) \ \triangleq \ \left[1 + (1-q)H_q(F, w)\right]^{\frac{1}{1-q}}\ .$$

In addition to the entropies of lower tails, we will also consider comparison of the lower tails of two distributions F and G. The Bregman divergence [8] is one natural approach for such comparison. The KL form of the Bregman divergence between two positive real numbers $x, y \in \mathbb{R}_+$ is defined as:

$$d_{\mathrm{KL}}(x, y) \triangleq x \ln\frac{x}{y} - x + y\,. \tag{1}$$

Using the Bregman KL divergence, one can compute the divergence between the lower tails of two distributions with CDFs F and G.

Table 1: Asymptotic relationships between normalized tail entropy variants and local intrinsic dimensionality.

Entropy Variant	**Normalized Tail Entropy**	**Limit as w → $\mathbf{0}^+$**
Cumulative Entropy	$\mathrm{ncH}(F,w) \triangleq \frac{1}{w}\mathrm{cH}(F,w)$	$\frac{\mathrm{ID}_F^*}{(\mathrm{ID}_F^*+1)^2}$
Cumulative q-Entropy	$\mathrm{ncH}_q(F,w) \triangleq \frac{1}{w}\mathrm{cH}_q(F,w)$	$\frac{\mathrm{ID}_F^*}{(\mathrm{ID}_F^*+1)(q\,\mathrm{ID}_F^*+1)}$
Entropy Power	$\mathrm{nHP}(F,w) \triangleq \frac{1}{w}\mathrm{HP}(F,w)$	$\frac{1}{\mathrm{ID}_F^*}\exp\left(1-\frac{1}{\mathrm{ID}_F^*}\right)$
q-Entropy Power	$\mathrm{nHP}_q(F,w) \triangleq \frac{1}{w}\mathrm{HP}_q(F,w)$	$\left(\frac{(\mathrm{ID}_F^*)^q}{q\,\mathrm{ID}_F^*-q+1}\right)^{\frac{1}{1-q}}$
C. B. KL Divergence	$d_{\mathrm{nKL}}(F,G,w) \triangleq \frac{1}{w}d_{\mathrm{KL}}(F,G,w)$	$\frac{(\mathrm{ID}_F^*-\mathrm{ID}_G^*)^2}{(\mathrm{ID}_F^*+1)^2(\mathrm{ID}_G^*+1)}$

Definition 8 (Cumulative Tail Bregman KL Divergence). *The cumulative (Bregman) KL divergence between F and G, conditioned on $[0,w]$, is given by*

$$\begin{aligned} d_{\mathrm{KL}}(F,G,w) &\triangleq \int_0^w d_{\mathrm{KL}}\left(F_w(t),G_w(t)\right)\,\mathrm{d}t \\ &= \int_0^w F_w(t)\ln\frac{F_w(t)}{G_w(t)} - F_w(t) + G_w(t)\,\mathrm{d}t\,. \end{aligned}$$

Observe that the divergence can be written as

$$\begin{aligned} d_{\mathrm{KL}}(F,G,w) &= \int_0^w F_w(t)\ln F_w(t)\,\mathrm{d}t - \int_0^w F_w(t)\ln G_w(t)\,\mathrm{d}t \\ &\qquad - \int_0^w F_w(t)\,\mathrm{d}t + \int_0^w G_w(t)\,\mathrm{d}t\,, \end{aligned}$$

where the first term is the cumulative entropy of F, and the second term can be regarded as a 'cross cumulative entropy' between F_w and G_w. The 'cross cumulative entropy' has been termed the *cumulative inaccuracy* in [49].

For the cumulative tail entropy and divergence variants, and the tail entropy power variants, we will also consider a normalization given by the ratio with w, the length of the tail. In the remainder of this section, we will show that as w tends to zero, the limits of these normalized entropies can be expressed in terms of the local intrinsic dimensionality of F. The notation for these normalized variants, and our theorems for their limits in terms of LID, are summarized in Table 1. The following subsections explain and prove these results.

4.2. Technical Preliminaries

Before presenting the main theoretical results of the paper, we begin with several technical lemmas. The first lemma concerns a slight generalization of the cumulative entropy formulation, that allows it to greatly facilitate the proofs for two tail entropy variants, the cumulative entropy and the entropy power.

Lemma 1. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ and $G : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be functions such that $F(0) = G(0) = 0$, and assume that ID^*_F and ID^*_G exist and are positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F and G are continuously differentiable and strictly monotonically increasing. Let ϕ and ψ be functions over $(0, r)$, with ψ positive. Then for any constants u and v such that $u < \mathrm{ID}^*_F$,*

$$\lim_{w \to 0^+} \phi(w) \int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v} \, \mathrm{d}t$$
$$= \lim_{w \to 0^+} \frac{w^{1-u}\phi(w)}{\mathrm{ID}^*_F + 1 - u} \left[\ln \frac{\psi(w)}{w^v} - \frac{\mathrm{ID}^*_G - v}{\mathrm{ID}^*_F + 1 - u} \right]$$

whenever the right-hand limit exists or diverges to $+\infty$ or $-\infty$.

Proof: Since the limit $\mathrm{ID}^*_F = \lim_{x \to 0^+} \mathrm{ID}_F(x)$ is assumed to exist, we have that for any real value $\epsilon > 0$ satisfying $\epsilon < \min\{r, \mathrm{ID}^*_F - u, |\mathrm{ID}^*_G - v|\}$, there must exist a value $0 < \delta < \epsilon$ such that $x < \delta$ implies that $|\mathrm{ID}_F(x) - \mathrm{ID}^*_F| < \epsilon$ and $|\mathrm{ID}_G(x) - \mathrm{ID}^*_G| < \epsilon$. Therefore, when $0 < t \leq w < \delta$,

$$|\ln A_F(t, w)| \;=\; \left| \int_t^w \frac{\mathrm{ID}^*_F - \mathrm{ID}_F(x)}{x} \, \mathrm{d}x \right| \;<\; \epsilon \cdot \left| \int_t^w \frac{1}{x} \, \mathrm{d}x \right| \;=\; \epsilon \cdot \ln \frac{w}{t} \, ,$$

and similarly,

$$|\ln A_G(t, w)| \;<\; \epsilon \cdot \ln \frac{w}{t} \, .$$

Exponentiating, we obtain the bounds

$$\left(\frac{w}{t} \right)^{-\epsilon} \;<\; \{ A_F(t, w) \, , \; A_G(t, w) \} \;<\; \left(\frac{w}{t} \right)^{\epsilon} . \tag{2}$$

Applying Theorem 2 to $G_w(t)$ in the theorem statement, and making use of the upper bound on A_G from Inequality 2, the integral becomes

$$\int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v} \, \mathrm{d}t$$
$$= \int_0^w \frac{F_w(t)}{t^u} \cdot \ln \left(\frac{\psi(w)}{t^v} \left(\frac{t}{w} \right)^{\mathrm{ID}^*_G} A_G(t, w) \right) \mathrm{d}t$$
$$< \int_0^w \frac{F_w(t)}{t^u} \cdot \left[\ln \left(\frac{\psi(w)}{t^v} \left(\frac{t}{w} \right)^{\mathrm{ID}^*_G} \right) + \ln \left(\frac{t}{w} \right)^{-\epsilon} \right] \mathrm{d}t \, .$$

Similarly, using the lower bound on A_G, we obtain a lower bound on the integral:

$$\int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v}\, \mathrm{d}t$$
$$> \int_0^w \frac{F_w(t)}{t^u} \cdot \left[\ln \left(\frac{\psi(w)}{t^v} \left(\frac{t}{w} \right)^{\mathrm{ID}_G^*} \right) + \ln \left(\frac{t}{w} \right)^{\epsilon} \right] \mathrm{d}t\,.$$

Since ϵ can be chosen arbitrarily close to 0, the upper and lower bounds converge to one another. The squeeze theorem for integrals therefore gives us the following equivalence.

$$\begin{aligned}
&\int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v}\, \mathrm{d}t \\
&\quad = \int_0^w \frac{F_w(t)}{t^u} \cdot \ln \left(\frac{\psi(w)}{t^v} \left(\frac{t}{w} \right)^{\mathrm{ID}_G^*} \right) \mathrm{d}t \\
&\quad = \int_0^w \frac{F_w(t)}{t^u} \cdot \left[\ln \frac{\psi(w)}{w^{\mathrm{ID}_G^*}} + (\mathrm{ID}_G^* - v) \ln t \right] \mathrm{d}t\,. \qquad (3)
\end{aligned}$$

Continuing along the same lines, applying Theorem 2 to $F_w(t)$ in Equation 3, the bound on A_F from Inequality 2 lead us to further simplifications. However, the argument is complicated by the fact that the sum of logarithmic terms produces a factor that could have different signs for different choices of t and w. Using the following notation, we distinguish the two cases in which the (non-zero) contribution of the logarithmic terms is positive or negative:

$$\begin{aligned}
\Psi_w(t) &\triangleq \ln \frac{\psi(w)}{w^{\mathrm{ID}_G^*}} + (\mathrm{ID}_G^* - v) \ln t \\
\Psi_w^+(t) &\triangleq \max \{\Psi_w(t), 0\} \\
\Psi_w^-(t) &\triangleq \min \{\Psi_w(t), 0\}\ .
\end{aligned}$$

