Expansion of the Kullback-Leibler Divergence, and a New Class of Information Metrics

David J. Galas [1,*], Gregory Dewey [2], James Kunert-Graf [1] and Nikita A. Sakhanenko [1]

[1] Pacific Northwest Research Institute, 720 Broadway, Seattle, WA 98122, USA; jkunert@pnri.org (J.K.-G.); nsakhanenko@pnri.org (N.A.S.)

[2] Albany College of Pharmacy and Health Sciences, 106 New Scotland Avenue, Albany, NY 12208, USA; greg.dewey@acphs.edu

[*] Correspondence: dgalas@pnri.org

Abstract: Inferring and comparing complex, multivariable probability density functions is fundamental to problems in several fields, including probabilistic learning, network theory, and data analysis. Classification and prediction are the two faces of this class of problem. This study takes an approach that simplifies many aspects of these problems by presenting a structured, series expansion of the Kullback-Leibler divergence—a function central to information theory—and devise a distance metric based on this divergence. Using the Möbius inversion duality between multivariable entropies and multivariable interaction information, we express the divergence as an additive series in the number of interacting variables, which provides a restricted and simplified set of distributions to use as approximation and with which to model data. Truncations of this series yield approximations based on the number of interacting variables. The first few terms of the expansion-truncation are illustrated and shown to lead naturally to familiar approximations, including the well-known Kirkwood superposition approximation. Truncation can also induce a simple relation between the multi-information and the interaction information. A measure of distance between distributions, based on Kullback-Leibler divergence, is then described and shown to be a true metric if properly restricted. The expansion is shown to generate a hierarchy of metrics and connects this work to information geometry formalisms. An example of the application of these metrics to a graph comparison problem is given that shows that the formalism can be applied to a wide range of network problems and provides a general approach for systematic approximations in numbers of interactions or connections, as well as a related quantitative metric.

1. Introduction

The problem of representing or inferring dependencies among variables is central to many fields. It is fundamental to data analysis of large data sets, as well as describing and approximating the behavior of physical, chemical, and biological systems with many modes, particles, or component interactions. These complex systems are usually modeled by graphs or hypergraphs, and their inference from data represents a central problem. Statistical inference and machine learning approaches have been directed at this general class of inference problem for many years, and the literature of physical chemistry, among other, related fields, abounds with approaches to the general representation problem [1–4]. A key related problem is that of measuring the difference between approximations, a useful metric of probability distributions. The relationships between neural networks, statistical mechanics, and this general class of problem have also been explored [5]. While certainly not the

only indication of complexity, the number of variables that interact or are functionally interdependent is a very important characteristic of the complexity of a system. This work engages a number of these problems.

A central function of information theory, the Kullback-Leibler divergence, can be shown to be close to the heart of these problems. It is the goal of this paper to describe an approach that simplifies some aspects of these problems in a different way, by focusing on interesting and useful symmetries of entropy and "relative entropy" and the Kullback-Leibler divergence (K-L, referred to as the "divergence" in the rest of this paper.) Particularly important in practical applications is the divergence between the "true," multivariable probability density function (pdf) and any approximation of it [6] as are specific metric measures of the distances between approximations.

The paper is structured around recognition and exploitation of several properties of this divergence, a central function in information theory. The divergence admits of a simple series expansion with increasing numbers of variables in each successive term. This study affects this expansion of the multivariable cross-entropy (or relative entropy) term of the divergence using the Möbius duality between multivariable entropies and multivariable interaction information [7–10]. This allows a series expansion in the number of interacting variables, which can be used as an approximation parameter: the more interactions considered, the more accurate the approximation. The derivation of some known factorizations of the pdf is then illustrated by truncating the expansion at small numbers of variables. Well-known simple approximations emerge, including the Kirkwood superposition approximation at three variables. This is a widely used approximation in the theory of liquids [1,2,11]. Other approximations, like the seminal approximation method of Chow and Liu [12], is closely connected to the expansion. This specific approach will not be expanded upon here, but will be explored and this connection will be extended in future work.