Expanding Equation 3 and applying the bounds of Inequality 2,

$$\begin{aligned}
&\int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v}\, \mathrm{d}t \\
&\quad = \int_0^w \frac{1}{t^u} \left(\frac{t}{w} \right)^{\mathrm{ID}_F^*} \cdot A_F(t, w) \cdot \left[\ln \Psi_w^+(t) + \ln \Psi_w^-(t) \right] \mathrm{d}t \\
&\quad < \int_0^w \frac{1}{t^u} \left(\frac{t}{w} \right)^{\mathrm{ID}_F^*} \cdot \left[\ln \Psi_w^+(t) \cdot \left(\frac{t}{w} \right)^{-\epsilon} + \ln \Psi_w^-(t) \cdot \left(\frac{t}{w} \right)^{\epsilon} \right] \mathrm{d}t \\
&\quad \text{and} \\
&\quad > \int_0^w \frac{1}{t^u} \left(\frac{t}{w} \right)^{\mathrm{ID}_F^*} \cdot \left[\ln \Psi_w^+(t) \cdot \left(\frac{t}{w} \right)^{\epsilon} + \ln \Psi_w^-(t) \cdot \left(\frac{t}{w} \right)^{-\epsilon} \right] \mathrm{d}t\,.
\end{aligned}$$

Once again, the squeeze theorem for integrals yields an exact relationship:

$$\begin{aligned}
&\int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v}\, \mathrm{d}t \\
&\quad = \int_0^w \frac{1}{t^u} \left(\frac{t}{w}\right)^{\mathrm{ID}^*_F} \cdot \ln \Psi_w(t)\, \mathrm{d}t \\
&\quad = \frac{1}{w^{\mathrm{ID}^*_F}} \int_0^w t^{\mathrm{ID}^*_F - u} \cdot \left[\ln \frac{\psi(w)}{w^{\mathrm{ID}^*_G}} + (\mathrm{ID}^*_G - v) \ln t \right] \mathrm{d}t\,. \qquad (4)
\end{aligned}$$

Noting that $u < \mathrm{ID}^*_F$ implies that $\lim_{t\to 0} t^{\mathrm{ID}^*_F - u} \ln t = 0$, integration by parts of the right-hand side of Equation 4 yields a closed expression that depends on F and G only through their LID values. Letting $m \triangleq \mathrm{ID}^*_F - u$ and $n \triangleq \mathrm{ID}^*_G - v$,

$$\begin{aligned}
&\lim_{w\to 0^+} \phi(w) \int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v}\, \mathrm{d}t \\
&\quad = \lim_{w\to 0^+} \frac{\phi(w)}{w^{m+u}} \int_0^w t^m \cdot \left[\ln \frac{\psi(w)}{w^{n+v}} + n \ln t \right] \mathrm{d}t \\
&\quad = \lim_{w\to 0^+} \frac{\phi(w)}{w^{m+u}} \left[\frac{nt^{m+1}}{m+1} \ln t \,\Big|_0^w - \int_0^w \frac{nt^{m+1}}{m+1} \cdot \frac{1}{t}\, \mathrm{d}t \,+\, \frac{w^{m+1}}{m+1} \ln \frac{\psi(w)}{w^{n+v}} \right] \\
&\quad = \lim_{w\to 0^+} \frac{\phi(w)}{w^{m+u}} \left[\frac{nw^{m+1}}{m+1} \ln w - \frac{nw^{m+1}}{(m+1)^2} + \frac{w^{m+1}}{m+1} \ln \frac{\psi(w)}{w^{n+v}} \right] \\
&\quad = \lim_{w\to 0^+} \frac{w^{1-u}\phi(w)}{m+1} \left[n \ln w - \frac{n}{m+1} + \ln \frac{\psi(w)}{w^{n+v}} \right] \\
&\quad = \lim_{w\to 0^+} \frac{w^{1-u}\phi(w)}{\mathrm{ID}^*_F + 1 - u} \left[\ln \frac{\psi(w)}{w^v} - \frac{\mathrm{ID}^*_G - v}{\mathrm{ID}^*_F + 1 - u} \right]
\end{aligned}$$

whenever the limit exists, or diverges to $+\infty$ or $-\infty$. □

From this result, we observe that the existence of a non-trivial (finite but non-zero) limit imposes strong conditions on both ϕ and ψ. For the former, the limit $\lim_{w\to 0^+} w^{1-u}\phi(w)$ must exist and be non-zero; for the latter, the limit $\lim_{w\to 0^+} w^{-v}\psi(w)$ must exist and be positive.

The second and third technical lemmas are obtained as byproducts of the proof of Lemma 1.

Corollary 1. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ and $G : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be functions such that $F(0) = G(0) = 0$, and assume that ID^*_F and ID^*_G exist and are positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F and G are continuously differentiable and strictly monotonically increasing. Let ϕ and ψ be functions over $(0, r)$, with ψ positive. Then for any constants u and v such that*

$u < \mathrm{ID}^*_F$,

$$\lim_{w\to 0^+} \phi(w) \int_0^w \frac{F_w(t)}{t^u} \ln \frac{\psi(w)\, G_w(t)}{t^v}\, \mathrm{d}t$$
$$= \lim_{w\to 0^+} \phi(w) \int_0^w \frac{F_w(t)}{t^u} \cdot \left[\ln \frac{\psi(w)}{w^{\mathrm{ID}^*_G}} + (\mathrm{ID}^*_G - v) \ln t \right] \mathrm{d}t$$

whenever the right-hand limit exists or diverges to $+\infty$ or $-\infty$.

Proof: Follows directly from Equation 3. □

Corollary 2. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be a function such that $F(0) = 0$, and assume that ID^*_F exists and is positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F is continuously differentiable and strictly monotonically increasing. Let ϕ be a functions over $(0, r)$. Then for any constant u,*

$$\lim_{w\to 0^+} \phi(w) \int_0^w \frac{F_w(t)}{t^u}\, \mathrm{d}t \quad = \quad \lim_{w\to 0^+} \frac{w^{1-u}\phi(w)}{\mathrm{ID}^*_F + 1 - u}$$

whenever the right-hand limit exists or diverges to $+\infty$ or $-\infty$.

Proof: Omitted, since the result follows from bounding arguments very similar to (but much simpler than) those found in Lemma 1. □

The final technical lemma also follows as a corollary of Lemma 1, since it uses much of the same proof strategy, albeit more simply and directly. Analogous with Lemma 1, it concerns a slight generalization of the cumulative q-entropy formulation that facilitates the proof of the results for the q-entropy and q-entropy power variants.

Corollary 3. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be a function such that $F(0) = 0$, and assume that ID^*_F exists and is positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F is continuously differentiable and strictly monotonically increasing. Let ϕ be a positive function over $(0, r)$. Then for any constants $u < \mathrm{ID}^*_F$ and $z > 0$,*

$$\lim_{w\to 0^+} w^{zu-1} \int_0^w \left(\frac{\phi(w)\, F_w(t)}{t^u} \right)^z \mathrm{d}t \quad = \quad \frac{\lim_{w\to 0^+} \phi^z(w)}{z\, \mathrm{ID}^*_F - zu + 1}$$

whenever the right-hand limit exists, or diverges to $+\infty$ or $-\infty$.

Proof: Following the same proof strategy of Lemma 1 that led to Equation 4, we arrive at the following upper bound on the integral:

$$\int_0^w \left(\frac{\phi(w)\, F_w(t)}{t^u} \right)^z \mathrm{d}t \quad < \quad \frac{\phi^z(w)}{w^{z(m+u-\epsilon)}} \int_0^w t^{z(m-\epsilon)}\, \mathrm{d}t \quad = \quad \frac{\phi^z(w)}{(zm - z\epsilon + 1) w^{zu-1}}\,,$$

where $m = \mathrm{ID}^*_F - u$ as before.

Continuing according to the proof strategy of Lemma 1, we use the lower bound from Equation 2, and let ϵ vanish by applying the limit $w \to 0^+$ with an introduced factor of w^{zu-1}. This brings us to

$$\lim_{w \to 0^+} w^{zu-1} \int_0^w \left(\frac{\phi(w)\, F_w(t)}{t^u} \right)^z \mathrm{d}t$$
$$= \lim_{w \to 0^+} w^{zu-1} \frac{\phi^z(w)}{(z\,\mathrm{ID}^*_F - zu + 1) w^{zu-1}} = \frac{\lim_{w \to 0^+} \phi^z(w)}{z\,\mathrm{ID}^*_F - zu + 1},$$

as required. □

4.3. Cumulative Tail Entropy and LID

Using the technical lemmas established in Section 4.2, we present the main results for the cumulative tail entropy variants. The first result shows that as the tail length w tends to zero, the normalized cumulative entropy $\mathrm{ncH}(F, w) \triangleq \frac{1}{w}\mathrm{cH}(F, w)$ tends to a value entirely determined by the local intrinsic dimensionality associated with F.

Theorem 3. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be a function such that $F(0) = 0$, and assume that ID^*_F exists and is positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F is continuously differentiable and strictly monotonically increasing. We have*

$$\lim_{w \to 0^+} \mathrm{ncH}(F, w) = \lim_{w \to 0^+} -\frac{1}{w} \int_0^w F_w(t) \ln F_w(t)\, \mathrm{d}t = \frac{\mathrm{ID}^*_F}{(\mathrm{ID}^*_F + 1)^2}.$$

Proof: Follows directly from Lemma 1, for the choices $G = F$, $u = v = 0$, $\psi(w) = 1$, and $\phi(w) = w^{-1}$. □

The second result uses Corollary 3 to show that as the tail length w tends to zero, the normalized cumulative q-entropy $\mathrm{ncH}_q(F, w) \triangleq \frac{1}{w}\mathrm{cH}_q(F, w)$ tends to a value determined by q together with the local intrinsic dimensionality associated with F.