The divergence expansion proposed herein is entirely general, can be extended to any degree, and leads to a number of useful relationships with other information theory measures. The following section defines a new simple metric between probability density functions and show that it meets all the requirements of a true metric.

Unlike the approach of the Jensen-Shannon divergence—which is a measure based on symmetrizing the K-L divergence [13]—or that which uses the Fisher metric to embed the functions in a Riemannian manifold [14], this metric provides a large class of information metrics that calculate distances directly, and thereby easily measure the relations between approximations, among other applications. This work examines a few cases of specific pdf function classes (e.g., Gaussian, Poisson) and finds explicit forms for the functions. Finally, it examines briefly the metric distances implied by different truncations of the divergence expansion, and describes an application to the character description of networks.

2. Expanding the Divergence

Consider a set of variables $\nu = \{_i\}$, for which we have many values constituting a data set. The concepts of maximum entropy and minimum divergence have been used to devise approaches to the inference of the best estimate of the true probability density function from a data set. The relation between the "true" and an approximate probability density function (pdf) is best characterized by the Kullback-Leibler divergence. If the true pdf is $P(\nu)$ and an approximation to it is $P'(\nu)$ then the divergence is given by

$$D\left(P \parallel P'\right) = \sum_s P(s) \log\left(\frac{P(s)}{P'(s)}\right) \tag{1}$$

where s traverses all possible states of ν. The approximated entropy (called the cross-entropy) is defined as

$$H'(\nu) = -\sum_s P(s) \log\left(P'(s)\right), \tag{2}$$

so the divergence is simply the difference between the true entropy and the cross entropy:

$$D(P \parallel P') = H'(\nu) - H(\nu) \tag{3}$$

In this form, it is clear that the approximate joint entropy must be greater than $H(\nu)$ since the divergence is always non-negative [6]. This is a consequence of the well-known Jensen's inequality. If P' is an approximation to P, then as the approximation gets better and better the divergence converges to zero. The approximation of the joint entropy is the measure of the accuracy of the approximation and minimizing H' (under some set of constraints or assumptions) must be optimum. Using other information theory measures related to the joint entropies in Equation (3), however, can also be used to good effect.

Specifically, the Möbius inversion relation between the entropy and interaction information is used [8–10]. This relationship can be written

$$H(\nu) = \sum_{\tau \subseteq \nu} (-1)^{|\tau|+1} I(\tau) \tag{4a}$$

where the sum is over all subsets of ν. H and I can be exchanged in this symmetric form of the relation and the equation still holds.

$$I(\nu) = \sum_{\tau \subseteq \nu} (-1)^{|\tau|+1} H(\tau) \tag{4b}$$

The symmetry derives from the inherent structure of the subset lattice, which is a hypercube [15]. Inserting the joint entropy expression into Equation (3) gives a sum over all subsets of the variables

$$D(P \parallel P') = \sum_{\tau \subset \nu} (-1)^{|\tau|+1} I'(\tau) - H(\nu) \tag{5a}$$

Now if terms are grouped by the number of variables in the subset and introduce notation to indicate the size of each of the subsets, the sum is rearranged as an expansion.

$$D(P \parallel P') = \sum_{m=0}^{|\nu|} \sum_{\tau_m \subset \nu} (-1)^{|\tau_m|+1} I'(\tau_m) - H(\nu) \tag{5b}$$

The symbol τ_m indicates a subset of variables of cardinality m ($|\tau_m| = m$). This then becomes an expansion in degrees, m, the number of variables. The full expansion includes, and terminates with, the full set of variables, ν.

3. Truncations of the Series

If the expansion is truncated at various degrees (numbers of variables), setting all interaction information terms above the truncation point equal to zero, a series of increasingly accurate, but ever more complex, approximations is generated. Truncation generates a specific probability density function relation, in the form of a specific factorization, by setting to zero an interaction information expression in the form of a sum of entropies (Note that setting $I(\tau_m) = 0$ does not imply that higher terms, $I(\tau_{m+1})$ etc., are also zero. The truncation approximation necessarily sets all higher terms to zero). This is a key result of the expansion of the divergence. Truncation, and a factorization of the probability density function, results from setting all the higher interaction informations to zero. Thus, the expansion represents a method for approximation and simplification that specifically limits the degree of variable dependencies.

The approximations that result from truncating the expansion of the cross entropy at the first few degrees are familiar ones. Since the number of dependent variables is the driver of the complexity, we begin with pairwise approximations and stepwise increase the number. The first few truncations show the character of this expansion process.

3.1. Truncation at m = 1

Considering the simplest possible truncation, setting all but the first term equal to zero:

$$D\left(P \parallel P'\right) = D_1\left(P \parallel P'\right) = \sum_i H'(X_i) - H(\nu) \equiv A_1 - H(\nu). \tag{6a}$$

This truncation requires that all of the *m* = 2 terms, for all pairs, are zero

$$I'\left(X_i, X_j\right) = 0 \; \forall i, j \tag{6b}$$

that from Equations (2) and (4b) implies independence of all pairs of variables, and the simplest factorization is

$$P\left(X_i, X_j\right) = P(X_i)P\left(X_j\right), \; \forall X_i, X_j \in \nu \tag{6c}$$

This determines all pairwise probability functions, but note that it actually does not determine the form of a three-way or higher pdf. The truncation requires, however, that the three-variable interaction information is zero: $I\left(X_i, X_j, X_k\right) = 0$. This fact combined with Equations (4b) and (6b) results in a full three-way factorization of the pdf:

$$P\left(X_i, X_j, X_k\right) = P(X_i)P\left(X_j\right)P(X_k) \tag{7a}$$

As variables are added we can use the interaction information recursion relation (Equation (12) to derive the higher pdf's implied by the truncation. Finally, the *m* = 1 truncation yields the fully factored pdf of

$$P(\nu) = \prod_i P(X_i) \tag{7b}$$

Note that this same result derives from minimizing the expression for the divergence in (6a), since this expression is a minimum when $\sum_i H(X_i) = H(\nu)$.

The implication of this result for data analysis is simply the solution of the classic problem of determining the optimal pdf under the assumption of the independence of all variables, fixed expectation values being defined by parameters usually represented by Lagrange multipliers. The physics implication would be simply that of independent particles, observables, etc., which leads to a simple Boltzmann distribution in equilibrium. The pdf becomes more complex, of course, if we truncate the expansion at a higher level.

3.2. Truncation at m = 2

This truncation requires that $I'\left(X_i, X_j, X_k\right) = 0$, which from Equation (4b) implies this factorization

$$P_2'(\tau_2) = \prod_{i>j} \frac{P(X_i, X_j)}{P(X_i)} = \prod_{i>j} P(X_i | X_j) \quad \forall \tau_2 \subset \nu \tag{8a}$$

Let us denote the cross entropy term for this truncation as A_2. Then we have

$$D_2\left(P \parallel P'\right) = \sum_i H'(X_i) - \sum_{\tau_2 \subseteq \nu} I'(\tau_2) - H(\nu) \equiv A_2 - H(\nu) \tag{8b}$$

$$A_2 = -\sum_{i>j}\left(H'(X_i) + H'\left(X_j\right) - H'\left(X_i, X_j\right)\right) + \sum_i H'(X_i). \tag{8c}$$

The cross entropy term A_2 is determined by the pdf P', and from (8b) above it can be seen that the minimization of the divergence is the same as truncation of the expansion. This is equivalent to the approximation made by Chow-Liu [12]. In physical terms, this is the same as ignoring all but pairwise interaction terms in a Hamiltonian, and is precisely the probabilistic version of the

Kirkwood superposition approximation [1,2,11]. This approximation is used in the physics of dense multiparticle systems, like liquids. The resulting pair correlation function is used in deriving many of the thermodynamic properties of liquids. Singer [1] related this to the more general theoretical constructs like the Percus-Yevick approximation and the Bogoliubov-Born-Green-Kirkwood-Yvon (BBGKY) hierarchy.