Theorem 4. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be a function such that $F(0) = 0$, and assume that ID^*_F exists and is positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F is continuously differentiable and strictly monotonically increasing. Then for $q > 0$ with $q \neq 1$,*

$$\lim_{w \to 0^+} \mathrm{ncH}_q(F, w)$$
$$= \lim_{w \to 0^+} \frac{1}{w(q-1)} \int_0^w F_w(t) - (F_w(t))^q\, \mathrm{d}t = \frac{\mathrm{ID}^*_F}{(\mathrm{ID}^*_F + 1)(q\,\mathrm{ID}^*_F + 1)}.$$

Proof: Separating the integral and applying Corollary 3 twice,

$$\lim_{w\to 0^+} \frac{1}{w(q-1)} \int_0^w F_w(t) - (F_w(t))^q \,\mathrm{d}t$$

$$= \frac{1}{q-1}\left(\frac{1}{\mathrm{ID}^*_F + 1} - \frac{1}{q\,\mathrm{ID}^*_F + 1}\right) = \frac{\mathrm{ID}^*_F}{(\mathrm{ID}^*_F + 1)(q\,\mathrm{ID}^*_F + 1)}$$

follows for the choices $u = 0$, $\phi(w) = 1$, and (respectively) $z = 1$ and $z = q$. □

Observe that as q tends to 1, the cumulative q-entropy variant $\mathrm{ncH}_q(F, w)$ does tend to the cumulative entropy $\mathrm{ncH}(F, w)$, as one would expect.

4.4. Tail Entropy Power and LID

We find that we encounter convergence issues when attempting to use the machinery of Lemma 1 to formulate a relationship between LID and either the tail entropy $\mathrm{H}(F, w)$ or the normalized tail entropy $\mathrm{nH}(F, w)$, in that the limits diverge as the tail size tends to zero.

Instead, we show that the entropy power, when normalized, does have a limit expressed as a function of the LID of F.

Theorem 5. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be a function such that $F(0) = 0$, and assume that ID^*_F exists and is greater than 1. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F is continuously differentiable and strictly monotonically increasing. Then*

$$\lim_{w\to 0^+} \mathrm{nHP}(F, w)$$

$$= \lim_{w\to 0^+} \frac{1}{w} \exp\left(-\int_0^w F'_w(t) \ln F'_w(t) \,\mathrm{d}t\right) = \frac{1}{\mathrm{ID}^*_F} \exp\left(1 - \frac{1}{\mathrm{ID}^*_F}\right).$$

Proof: For convenience of expression, we make use of the notation $\mathrm{xlnx}(x) \triangleq x \ln x$. Applying the formula of Theorem 1,

$$\lim_{w\to 0^+} \frac{\mathrm{HP}(F, w)}{w} = \lim_{w\to 0^+} \frac{1}{w} \exp\left(-\int_0^w \mathrm{xlnx}\left(\frac{F'(t)}{F(w)}\right) \mathrm{d}t\right)$$

$$= \lim_{w\to 0^+} \frac{1}{w} \exp\left(-\int_0^w \mathrm{xlnx}\left(\frac{\mathrm{ID}_F(t) F(t)}{t F(w)}\right) \mathrm{d}t\right).$$

Since ID^*_F is assumed to exist, for any real value $\epsilon > 0$ satisfying $\epsilon < \min\{r, \mathrm{ID}^*_F - u\}$, there must exist a value $0 < \delta < \epsilon$ such that $v < \delta$ implies that $|\,\mathrm{ID}_F(v) - \mathrm{ID}^*_F\,| < \epsilon$. Therefore, when $0 < t \leq w < \delta$, $\mathrm{ID}_F(t)$ falls within the interval $(\mathrm{ID}^*_F - \epsilon, \mathrm{ID}^*_F + \epsilon)$ over the entire integral. Since ϵ can be chosen to be arbitrarily small, $\mathrm{ID}_F(t)$ can be replaced by ID^*_F in the limit.

Next, we apply Lemma 1 for the choices $G = F$, $u = v = 1$ and $\phi(w) = \psi(w) = \mathrm{ID}_F^*$. The choice of u is valid for Lemma 1 since by assumption $\mathrm{ID}_F^* > 1 = u$.

$$\begin{aligned}
\lim_{w\to 0^+} \frac{\mathrm{HP}(F,w)}{w} &= \lim_{w\to 0^+} \frac{1}{w} \exp\left(-\int_0^w \mathrm{xlnx}\left(\frac{\mathrm{ID}_F^*\, F_w(t)}{t}\right) \mathrm{d}t\right) \\
&= \lim_{w\to 0^+} \frac{1}{w} \exp\left(-\frac{\mathrm{ID}_F^*}{\mathrm{ID}_F^* + 1 - 1}\left[\ln \frac{\mathrm{ID}_F^*}{w^1} - \frac{\mathrm{ID}_F^* - 1}{\mathrm{ID}_F^* + 1 - 1}\right]\right) \\
&= \lim_{w\to 0^+} \frac{1}{w} \exp\left(1 - \frac{1}{\mathrm{ID}_F^*} - \ln \mathrm{ID}_F^* + \ln w\right) \\
&= \frac{1}{\mathrm{ID}_F^*} \exp\left(1 - \frac{1}{\mathrm{ID}_F^*}\right) .
\end{aligned}$$

□

For the case of the normalized tail q-entropy power $\mathrm{nHP}_q(F,w)$, we have the following result.

Theorem 6. *Let $F : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be a function such that $F(0) = 0$, and assume that ID_F^* exists and is greater than 1. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F is continuously differentiable and strictly monotonically increasing. Then for $q > 0$ $(q \neq 1)$,*

$$\begin{aligned}
&\lim_{w\to 0^+} \mathrm{nHP}_q(F,w) \\
&\quad = \lim_{w\to 0^+} \frac{1}{w} \exp_q\left(\frac{1}{q-1}\left[1 - \int_0^w (F'_w(t))^q \,\mathrm{d}t\right]\right) = \left[\frac{(\mathrm{ID}_F^*)^q}{q\,\mathrm{ID}_F^* - q + 1}\right]^{\frac{1}{1-q}} .
\end{aligned}$$

Proof: Applying the formula of Theorem 1, and using arguments similar to that of the proof of Theorem 5, we obtain

$$\lim_{w\to 0^+} \frac{\mathrm{HP}_q(F,w)}{w} = \lim_{w\to 0^+} \frac{1}{w} \exp_q\left(\frac{1}{q-1}\left[1 - \int_0^w \left(\frac{\mathrm{ID}_F^*\, F(t)}{tF(w)}\right)^q \mathrm{d}t\right]\right) .$$

Since ID_F^* is assumed to exist, for any real value $\epsilon > 0$ satisfying $\epsilon < \min\{r, \mathrm{ID}_F^* - u\}$, there must exist a value $0 < \delta < \epsilon$ such that $v < \delta$ implies that $|\,\mathrm{ID}_F(v) - \mathrm{ID}_F^*\,| < \epsilon$. Therefore, when $0 < t \leq w < \delta$, $\mathrm{ID}_F(t)$ falls within the interval $(\mathrm{ID}_F^* - \epsilon, \mathrm{ID}_F^* + \epsilon)$ over the entire integral. Since ϵ can be chosen to be arbitrarily small, $\mathrm{ID}_F(t)$ can be replaced by ID_F^* in the limit.

Applying Corollary 3 for the choices $u = 1$, $\phi(w) = \mathrm{ID}_F^*$, and $z = q$, we arrive at the following. The choice of u is valid for Corollary 3 since by assumption $\mathrm{ID}_F^* > 1 = u$. Here, we also make use of the definition of the q-exponential,

$\exp_q = [1 + (1-q)x]^{\frac{1}{1-q}}$.

$$\begin{aligned}
\lim_{w\to 0^+} \frac{\mathrm{HP}_q(F,w)}{w} &= \lim_{w\to 0^+} \frac{1}{w} \exp_q\left(\frac{1}{q-1}\left[1 - \frac{w^{1-q}(\mathrm{ID}_F^*)^q}{q\,\mathrm{ID}_F^* - q + 1}\right]\right) \\
&= \lim_{w\to 0^+} \frac{1}{w} \left(\frac{w^{1-q}(\mathrm{ID}_F^*)^q}{q\,\mathrm{ID}_F^* - q + 1}\right)^{\frac{1}{1-q}} \\
&= \left(\frac{(\mathrm{ID}_F^*)^q}{q\,\mathrm{ID}_F^* - q + 1}\right)^{\frac{1}{1-q}} .
\end{aligned}$$

□

4.5. Tail Bregman KL Divergence and LID

We also consider a normalization by a factor of w for the limit of the cumulative tail Bregman KL divergence, as w tends to zero.

Theorem 7. *Let $F, G : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be functions such that $F(0) = 0$, and assume that ID_F^* and ID_G^* exist and are positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F and G are continuously differentiable and strictly monotonically increasing. Then*

$$\begin{aligned}
&\lim_{w\to 0^+} \frac{1}{w} d_{\mathrm{KL}}(F,G,w) \\
&\qquad = -\lim_{w\to 0^+} \frac{1}{w} \int_0^w F_w(t) \cdot \mathrm{ID}_G^* \ln \frac{t}{w}\,\mathrm{d}t \;-\; \frac{2\,\mathrm{ID}_F^* + 1}{(\mathrm{ID}_F^* + 1)^2} \;+\; \frac{1}{\mathrm{ID}_G^* + 1} .
\end{aligned}$$

Proof: Expanding the Bregman KL divergence gives us a sum of limits of integrals:

$$\begin{aligned}
\lim_{w\to 0^+} \frac{1}{w} d_{\mathrm{KL}}(F,G,w) &= -\lim_{w\to 0^+} \frac{1}{w} \int_0^w F_w(t) \ln G_w(t)\,\mathrm{d}t \\
&\quad + \lim_{w\to 0^+} \frac{1}{w} \int_0^w F_w(t) \ln F_w(t)\,\mathrm{d}t \\
&\quad - \lim_{w\to 0^+} \frac{1}{w} \int_0^w F_w(t)\,\mathrm{d}t \;+\; \lim_{w\to 0^+} \frac{1}{w} \int_0^w G_w(t)\,\mathrm{d}t .
\end{aligned}$$

Of the three terms in the result, the first is derived from the first integral in the expansion, using Corollary 1 with $u = v = 0$, $\psi(w) = 1$, and $\phi(w) = \frac{1}{w}$. The second term is obtained from the sum of the second and third integrals; the second integral can be expressed in terms of ID_F^* using Theorem 3 directly, and Corollary 2 can be applied to the third integral with $u = 0$ and $\phi(w) = \frac{1}{w}$. The third term in the result is obtained from the fourth integral of the expansion, again from Corollary 2 with $u = 0$ and $\phi(w) = \frac{1}{w}$. □

Corollary 4. *Let $F, G : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$ be functions such that $F(0) = 0$, and assume that ID^*_F and ID^*_G exist and are positive. For some value of $r > 0$, let us further assume that within the interval $[0, r)$, F and G are continuously differentiable and strictly monotonically increasing. Then*

$$\lim_{w \to 0^+} \frac{1}{w} d_{\mathrm{KL}}(F, G, w) \quad = \quad \frac{(\mathrm{ID}^*_F - \mathrm{ID}^*_G)^2}{(\mathrm{ID}^*_F + 1)^2 (\mathrm{ID}^*_G + 1)} \, .$$

Proof: The proof reduces to finding an expression in terms of ID^*_F and ID^*_G for the limit of the integral in the statement of Theorem 7. Rewriting this limit integral with $\phi(w) = \frac{\mathrm{ID}^*_G}{w}$, $\psi(w) = 1$, and $B_w(t) = \frac{t}{w}$, Lemma 1 can be applied with $u = v = 0$ to give

$$\begin{aligned} \lim_{w \to 0^+} \frac{1}{w} \int_0^w F_w(t) \cdot \mathrm{ID}^*_G \ln \frac{t}{w} \, \mathrm{d}t \quad &= \quad \lim_{w \to 0^+} \phi(w) \int_0^w F_w(t) \cdot \ln B_w(t) \, \mathrm{d}t \\ &= \quad -\frac{\mathrm{ID}^*_G}{\mathrm{ID}^*_F + 1} \cdot \frac{\mathrm{ID}^*_B}{\mathrm{ID}^*_F + 1} \quad = \quad -\frac{\mathrm{ID}^*_G}{(\mathrm{ID}^*_F + 1)^2} \, . \end{aligned}$$

Substituting this into the statement of Theorem 7 and then after some manipulation, the result follows. □

As one might expect, Corollary 4 confirms that the divergence is non-negative; moreover, if the LID values of F and G are the same, the tail limit of the cumulative Bregman KL divergence is zero.

5. Estimation

In the previous sections, we have established relationships between local intrinsic dimensionality and variants of tail entropies. We now investigate how these concepts and relationships can be leveraged to develop novel estimators for local intrinsic dimensionality.

We propose four new estimators for local intrinsic dimensionality:

- an estimator based on cumulative entropy;
- an estimator based cumulative q-entropy;
- an estimator based on entropy (power);
- an estimator based on minimization of the cumulative Bregman KL divergence.

In our formulation, we will assume the availability of the k nearest neighbor distances from the query point whose LID to be estimated. These distances are assumed to be in the tail of the distribution whose interval is $[0, w]$, yielding the order statistics $0 = X_0 \leq X_1 \leq X_2 \leq, \ldots, X_k = w$.

We now discuss the estimators in turn. For a summary of the estimators, please see Table 2.

Estimation Method	**Estimator Formula**
Cumulative Entropy	$\widehat{\mathrm{ID}}^*_{F,1} = \alpha(F,w) \pm \sqrt{\alpha^2(F,w)-1}$ where $\alpha(F,w) \triangleq \frac{w}{2\widehat{\mathrm{cH}}(F,w)} - 1$ $\widehat{\mathrm{cH}}(F,w) = -\sum_{j=1}^{k-1}(X_{j+1}-X_j)\frac{j}{k}\ln\frac{j}{k}$
Cumulative $q-$Entropy	$\widehat{\mathrm{ID}}^*_{F,q} = \frac{1}{\sqrt{q}}\left[\beta(F,w,q) \pm \sqrt{\beta^2(F,w,q)-1}\right]$ for $q \neq 1$, where $\beta(F,w,q) \triangleq \frac{1}{2\sqrt{q}}\left[\frac{w}{\widehat{\mathrm{cH}}_q(F,w)} - (q+1)\right]$ $\widehat{\mathrm{cH}}_q(F,w) = \frac{1}{q-1}\sum_{j=1}^{k-1}(X_{j+1}-X_j)\left(\frac{j}{k} - \left(\frac{j}{k}\right)^q\right)$
Entropy Power	$\exp(\widehat{H}(F,w)) = \frac{w}{\widehat{\mathrm{ID}}^*_{F,\mathrm{HP}}}\exp(1-\frac{1}{\widehat{\mathrm{ID}}^*_{F,\mathrm{HP}}})$ Solve numerically for $\widehat{\mathrm{ID}}^*_{F,\mathrm{HP}}$. Evaluate $\widehat{H}(F,w)$ with a univariate entropy estimator.
Cumulative Bregman KL Divergence	$\widehat{\mathrm{ID}}^*_{G,\mathrm{KL}} = \frac{\sqrt{X_k} - \sqrt{\frac{1}{k}\sum_{j=1}^{k} d_{\mathrm{KL}}(X_j,X_k)}}{\sqrt{\frac{1}{k}\sum_{j=1}^{k} d_{\mathrm{KL}}(X_j,X_k)}}$ where $d_{\mathrm{KL}}(x,y) = x\ln\frac{x}{y} - x + y$

Table 2: Summary of proposed estimators. All estimators are based on nearest neighbor distances, corresponding to the order statistics $0 \leq X_1 \leq X_2 \leq \ldots \leq X_k = w$.

5.1. Estimation Based on Cumulative Entropy and Cumulative q-Entropy

Using the relationships from Theorems 3 and 4 which hold in the tail $[0,w]$, we can relate an estimator of LID with estimators of cumulative tail entropy and cumulative tail q-entropy.

$$\begin{aligned} \mathrm{cH}(F,w) &\rightarrow w \cdot \frac{\mathrm{ID}^*_F}{(\mathrm{ID}^*_F+1)^2} \\ \mathrm{cH}_q(F,w) &\rightarrow w \cdot \frac{\mathrm{ID}^*_F}{(\mathrm{ID}^*_F+1)(q\,\mathrm{ID}^*_F+1)} \quad \text{for } q \neq 1\,. \end{aligned}$$

An expression for $\widehat{\mathrm{ID}}^*_F$ is obtained by solving the quadratic equation in each case:

$$\begin{aligned} \widehat{\mathrm{ID}}^*_{F,1} &= \alpha(F,w) \pm \sqrt{\alpha^2(F,w)-1} \quad \text{and} \\ \widehat{\mathrm{ID}}^*_{F,q} &= \frac{1}{\sqrt{q}}\left[\beta(F,w,q) \pm \sqrt{\beta^2(F,w,q)-1}\right] \quad \text{for } q \neq 1\,, \end{aligned}$$

where

$$\begin{aligned} \alpha(F,w) &\triangleq \frac{w}{2\widehat{\mathrm{cH}}(F,w)} - 1 \quad \text{and} \\ \beta(F,w,q) &\triangleq \frac{1}{2\sqrt{q}}\left[\frac{w}{\widehat{\mathrm{cH}}_q(F,w)} - (q+1)\right]. \end{aligned}$$

Observe that as q approaches 1, the estimator $\widehat{\mathrm{ID}^*_{F,q}}$ tends to $\widehat{\mathrm{ID}^*_{F,1}}$.

The smaller roots are to be used when ID^*_F is assumed to be less than 1; otherwise, the larger roots should be used. Corollary 2 can be used to show that as w tends to zero, the mean value μ_w of F_w tends to an expression involving ID^*_F (see also [3]):

$$\mu_w \to w \frac{\mathrm{ID}^*_F}{\mathrm{ID}^*_F + 1} .$$

Hence, if $\mathrm{ID}^*_F \geq 1$, then $\frac{\mu_w}{w}$ tends to a value of at least $\frac{1}{2}$. As a decision rule, we can therefore use the larger root if $\frac{1}{k}\sum_{i=1}^{k} X_i \geq \frac{X_k}{2}$; otherwise, we use the smaller root.