3.3. Truncation at m = 3

Parallel to the above we can express the truncation approximation at the next level using three terms:

$$D_3\left(P \parallel P'\right) = \sum_i H'(X_i) - \sum_{\tau_2 \subseteq \nu} I'(\tau_2) + \sum_{\tau_3 \subseteq \nu} I'(\tau_3) - H(\nu) \equiv A_3 - H(\nu) \tag{9a}$$

In terms of the cross entropies the term A_3 becomes

$$\begin{aligned}
A_3 =\ & -\sum_{i>j>k}\left(H'(X_i) + H'(X_j) + H'(X_k) - H'(X_i, X_j) - H'(X_i, X_k) - H'(X_j, X_k) + H'(X_i, X_j, X_k)\right) \\
& -\sum_{i>j}\left(H'(X_i) + H'(X_j) - H'(X_i, X_j)\right) + \sum_j H'(X_j)
\end{aligned} \tag{9b}$$

As before we see that the truncation, assuming the four-variable cross-interaction information is zero, is the same as minimizing the divergence D_3. Both imply that the approximation to the pdf is

$$P'_3(\nu) = \prod_{i>j>k} \frac{P\left(X_i X_j X_k\right) P(X_k) P\left(X_j\right)}{P(X_i X_k) P\left(X_j X_k\right)} \tag{10}$$

Note that A_3 is also expressed simply in terms of the deltas used in the analysis of dependency and as a partial measure of complexity [16]. For three variables this quantity is the same as the conditional mutual information, as can be seen from the recursion relation, Equation (12).

$$A_3 - \sum_i H'(X_i) = -\sum_{k; i>j} I(X_i X_j | X_k) \tag{11}$$

The approximation indicated on the right-hand side is based on the cross entropy approximation, that $H = H'$. The truncation of the expansion, leading to more complex representations of the variable interactions, can be taken to higher levels, of course, which leads in turn to higher-level, more complex factorizations of the pdf. These factorizations are most simply seen by setting the cross interaction information for m variables equal to zero and inferring the implied pdf factors.

4. A Relation to the Deltas

The truncation relation implies another simple equivalence that has direct intuitive meaning, and connects in a simple way to the differential interaction information [10]. From the general recursion relation for the interaction information we can derive a set of simple equivalences. For the set ν_n of n variables the general, multi-variable recursion relation for the interaction information is

$$I(\nu_n) = I(\nu_{n-1}) - I(\nu_{n-1} | X_n) \tag{12}$$

for all n choices of X_n, where the set ν_{n-1} is the set missing X_n. Thus the truncation, setting the left side to zero, implies exactly n relations, one for each choice of i:

$$I(\nu_{n-1} | X_i) = I(\nu_{n-1}) \tag{13}$$

The implication of the truncation criterion for the divergence at $m = n$, then, is that the interaction information, conditioned on each variable of a set ν_n is the same as the interaction information of

the remaining $n-1$ variables. Note that the conditional in Equation (12) is the same (within a sign) as the asymmetric delta function for n variables [16], so the truncation of the divergence is seen to be equivalent to a simplification and truncation of the asymmetric delta. For truncation at $m=2$, this would mean that all conditional mutual informations are equal to the mutual information itself: equivalent to specifying independence of the conditional variable.

5. Multi-Information

It is easy to show that the truncation embodied in Equation (11) also implies a simple relation between the "multi-information" (called "complete correlation" by Watanabe [3]) and the interaction information. The multi-information is defined as $\Omega(v_n) = \sum_i H(X_i) - H(v_n)$. This quantity is often used as a measure of overall multivariable dependence, since it goes to zero if all variables are independent. It is always positive, but has several drawbacks in that it does not distinguish at all the degrees of dependence (number of variables), and is not a metric.