We can leverage existing estimators with good convergence properties [4, 10] to estimate the cumulative entropy $\widehat{\mathrm{cH}}(F,w)$ and the cumulative q-entropy $\widehat{\mathrm{cH}}_q(F,w)$. For the cumulative entropy,

$$\begin{aligned}\mathrm{cH}(F,w) &= -\int_0^w F_w(t) \ln F_w(t)\,\mathrm{d}t \\ \widehat{\mathrm{cH}}(F,w) &= -\sum_{j=1}^{k-1} U_{j+1} \cdot \frac{j}{k} \ln \frac{j}{k} ,\end{aligned}$$

where $U_j = X_j - X_{j-1}$, (for $j = 1, \ldots, k$) are the spacings between the successive distance samples, and where for the j-th sample, $F_w(t)$ is straightforwardly estimated as $\widehat{F}_w(t) = \frac{j}{k}$.

For the cumulative q-entropy, we can use a similar estimator

$$\begin{aligned}\mathrm{cH}_q(F,w) &= \frac{1}{(q-1)} \int_0^w F_w(t) - (F_w(t))^q \;\mathrm{d}t \\ \widehat{\mathrm{cH}}_q(F,w) &= \frac{1}{(q-1)} \sum_{j=1}^{k-1} U_{j+1} \cdot \left[\frac{j}{k} - \left(\frac{j}{k}\right)^q \right] .\end{aligned}$$

5.2. Estimation Based on Entropy

Theorem 5 establishes a relationship between entropy (power) and local intrinsic dimensionality which holds in the tail $[0,w]$. We have

$$\exp\left(H(F,w)\right) \;\to\; w \cdot \frac{1}{\mathrm{ID}^*_F} \exp\left(1 - \frac{1}{\mathrm{ID}^*_F}\right) \tag{5}$$

We can compute an entropy estimate $\widehat{H}(F,w)$ using any univariate entropy estimator based on the on the order statistics $0 = X_0 \leq X_1 \leq X_2 \leq, \ldots, X_k = w$ — for example, the popular Kozachenko-Leonenko estimator [11] which is based on nearest neighbor distances. We can then numerically solve Equation 5 to obtain an LID estimate $\widehat{\mathrm{ID}^*_{F,\mathrm{HP}}}$.

5.3. Estimation Based on Cumulative KL Divergence

We can use the KL divergence between cumulative distributions as the basis for estimation. We follow a similar approach to Yari et al. [63], who used cumulative residual KL divergence between $\bar{F}_k$ and $\bar{F}_\theta$ to estimate the parameters of a Weibull distribution. In turn, their approach is similar to the density-based estimation strategy of Basu and Linday [64], who used KL divergence between the probability densities f_k and f_θ.

We let F_w be the true (unknown) empirical distribution conditioned on the lower tail $[0, w]$, from which k independent random observations $\{X_1, X_2, \ldots, X_k\}$ have been drawn. Without loss of generality, we may assume that $0 \leq X_1 \leq \ldots \leq X_k = w$. Next, we let G_w be an ideal distribution conditioned to the same lower tail, with (unknown) local intrinsic dimensionality parameter $\theta = \mathrm{ID}^*_G$. We wish to estimate the intrinsic dimensionality $\hat{\theta}$ as the value for which the KL divergence is minimized. Considering a normalization by w for the limit as w tends to zero, Theorem 7 states that

$$\lim_{w\to 0^+} \frac{d_{\mathrm{KL}}(F,G,w)}{w} \;=\; \frac{1}{\theta+1} - \frac{2\,\mathrm{ID}^*_F+1}{\left(\mathrm{ID}^*_F+1\right)^2} - \lim_{w\to 0^+} \frac{1}{w}\int_0^w F_w(x)\cdot\theta\ln\frac{x}{w}\,\mathrm{d}x\,. \quad (6)$$

In order to minimize $d_{\mathrm{KL}}(F_w, G_w)$ and thereby determine a value of θ that brings G_w as close as possible to the empirical distribution F_w, we therefore take the derivative of Equation 6 with respect to $\theta = \mathrm{ID}^*_G$, for some small positive choice of the tail boundary w. Setting this derivative to zero and then solving the resulting equation for θ,

$$\begin{aligned} 0 &= -\frac{1}{(\theta+1)^2} - \frac{1}{w}\int_0^w F_w(x)\cdot\ln\frac{x}{w}\,\mathrm{d}x \\ (\theta+1)^2 &= -\frac{w}{\int_0^w F_w(x)\cdot\ln\frac{x}{w}\,\mathrm{d}x}\,. \end{aligned} \quad (7)$$

Given our samples $0 \leq X_1 \leq \ldots \leq X_k$, we set the tail bound at $w = X_k$, and approximate F_w through the empirical distribution conditioned on the tail:

$$\hat{F}_k(x) \;=\; \sum_{j=0}^{k-1} \frac{j}{k} I_{[X_j,X_{j+1}]}(x)\,,$$

where $I_{[X_j,X_{j+1}]}$ is the indicator function for the interval between consecutive samples X_j and X_{j+1}.

Substituting $\hat{F}_w$ for F_w in Equation 7, we obtain an expression involving our estimator $\hat{\theta}$ of the local intrinsic dimensionality:

$$(\hat{\theta}+1)^2 \;=\; -\frac{w}{\int_0^w \hat{F}_w(x)\cdot\ln\frac{x}{w}\,\mathrm{d}x}\,.$$

Denoting the origin by $X_0 = 0$,

$$\begin{aligned}
(\hat{\theta}+1)^2 &= -\frac{w}{\sum_{j=0}^{k-1} \int_{X_j}^{X_{j+1}} \hat{F}_w(x) \ln \frac{x}{w} \,\mathrm{d}x} \\
&= -\frac{X_k}{\sum_{j=0}^{k-1} \frac{j}{k} \int_{X_j}^{X_{j+1}} (\ln x - \ln X_k) \,\mathrm{d}x} \\
&= -\frac{X_k}{\frac{1}{k} \sum_{j=1}^{k-1} j \cdot \Big(X_{j+1} \ln \frac{X_{j+1}}{X_k} - X_{j+1} - X_j \ln \frac{X_j}{X_k} + X_j \Big)} ,
\end{aligned}$$

via integration by parts. Simplifying through partial cancellation of terms, we arrive at

$$(\hat{\theta}+1)^2 \;=\; \frac{X_k}{\frac{1}{k} \sum_{j=1}^{k} \Big(X_j \ln \frac{X_j}{X_k} - X_j + X_k \Big)} \;=\; \frac{X_k}{\frac{1}{k} \sum_{j=1}^{k} d_{\mathrm{KL}}(X_j, X_k)} .$$

Solving for $\hat{\theta}$, and noting that Bregman divergences are always non-negative, we obtain the estimator

$$\widehat{\mathrm{ID}}^*_{G,\mathrm{KL}} \;=\; \hat{\theta} \;=\; \frac{\sqrt{X_k} - \sqrt{\frac{1}{k} \sum_{j=1}^{k} d_{\mathrm{KL}}(X_j, X_k)}}{\sqrt{\frac{1}{k} \sum_{j=1}^{k} d_{\mathrm{KL}}(X_j, X_k)}} .$$

Our Bregman KL divergence estimator of LID involves the square roots of two quantities with units of distance: the tail boundary (or tail length) X_k, and the average Bregman KL divergence between the samples and the tail boundary, $\frac{1}{k} \sum_{j=1}^{k} d_{\mathrm{KL}}(X_j, X_k)$. The LID estimator can therefore be interpreted as the relative error incurred when the root distance $\sqrt{X_k}$ is used as an approximation for the root of the average sample divergence.

6. Experimental Results

Our evaluation addresses two main questions regarding the practical impact of our theoretical results:

- How do our four proposed estimators for LID perform in comparison to existing nearest neighbor based estimation approaches?
- What benefits (if any) do the tail entropy power and the tail q-entropy power offer in a supervised learning scenario, as compared to using raw LID estimates? In particular, is using the entropy power as a feature better than the raw LID, for the purpose of training a classification model?

6.1. Estimation of LID

Our four proposed estimators are:

1. Estimation based on cumulative entropy ($\widehat{\mathrm{ID}}^*_{F,1}$, using cH).
2. Estimation based on cumulative q-entropy ($\widehat{\mathrm{ID}}^*_{F,q}$, using cH_q).
3. Estimation based on the entropy power of the tail ($\widehat{\mathrm{ID}}^*_{F,\mathrm{HP}}$, using HP).
4. Estimation based on the cumulative Bregman KL divergence ($\widehat{\mathrm{ID}}^*_{G,\mathrm{KL}}$, using d_{KL}).

We evaluate the performance of our estimators by comparing against two well known baselines from the literature:

- The maximum likelihood (MLE) estimator [27, 3] ($\widehat{\mathrm{ID}}^*_{F,\mathrm{MLE}}$).
- The method of moments (MoM) estimator [3] ($\widehat{\mathrm{ID}}^*_{F,\mathrm{MoM}}$).

The two baselines are simple, well-known approaches that (like all our proposed estimators) are based on nearest neighbor distances.