This study will not show the elementary proof of the general case of truncation at n variables here, but illustrate a simple expression for multi-information in terms of the interaction information with the three- and four-variable cases. For the case of truncation at $n=3$ ($I(v_3) = 0$), the relation is simply the sum of all three mutual informations:

$$\Omega(X, Y, Z) = I(X, Y) + I(X, Z) + I(Y, Z) \tag{14a}$$

This is easy to see by direct calculation using the marginal entropies. For $n=4$.

$$\begin{aligned}\Omega(X, Y, Z, W) &= I(X, Y, Z) + I(X, Y, W) + I(X, Z, W) + I(Y, Z, W) - I(X, Y) - I(X, Z)\\ &\quad -I(X, W) - I(Y, Z) - I(Y, W) - I(Z, W) = \sum_{\tau_3} I(\tau_3) - \sum_{\tau_2} I(\tau_2)\end{aligned} \tag{14b}$$

The relation (14a) is strongly intuitive in the sense that if the three-variable interaction information is zero, the multi-information is simply the sum of the mutual information for all three pairs. A similar, but less intuitive, relationship is embodied in the four-variable case, Equation (14b), and the general case is suggested.

The divergence expansion can also be expressed using the multi-information in a limited number of variables, as well as a series of truncation-approximate probability density functions, in the following way. Consider a series of functions $\{P_m\}$ related to the true, untruncated, probability density function, such that P_m is the pdf of m variables that results from setting the interaction information equal to zero for subsets τ_m. Then we have

$$I(\tau_m) = 0 \Rightarrow H(\tau_m) = \sum_{\eta \subset \tau_m} (-1)^{|\eta|+1} H(\eta); \ \tau_m \subseteq v, \ |v| = n \tag{15}$$

The divergence converges to zero for the series $\{P_m\}$ as the number of variables increases to n.

$$\lim_{m \to n} D(P \| P_m) = 0 \tag{16}$$

The divergence therefore induces a topology on the series of functions. The proof of (16) follows directly from the definitions.

Note that the multi-information is not a metric, and that a metric specifically gives a distance measure between different pdfs. This is a problem that has received much attention as a metric provides a clear measure of the function space, we can complete this formalism around the K-L divergence and its approximations by devising a simple pdf metric.

6. Information Geometry and a Simple Metric

Although it is sometimes thought of as a distance measure between probability distributions, the Kullback–Leibler divergence is not a true metric. Among the disqualifying properties is its asymmetry. Much work has, however, been devoted to the development of geometric measures of information, particularly in differential geometry [14], and symmetric divergences have been defined [13]. A derivative form, the Hessian, of the divergence does yield a metric tensor known as the Fisher information metric. This is a Riemannian metric tensor, and has been used extensively. While having a real metric is essential to a complete quantitative theory, it is even more useful if it is relatively simple and direct. Finite distances between functions in the differential manifold of the Fisher metric must be determined by integration along geodesics. Simpler metrics allow the direct calculation of the distance between probability density functions. We now describe such a simple information metric.

Consider the problem of comparing two approximate distributions, $R(\nu)$ and $S(\nu)$, using another pdf, $P(\nu)$, as a reference function. The K-L divergence is used to define a metric simply as the absolute value of the difference between two K-L divergences using the same reference function. This definition is embodied in the following equation.

$$\mathfrak{D}_P(R \parallel S) \equiv |D(P \parallel R) - D(P \parallel S)| = \left| \sum_s P(s) log\left(\frac{R(s)}{S(s)} \right) \right| \qquad (17)$$

It is next established that $\mathfrak{D}_P(R \parallel S)$ does indeed have the properties of a metric on a function space. A metric has the following four properties, which as shown are fulfilled by the following definition:

(1) Non-negativity: $\mathfrak{D}_P(R \parallel S) \geq 0$ is assured because $P(\nu) \geq 0$ and the absolute value in Equation (17) assures a summation that is non-negative.

(2) Identity of indiscernibles: $\mathfrak{D}_P(R \parallel S) = 0$ when $R(\nu) = S(\nu)$. $\mathfrak{D}_P(R \parallel S) = \left| \sum_s P(s)\left(log\left(\frac{R(s)}{P(s)} \right) - log\left(\frac{S(s)}{P(s)} \right) \right) \right| = 0$. For a metric it must also be true that $\mathfrak{D}_P(R \parallel S) \neq 0$, unless $R(\nu) = S(\nu)$, otherwise the metric is a pseudometric. This condition does not hold for all choices of P, R and S and therefore the metric property may apply only to specific spaces, and must be examined in each case. We illustrate this later for some specific cases.

(3) Symmetry: $\mathfrak{D}_P(R \parallel S) = \mathfrak{D}_P(S \parallel R)$.

$$\mathfrak{D}_P(R \parallel S) = \left| \sum_s P(s)\left\{ log\left(\frac{R(s)}{P(s)} \right) - log\left(\frac{S(s)}{P(s)} \right) \right\} \right| = \mathfrak{D}_P(S \parallel R)$$

(4) Subadditivity, obeying the triangle inequality:

$$\mathfrak{D}_P(R \parallel S) \leq \mathfrak{D}(R \parallel Q) + \mathfrak{D}(Q \parallel S)$$

$$\begin{aligned} \mathfrak{D}_P(R \parallel S) &= \left| \sum_s P(s) log\left(\frac{R(s)}{S(s)} \right) \right| = \left| \sum_s P(s) log\left(\frac{R(s)}{Q(s)} \right)\left(\frac{Q(s)}{S(s)} \right) \right| \\ &= \left| \sum_s P(s) log\left(\frac{R(s)}{Q(s)} \right) + \sum_s P(s) log\left(\frac{Q(s)}{S(s)} \right) \right| \\ &\leq \left| \sum_s P(s) log\left(\frac{R(s)}{Q(s)} \right) \right| + \left| \sum_s P(s) log\left(\frac{Q(s)}{S(s)} \right) \right| = \mathfrak{D}_P(R \parallel Q) + \mathfrak{D}_P(Q \parallel S) \end{aligned}$$

The inequality holds because the sums are real numbers, and the triangle inequality applies.

$\mathfrak{D}$ is therefore a true metric on the function space of pdfs, which can be used directly as a measure of information distance. In some cases and function spaces, however, there are subspaces that are true metrics and other that are pseudometrics, having some distinct functions that have zero distance from each other. Since the metric is determined by a reference function, $\mathfrak{D}$ represents a large class of metrics,

each determined by the choice of reference function. This study now examines some properties of these metrics.

An intriguing similarity of the metric, the distance between functions defined by a third function, lies in Bayesian statistics. It can be said that by defining the reference pdf, $P(\nu)$ as a prior pdf, $\mathfrak{D}(R \parallel S)$ measures the distance between two posterior functions, $R(\nu)$ and $S(\nu)$. By measuring the distance between successive posteriors, one can monitor the convergence of Bayesian updating to a steady state distribution. The distance measure can also be used to assess quantitatively how close different posterior models are to each other. A Dirichlet metric, $\mathfrak{D}_D$ for example, can be defined if the reference pdf, or prior, was a Dirichlet distribution, or a uniform, or a Gaussian metric if the reference was uniform or Gaussian.

7. Special Metrics

The fact that $P(\nu)$ defines a metric on a function space inspires the authors of this paper to ask what specific functional forms yield metric spaces with particular properties. We could define a uniform probability density over the variable set ν, which leads to the very simple expression for this metric, $\mathfrak{D}_0$.

$$\mathfrak{D}_0(R \parallel S) = \frac{1}{\mathbb{N}} \left| \sum_s log\left(\frac{R(s)}{S(s)} \right) \right| \tag{18}$$

where $\mathbb{N}$ is the number of values that the total set of variables can take on (consider it a vector). This is always a metric since $\mathfrak{D}_P(R \parallel S) \neq 0$, unless $R(\nu) = S(\nu)$. An interesting class of metrics is generated by choosing a Gaussian reference.

If the functions R and S are also Gaussian, a particularly simple expression can be illustrated for distances for the case of a single variable. Let the reference function be defined as a normal distribution with variance σ^2 and mean, μ, designated

$$P(x) = \frac{1}{(2\pi)^{1/2}\sigma} exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \tag{19a}$$

and the functions to be measured are:

$$R(x) = \frac{1}{(2\pi)^{1/2}\sigma_1} exp\left\{ -\frac{(x-\mu_1)^2}{2\sigma_1^2} \right\} \tag{19b}$$

$$S(x) = \frac{1}{(2\pi)^{1/2}\sigma_2} exp\left\{ -\frac{(x-\mu_2)^2}{2\sigma_2^2} \right\} \tag{19c}$$