For our evaluations, we use a variety of synthetic datasets from [65] created for benchmarking intrinsic dimensionality estimation, and which were also used in [3]. Their descriptions and dimensional characteristics are shown in Table 3. Each dataset consists of 10,000 samples. LID values were estimated for each data sample using $k = 100$ nearest neighbors, and then compared against the ground truth to compute the mean absolute error (MAE). The hyperparameter q was set to 0.9 for the estimator based on $\widehat{\mathrm{cH}_q}(F, w)$.

The estimation performances are shown in Table 3, from which we observe the following:

- The baseline $\widehat{\mathrm{ID}}^*_{F,\mathrm{MoM}}$ is consistently worse than the baseline $\widehat{\mathrm{ID}}^*_{F,\mathrm{MLE}}$. This is consistent with the findings in other studies (such as [3]).
- The Bregman KL estimator $\widehat{\mathrm{ID}}^*_{G,\mathrm{KL}}$ is stronger against these benchmarks than the most popular estimator, $\widehat{\mathrm{ID}}^*_{F,\mathrm{MLE}}$ (11 wins over MLE, 5 draws, one loss). These results provide evidence that the Bregman KL estimator may be a very competitive alternative choice to MLE.
- The entropy power estimator $\widehat{\mathrm{ID}}^*_{F,\mathrm{HP}}$ is always worse than $\widehat{\mathrm{ID}}^*_{F,\mathrm{MLE}}$. This is perhaps not surprising, since the entropy power requires the estimation of differential entropy, which is itself known to be a hard problem [66].
- The cumulative entropy estimator $\widehat{\mathrm{ID}}^*_{F,1}$ generally has less error than $\widehat{\mathrm{ID}}^*_{F,q}$ for the low dimensional datasets (m1-m8), but slightly higher error on the high dimensional datasets (m9, m10a, m10b, m10c, m12, m14 and m15).

- The estimators based on cumulative entropy, $\widehat{\mathrm{ID}}^*_{F,1}$ and $\widehat{\mathrm{ID}}^*_{F,q}$, usually incur a higher MAE cost than does $\widehat{\mathrm{ID}}^*_{F,\mathrm{MLE}}$. However, they can have lower error on datasets where the true intrinsic dimension is relatively high (m9, m10a, m10b, m10c, m12, m14 and m15).

Overall, we believe that these four new estimators are interesting additions to the family of LID estimation techniques that further increase the diversity of available approaches. The proposed Bregman KL estimator $\widehat{\mathrm{ID}}^*_{G,\mathrm{KL}}$, as well as being theoretically interesting, was seen to perform more effectively for the benchmark datasets as compared to the standard MLE estimator.

6.2. Use of Entropy Power as Classification Features

We examine how tail entropy power (nHP) can be used instead of LID as a classification feature for adversarial example detection, an application scenario where ID has demonstrated superior performance to measures such as kernel density [67]. Adversarial examples are test input instances that are intentionally engineered to fool deep neural networks. Adversarial detection, which trains a binary logistic regression classifier to decide whether an input sample is adversarial or normal, is one of the most effective defenses against adversarial examples [67, 12, 68]. Here we test the use of entropy power features instead of LID features for adversarial example detection.

We follow the experimental setting of [12] by training 4-layer, 6-layer and 8-layer Convolutional Neural Networks (CNNs) on MNIST [69], SVHN [70] and CIFAR-10 [71] datasets, respectively. We then craft adversarial examples for each CNN model using 3 state-of-the-art attack methods: Fast Gradient Sign Method (FGSM) [72], Projected Gradient Descent (PGD) [73] and Carlini and Wagner (CW) [74]. For the PGD attack, we consider three variants (denoted as PGD-s), where the number of perturbation steps is set at $s = 20$, 40, and 100. For the CW attack, we consider three variants (denoted as CW^c), with attack confidences set at $c = 0\%$, 40%, and 100%.

We compute the (normalized) tail entropy power (nHP) and LID values at each layer of the network for successful adversarial examples as well as their corresponding original (unperturbed) samples. The nHP and LID values for a size-100 minibatch of examples (either adversarial or original) are estimated based on the 20 nearest neighbors found within the same minibatch [12]. For each combination of attack method and CNN model, this process produces one nHP dataset and one LID dataset. We partition each dataset randomly into a training set (80% of the examples) and a test set (the remaining 20%), and train a Logistic Regression (LR) classifier on the training set. Following the same extraction procedure as with nHP and LID, we also test the use of tail q-entropy power (HP$_q$) as the classification feature, for values of q from 0.5 to 1.5. The detection AUC ('Area Under the Curve') results on the test sets are reported in Table 4.

From the results in Table 4, we see that across all attacks and datasets, the use of tail entropy power nHP brings a consistent improvement in detection over the use of LID features. We hypothesize that this is because the entropy power can be interpreted as a diversity, and thus has a natural doubling property making it more suitable as a feature for use in logistic regression. Observe that when F is a (univariate) uniform distance distribution ranging over the interval $[0, w]$, we have $\mathrm{ID}^*_F = 1$ and $\mathrm{nHP}(F, w) = w$. In other words, the entropy power is equal to the "effective diversity" of the distribution (the number of neighbor distance possibilities). Given two different queries, each with its own neighborhood, one query with tail entropy power equal to 2 and the other with tail entropy power equal to 4, we can say that the distance distribution of the second query is twice as diverse as that of the first query. The q-entropy power nHP_q provides more flexibility when used as a classification feature. In most cases, nHP_q can lead to better performance than either nHP or LID by varying q.

The best-performing choices of q reveal an interesting property of the neighborhood distribution of weak versus strong attacks: for the weaker attacks FGSM and CW^0, a smaller choice of q was better at identifying adversarial examples, whereas for strong attacks, larger choices (close to 1) performed better. This is likely due to the tendency for strong attacks to push examples to more sparse regions in the data domain, farther from the underlying data manifold. In these sparse neighborhoods, small choices of q can help increase the discriminability of the entropy power. In practice, for a supervised learning scenario, a value for q could be chosen using a hyperparameter optimization scheme, similar to other hyperparameters, such as trade-off factors in loss functions.

7. Conclusion

In this paper we have established an asymptotic relationship between tail entropy variants and the emerging theory of local intrinsic dimensionality. Our results provide insights into the complexity of data within local neighborhoods, and how they may be assessed. These fundamental discoveries open the door to cross-fertilization between intrinsic dimensionality research and entropy research. They emphasize that for a highly local neighborhood around a query point, (appropriately normalized) information-theoretic quantities are solely dependent on the underlying LID.

We have demonstrated immediate applications of our results for estimation and learning: proposing and evaluating four new estimators, and evaluating the use of entropy power as a representation feature as an alternative to raw LID values. We believe there is considerable scope for use of the normalized entropy power in addition to the local intrinsic dimensionality, as a measure for understanding and assessing changes in time and space.

As future work, we plan to further investigate the generalization and learning behaviors of deep neural networks in light of both local intrinsic dimensionality and tail entropy variants.

Acknowledgments

Michael E. Houle acknowledges the financial support of JSPS Kakenhi Kiban (B) Research Grant 18H03296.

Table 3: Mean absolute error for estimators, with dataset size=10000, and $k = 100$ nearest neighbors as samples for estimation. d is the true intrinsic dimension, and D is the representational dimension. For the sake of conciseness, for the four proposed estimators, the table headings refer to the tail entropy variants employed. The q parameter is fixed at 0.9 for estimation based on cH_q. Lowest mean absolute error shown in bold for each dataset.

Dataset	d	D	MLE	MoM	cH	cH_q	HP	d_{KL}
m1: Uniformly sampled sphere	10	11	1.05	1.06	1.10	1.12	1.29	**1.03**
m2: Affine space	3	5	**0.29**	0.30	0.37	0.39	0.41	**0.29**
m3: Fused figures, concentrated & 3d	4	6	0.57	0.59	0.61	0.63	0.67	**0.56**
m4: Non-linear manifold	4	8	0.45	0.47	0.52	0.53	0.57	**0.43**
m5: 2-d Helix	2	3	**0.18**	0.19	0.29	0.31	0.33	**0.18**
m6: Non-linear manifold	6	36	1.02	1.05	1.03	1.06	1.11	**0.95**
m7: Swiss-Roll	2	3	**0.18**	0.20	0.29	0.31	0.33	**0.18**
m8: Non-linear manifold	12	72	**2.13**	**2.13**	2.62	2.74	2.19	2.19
m9: Affine space	20	20	5.43	5.45	4.63	**4.47**	5.57	5.23
m10a: Uniform distribution	10	11	1.75	1.77	1.53	**1.50**	1.89	1.69
m10b: Uniform distribution	17	18	4.14	4.16	3.51	**3.39**	4.31	3.98
m10c: Uniform distribution	24	25	7.21	7.24	6.19	**5.99**	7.36	6.95
m11: Moebius band with 10 twists	2	3	**0.18**	0.19	0.29	0.31	0.32	**0.18**
m12: Isotropic multivariate Gaussian	20	20	4.39	4.43	3.51	**3.38**	4.58	4.11
m13: Curve	1	13	**0.09**	0.11	0.22	0.32	0.16	**0.09**
m14: Non-linear manifold	18	72	3.21	3.23	2.87	**2.82**	3.46	3.09
m15: Non-linear manifold	24	96	5.18	5.20	4.41	**4.28**	5.42	4.95
Wins-draws-losses over MLE	–	–	–	0-1-16	7-0-10	7-0-10	0-0-17	11-5-1

Table 4: Detection AUC (%) of Logistic Regression (LR) classifiers trained on Local Intrinsic Dimensionality (LID), tail entropy power (nHP) and tail q-entropy power (HP$_q$) features on different types of adversarial examples.