The distance between R and S then is:

$$\mathfrak{D}_G(R \parallel S) = \left| \int P(x) \left\{ log\left(\frac{R(x)}{S(x)} \right) \right\} dx \right|$$

which can easily be evaluated. Using the simple properties of Gaussians we have

$$\mathfrak{D}_G(R \parallel S) = \left| log\left(\frac{\sigma_2}{\sigma_1} \right) - \left\{ \frac{\sigma^2}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) + \frac{(\mu-\mu_1)^2}{\sigma_1^2} - \frac{(\mu-\mu_2)^2}{\sigma_2^2} \right\} \right| \tag{20}$$

This explicit expression for distance has some simple special cases. First, if the standard deviations of R and S are the same, then the distance is dependent only on their mean values, independent of the standard deviation of the reference function. Likewise, if the means of R and S are the same, then the distance depends only on the standard deviation, independent of the mean of the reference function.

There is another, special case worth pointing out. If the reference function is chosen to be a Dirac delta function, which could be considered to be the limiting case of a Gaussian with vanishing standard deviation, the expression of Equation (20) simplifies further. The key property of the Dirac delta function, $\delta(x - x_0)$, is that the integral over x with any function yields a specific value of the function, $\int \delta(x - x_0)f(x)dx = f(x_0)$. The metric space is defined by the single parameter of the mean of the reference function, μ. The distance expression, $\mathfrak{D}_\delta$, is then

$$\mathfrak{D}_\delta(R \parallel S) = \left| \left\{ \frac{(\mu_1 - \mu)^2}{\sigma_1^2} - \frac{(\mu_2 - \mu)^2}{\sigma_2^2} \right\} - log\left(\frac{\sigma_2}{\sigma_1}\right) \right| \tag{21}$$

If the standard deviations of R and S are equal, the expression is extremely simple. To assure that this is a metric rather than a pseudometric, the function space and the reference function can be chosen such that $\mu \leqq \mu_1, \mu_2$, for example. In this case, the distance between R and S is proportional to the squares of the distance from the reference mean. If the function space includes those with different σ the ratios of μ to σ defines the distance. It should be noted that in this case there are many functions that are zero distance apart, but they are a very restricted class. If the reference mean is set at zero, it is clear that the relevant measure of distance is just the squares of the ratios of the mean to standard deviation. This one-dimensional case has a simple geometric interpretation, which is illustrated in Figure 1 in two different ways.

(**a**)   (**b**)
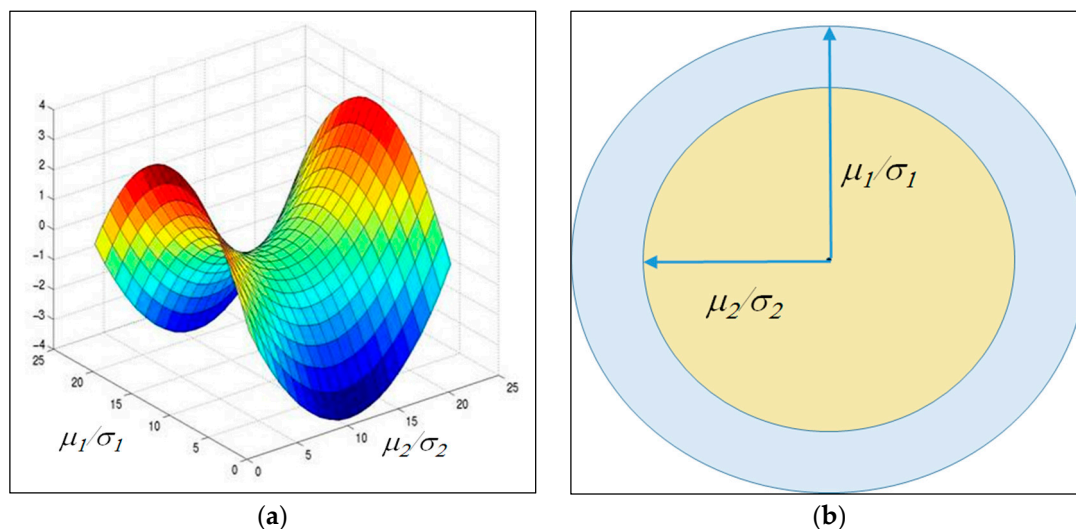
Figure 1. (**a**) The metric distance between Gaussian's R and S (Equation (21)) for the Dirac delta function reference metric with mean at zero can be represented as a hyperbolic function (distance is the vertical axis where for simplicity the metric distance is the deviation from the zero plane—the absolute value) with a saddle point. (**b**) Another geometric metaphor for the distance is the area of the blue annular region divided by π is the distance, where $\frac{\mu_1}{\sigma_1} > \frac{\mu_2}{\sigma_2}$. The expression for this area is simply: Area = $\frac{1}{\pi}\left(\frac{\mu_1}{\sigma_1} + \frac{\mu_2}{\sigma_2}\right)\left(\frac{\mu_1}{\sigma_1} - \frac{\mu_2}{\sigma_2}\right)$.