Dataset	Feature	FGSM	PGD-20	PGD-40	PGD-100	CW^{0}	CW^{40}	CW^{100}
MNIST	ID	99.99	99.95	99.95	99.99	96.90	99.96	99.99
	nHP	**100.00**	**100.00**	**100.00**	**100.00**	**99.87**	**100.00**	**100.00**
	nHP$_{q=0.5}$	**100.00**	**100.00**	**100.00**	**100.00**	96.24	99.94	**100.00**
	nHP$_{q=0.6}$	**100.00**	**100.00**	**100.00**	**100.00**	96.35	99.98	**100.00**
	nHP$_{q=0.7}$	**100.00**	**100.00**	**100.00**	**100.00**	96.51	99.99	**100.00**
	nHP$_{q=0.8}$	**100.00**	**100.00**	**100.00**	**100.00**	96.69	**100.00**	**100.00**
	nHP$_{q=0.9}$	**100.00**	**100.00**	**100.00**	**100.00**	96.89	**100.00**	**100.00**
	nHP$_{q=1.1}$	**100.00**	**100.00**	**100.00**	**100.00**	97.26	**100.00**	**100.00**
	nHP$_{q=1.2}$	**100.00**	**100.00**	**100.00**	**100.00**	96.79	**100.00**	**100.00**
	nHP$_{q=1.3}$	**100.00**	**100.00**	**100.00**	**100.00**	97.15	**100.00**	**100.00**
	nHP$_{q=1.4}$	**100.00**	**100.00**	**100.00**	**100.00**	97.73	**100.00**	**100.00**
	nHP$_{q=1.5}$	**100.00**	**100.00**	**100.00**	**100.00**	98.14	**100.00**	**100.00**
SVHN	ID	91.21	94.64	95.86	96.69	95.12	99.90	99.99
	nHP	92.18	95.03	96.10	96.79	**98.62**	**99.92**	**100.00**
	nHP$_{q=0.5}$	**92.47**	94.49	95.72	96.43	95.29	99.88	**100.00**
	nHP$_{q=0.6}$	92.45	94.67	95.81	96.52	94.54	99.89	**100.00**
	nHP$_{q=0.7}$	92.40	94.77	95.90	96.60	93.75	99.90	**100.00**
	nHP$_{q=0.8}$	92.29	94.87	95.98	96.67	92.90	99.91	**100.00**
	nHP$_{q=0.9}$	92.24	94.95	96.04	96.74	91.86	99.91	**100.00**
	nHP$_{q=1.1}$	92.13	**95.09**	**96.15**	**96.84**	89.70	**99.92**	**100.00**
	nHP$_{q=1.2}$	91.92	94.29	95.50	96.13	78.45	99.86	**100.00**
	nHP$_{q=1.3}$	91.86	94.57	95.69	96.35	78.03	99.89	**100.00**
	nHP$_{q=1.4}$	91.88	94.55	95.78	96.38	78.25	99.90	**100.00**
	nHP$_{q=1.5}$	91.94	94.86	96.01	96.65	78.78	**99.92**	**100.00**
CIFAR-10	ID	88.38	98.52	98.87	99.18	85.25	95.84	99.99
	nHP	90.50	**98.99**	99.29	99.47	87.06	97.43	**100.00**
	nHP$_{q=0.5}$	**90.55**	98.90	99.22	99.43	**88.64**	97.28	**100.00**
	nHP$_{q=0.6}$	90.53	98.94	99.26	99.45	87.02	97.36	**100.00**
	nHP$_{q=0.7}$	90.52	98.97	99.29	99.47	85.37	97.42	**100.00**
	nHP$_{q=0.8}$	90.52	**98.99**	**99.30**	**99.48**	83.61	97.46	**100.00**
	nHP$_{q=0.9}$	90.51	**98.99**	**99.30**	**99.48**	81.81	**97.47**	**100.00**
	nHP$_{q=1.1}$	90.50	98.27	98.69	99.06	80.17	95.27	**100.00**
	nHP$_{q=1.2}$	90.51	98.51	98.88	72.17	72.73	95.92	**100.00**
	nHP$_{q=1.3}$	90.50	98.78	99.13	99.37	74.36	96.65	**100.00**
	nHP$_{q=1.4}$	90.52	98.83	99.14	99.38	76.57	96.86	**100.00**
	nHP$_{q=1.5}$	90.50	98.77	99.09	99.33	79.07	96.67	**100.00**

References

[1] M. E. Houle, Dimensionality, discriminability, density and distance distributions, in: IEEE 13th International Conference on Data Mining Workshops, 2013, pp. 468–473.

[2] M. E. Houle, Local intrinsic dimensionality I: an extreme-value-theoretic foundation for similarity applications, in: International Conference on Similarity Search and Applications, 2017, pp. 64–79.

[3] L. Amsaleg, O. Chelly, T. Furon, S. Girard, M. E. Houle, K. Kawarabayashi, M. Nett, Extreme-value-theoretic estimation of local intrinsic dimensionality, Data Mining and Knowledge Discovery 32 (6) (2018) 1768–1805.

[4] A. Di Crescenzo, M. Longobardi, On cumulative entropies, Journal of Statistical Planning and Inference 139 (12) (2009) 4072–4087.

[5] M. Rao, Y. Chen, B. C. Vemuri, F. Wang, Cumulative residual entropy: a new measure of information, IEEE Transactions on Information Theory 50 (6) (2004) 1220–1228.

[6] S. Park, M. Rao, D. W. Shin, On cumulative residual kullback–leibler information, Statistics & Probability Letters 82 (11) (2012) 2025–2032. doi:https://doi.org/10.1016/j.spl.2012.06.015.
URL https://www.sciencedirect.com/science/article/pii/S016771521200226X

[7] A. D. Crescenzo, M. Longobardi, Some properties and applications of cumulative kullback–leibler information, Applied Stochastic Models in Business and Industry 31 (6) (2015) 875–891. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/asmb.2116, doi:https://doi.org/10.1002/asmb.2116.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2116

[8] L. M. Bregman, The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming, USSR Computational Mathematics and Mathematical Physics 7 (1967) 200–217.

[9] C. Tsallis, Possible generalization of Boltzmann-Gibbs statistics, Journal of Statistical Physics 52 (1988) 479–487. doi:10.1007/BF01016429.

[10] C. Calì, M. Longobardi, J. Ahmadi, Some properties of cumulative Tsallis entropy, Physica A: Statistical Mechanics and its Applications 486 (2017) 1012–1021.

[11] L. F. Kozachenko, N. N. Leonenko, A statistical estimate for the entropy of a random vector, Problemy Peredachi Informatsii 23 (1987) 9–16.

[12] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, J. Bailey, Characterizing adversarial subspaces using local intrinsic dimensionality, in: International Conference on Learning Representations, 2018, pp. 1–15.

[13] X. Ma, Y. Wang, M. E. Houle, S. Zhou, S. M. Erfani, S. Xia, S. N. R. Wijewickrema, J. Bailey, Dimensionality-driven learning with noisy labels, in: International Conference on Machine Learning, 2018, pp. 3361–3370.

[14] F. Camastra, A. Staiano, Intrinsic dimension estimation: Advances and open problems, Information Sciences 328 (2016) 26–41.

[15] P. Campadelli, E. Casiraghi, C. Ceruti, A. Rozza, Intrinsic dimension estimation: Relevant techniques and a benchmark framework, Mathematical Problems in Engineering (2015).

[16] P. J. Verveer, R. P. W. Duin, An evaluation of intrinsic dimensionality estimators, IEEE TPAMI 17 (1) (1995) 81–86.

[17] J. Bruske, G. Sommer, Intrinsic dimensionality estimation with optimally topology preserving maps, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (5) (1998) 572–575.

[18] K. W. Pettis, B. T. A., A. K. Jain, R. C. Dubes, An intrinsic dimensionality estimator from near-neighbor information, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1979) 25—37.

[19] G. Navarro, R. Paredes, N. Reyes, C. Bustos, An empirical evaluation of intrinsic dimension estimators, Inf. Syst. 64 (2017) 206–218. doi:10.1016/j.is.2016.06.004.
URL https://doi.org/10.1016/j.is.2016.06.004

[20] I. T. Jolliffe, Principal Component Analysis, Springer, 2002.

[21] J. A. Costa, A. O. Hero III, Entropic graphs for manifold learning, in: The 37th Asilomar Conference on Signals, Systems & Computers, Vol. 1, 2003, pp. 316–320.

[22] M. Hein, J. Y. Audibert, Intrinsic dimensionality estimation of submanifolds in R^d, in: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 289–296.

[23] A. Rozza, G. Lombardi, M. Rosa, E. Casiraghi, P. Campadelli, IDEA: Intrinsic dimension estimation algorithm, in: International Conference on Image Analysis and Processing, 2011, pp. 433–442.

[24] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, P. Campadelli, Novel high intrinsic dimensionality estimators, Machine Learning 89 (1–2) (2012) 37–65.

[25] C. Ceruti, S. Bassis, A. Rozza, G. Lombardi, E. Casiraghi, , P. Campadelli, DANCo: An intrinsic dimensionality estimator exploiting angle and norm concentration, Pattern Recognition 47 (2014) 2569–2581.

[26] E. Facco, M. d'Errico, A. Rodriguez, A. Laio, Estimating the intrinsic dimension of datasets by a minimal neighborhood information, Scientific Reports 7 (12140) (2017).

[27] E. Levina, P. J. Bickel, Maximum likelihood estimation of intrinsic dimension, in: Advances in Neural Information Processing Systems, 2004, pp. 777–784.

[28] B. M. Hill, A simple general approach to inference about the tail of a distribution, The Annals of Statistics 3 (5) (1975) 1163–1174.

[29] K. Johnsson, C. Soneson, M. Fontes, Low bias local intrinsic dimension estimation from expected simplex skewness, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (1) (2015) 196–202.

[30] L. Amsaleg, O. Chelly, M. E. Houle, K. Kawarabayashi, M. Radovanović, W. Treeratanajaru, Intrinsic dimensionality estimation within tight localities, in: Proceedings of the 2019 SIAM International Conference on Data Mining, 2019, p. 181–189.

[31] A. M. Farahmand, C. Szepesvári, J. Y. Audibert, Manifold-adaptive dimension estimation, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 265–272.

[32] P. Tempczyk, R. Michaluk, L. Garncarek, P. Spurek, J. Tabor, A. Golinski, LIDL: local intrinsic dimension estimation using approximate likelihood, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, Vol. 162 of Proceedings of Machine Learning Research, PMLR, 2022, pp. 21205–21231.
URL https://proceedings.mlr.press/v162/tempczyk22a.html

[33] S. Zhou, A. Tordesillas, M. Pouragha, J. Bailey, H. Bondell, On local intrinsic dimensionality of deformation in complex materials, Nature Scientific Reports 11 (10216) (2021).

[34] J. B. Antoinette Tordesillas, Shuo Zhou, H. Bondell, A representation learning framework for detection and characterization of dead versus strain localization zones from pre- to post- failure, Granular Matter 24 (2022).

[35] N. Kambhatla, T. K. Leen, Dimension reduction by local principal component analysis, Neural Computation 9 (7) (1997) 1493–1516.

[36] E. Santos-Fernandez, F. Denti, K. Mengersen, A. Mira, The role of intrinsic dimension in high-resolution player tracking data—Insights in basketball, The Annals of Applied Statistics 16 (1) (2022) 326 – 348. doi:10.1214/21-AOAS1506.
URL `https://doi.org/10.1214/21-AOAS1506`

[37] D. Faranda, G. Messori, P. Yiou, Dynamical proxies of north atlantic predictability and extremes, Sci. Rep. 7 (41278) (2017).

[38] A. Varghese, E. Santos-Fernandez, F. Denti, A. Mira, K. Mengersen, On the intrinsic dimensionality of covid-19 data: a global perspective (2022). doi:10.48550/ARXIV.2203.04165.
URL `https://arxiv.org/abs/2203.04165`

[39] M. E. Houle, X. Ma, M. Nett, V. Oria, Dimensional testing for multi-step similarity search, in: IEEE 12th International Conference on Data Mining, 2012, pp. 299–308.

[40] M. E. Houle, E. Schubert, A. Zimek, On the correlation between local intrinsic dimensionality and outlierness, in: International Conference on Similarity Search and Applications, 2018, p. 177–191.

[41] L. Amsaleg, J. Bailey, D. Barbe, S. M. Erfani, M. E. Houle, V. Nguyen, M. Radovanović, The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality, in: IEEE Workshop on Information Forensics and Security, 2017, pp. 1–6.

[42] L. Amsaleg, J. Bailey, A. Barbe, S. M. Erfani, T. Furon, M. E. Houle, M. Radovanović, X. V. Nguyen, High intrinsic dimensionality facilitates adversarial attack: Theoretical evidence, IEEE Transactions on Information Forensics and Security 16 (2021) 854–865.

[43] A. Ansuini, A. Laio, J. H. Macke, D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, in: Advances in Neural Information Processing Systems, 2019, pp. 6111–6122.
URL `http://papers.nips.cc/paper/8843-intrinsic-dimension-of-data-representations-in-deep-neural-networks.pdf`

[44] P. Campadelli, E. Casiraghi, C. Ceruti, G. Lombardi, A. Rozza, Local intrinsic dimensionality based features for clustering, in: International Conference on Image Analysis and Processing, 2013, pp. 41–50.

[45] K. M. Carter, R. Raich, W. G. Finn, A. O. Hero III, FINE: Fisher information non-parametric embedding, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (11) (2009) 2093–2098.

[46] P. Pope, C. Zhu, A. Abdelkader, M. Goldblum, T. Goldstein, The intrinsic dimension of images and its impact on learning, in: International Conference on Learning Representations, 2021.

[47] H. V. Nguyen, P. Mandros, J. Vreeken, Universal dependency analysis, in: Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM, 2016, pp. 792–800. doi:10.1137/1.9781611974348.89.
URL `https://doi.org/10.1137/1.9781611974348.89`

[48] K. Böhm, F. Keller, E. Müller, H. V. Nguyen, J. Vreeken, CMI: an information-theoretic contrast measure for enhancing subspace cluster and outlier detection, in: Proceedings of the 13th SIAM International Conference on Data Mining, 2013, pp. 198–206. doi:10.1137/1.9781611972832.22.
URL `https://doi.org/10.1137/1.9781611972832.22`

[49] A. Di Crescenzo, M. Longobardi, Stochastic Comparisons of Cumulative Entropies, Springer New York, New York, NY, 2013, pp. 167–182.

[50] S. Baratpour, A. H. Rad, Testing goodness-of-fit for exponential distribution based on cumulative residual entropy, Communications in Statistics - Theory and Methods 41 (8) (2012) 1387–1396. arXiv:https://doi.org/10.1080/03610926.2010.542857, doi:10.1080/03610926.2010.542857.
URL `https://doi.org/10.1080/03610926.2010.542857`

[51] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, Clustering with bregman divergences, J. Machine Learning Research 6 (2005) 1705–1749.

[52] F. Nielsen, R. Nock, Sided and symmetrized bregman centroids, IEEE Trans. Inf. Theory 55 (6) (2009) 2882–2904.

[53] D. T. Pele, E. Lazar, M. Mazurencu-Marinescu-Pele, Modeling expected shortfall using tail entropy, Entropy 21 (12) (2019) 1204.

[54] J. Bailey, M. E. Houle, X. Ma, Relationships between local intrinsic dimensionality and tail entropy, in: Similarity Search and Applications - Proc. of the 14th International Conference, SISAP 2021, Dortmund, Germany, September 29 - October 1, 2021., 2021.

[55] J. Bailey, M. E. Houle, X. Ma, Local intrinsic dimensionality, entropy and statistical divergences, Entropy 24 (9) (2022). doi:10.3390/e24091220.
URL `https://www.mdpi.com/1099-4300/24/9/1220`

[56] M. E. Houle, H. Kashima, M. Nett, Generalized expansion dimension, in: IEEE 12th International Conference on Data Mining Workshops, 2012, pp. 587–594.

[57] D. R. Karger, M. Ruhl, Finding nearest neighbors in growth-restricted metrics, in: Proceedings of the 34th Annual ACM Symposium on Theory of Computing, 2002, pp. 741–750.

[58] J. Karamata, Sur un mode de croissance régulière. Théorèmes fondamentaux, Bulletin de la Société Mathématique de France 61 (1933) 55–62.

[59] S. Coles, J. Bawa, L. Trenner, P. Dorazio, An introduction to statistical modeling of extreme values, Vol. 208, Springer, 2001.

[60] M. E. Houle, Local intrinsic dimensionality II: multivariate analysis and distributional support, in: International Conference on Similarity Search and Applications, 2017, pp. 80–95.

[61] L. Kostal, P. Lansky, O. Pokora, Measures of statistical dispersion based on Shannon and Fisher information concepts, Information Sciences (2013).

[62] A. Anastasiadis, Special issue: Tsallis entropy, Entropy 14 (2) (2012) 174–176. doi:10.3390/e14020174.
URL `https://www.mdpi.com/1099-4300/14/2/174`

[63] G. Yari, A. Mirhabibi, A. Saghafi, Estimation of the weibull parameters by kullback-leibler divergence of survival functions, Appl. Math. Inf. Sci 7 (1) (2013) 187–192.

[64] A. Basu, B. Lindsay, Minimum disparity estimation for continuous models: Efficiency, distributions and robustness, Ann Inst Stat Math 46 (1994) 683–705.

[65] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, P. Campadelli, Novel high intrinsic dimensionality estimators, Machine learning 89 (1-2) (2012) 37–65.

[66] J. Jiao, W. Gao, Y. Han, The nearest neighbor information estimator is adaptively near minimax rate-optimal, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018.
URL `https://proceedings.neurips.cc/paper/2018/file/e9fd7c2c6623306db59b6aef5c0d5cac-Paper.pdf`

[67] R. Feinman, R. R. Curtin, S. Shintre, A. B. Gardner, Detecting adversarial samples from artifacts, arXiv preprint arXiv:1703.00410 (2017).

[68] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, in: Advances in Neural Information Processing Systems, 2018, pp. 7167–7177.

[69] Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, L. D. Jackel, Handwritten digit recognition with a back-propagation network, in: Advances in Neural Information Processing Systems, 1990, pp. 396–404.

[70] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, Reading digits in natural images with unsupervised feature learning, in: NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2011, p. 5.

[71] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images (2009).

[72] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015.

[73] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.

[74] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: IEEE Symposium on Security and Privacy, IEEE, 2017, pp. 39–57.