Notice that with the exception of the single log term on the right-hand side of Equation (20), the expression is a quadratic form in the ratios of mean to standard deviation of R and S, and of the ratios of each of these standard deviations to the reference standard deviation.

In general, the Dirac delta reference function metric does not carry much information about the functions themselves, but if the function class is restricted, it becomes both more interesting and useful. These logs, whose difference is the metric in Equation (16), are often called "surprisals" in information theory. So in this case, the metric is essentially how much more surprising R is than S at any specific point. It should be mentioned that if multiple delta function metrics are used where the distance

coordinates for each surprisal point t distances between R and S leads naturally to a multi-dimensional space representation of the log ratios. A three-dimensional representation, for example, reflects the three chosen points where the functions are compared.

Another interesting metric space results from selecting all three functions, the reference and the measured functions, as Poisson distributions. These discrete valued functions, $P(k, \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$, yield a particularly simple metric distance. If the reference function has parameter λ, and the other two λ_1 and λ_2 the distance is simply

$$\mathfrak{D}_P(\lambda_1 \parallel \lambda_2) = |(\lambda_1 - \lambda \log \lambda_1) - (\lambda_2 - \lambda \log \lambda_2)| \tag{22}$$

Of course, the distance vanishes when λ_1 goes to λ_2. If the reference λ, is much smaller than the other two, $\lambda \ll \lambda_1, \lambda_2$ the distance is linear in the difference between them, while if it is very much larger, $\lambda \gg \lambda_1, \lambda_2$ the distance is proportional to the difference of the logs of the λ's. In these cases the distance is a true metric, with no distinct functions at zero distance. If the reference λ were set to one, on the other hand, it is easy to see that there are pairs of functions, on either sides of one that have zero distance. For that choice of reference function, the result is a pseudometric.

There are a very large number of possible special metrics based on a wide range of possible continuous distributions that could be used as reference functions, many of which lead to interesting functional expressions. To explore these further see the comprehensive list of such functions in the compendium "Field Guide to Continuous Probability Distributions," which is available from Gavin Crooks's website (http://threeplusone.com/FieldGuide.pdf).

8. Measuring the Independence of Variable Subsets

Next consider comparing the probability of a given, single variable, $R(X_1)$ with that of the conditional probability of that variable given the remaining set $\nu - \{X_1\}$ of $|\nu| - 1$ variables, $R(X_1|\nu - \{X_1\})$. If the chosen variable is independent of the others, so that we have $R(X_1) = R(X_1|\nu - \{X_1\})$, then the distance is zero: $\mathfrak{D}(R(X_1) \parallel R(X_1|\nu - \{X_1\})) = 0$. This result is independent of the reference function, $P(\nu)$.

Also, we have

$$\begin{aligned}
\mathfrak{D}(R(X_1) \parallel R(X_1|\nu - \{X_1\})) &= \left| \sum_s P(s) \left\{ log\left(\frac{R(s_1)}{R(s_1|s - \{s_1\})} \right) \right\} \right| \\
&= \left| \sum_s P(s) \left\{ log\left(\frac{R(s_1)R(s - \{s_1\})}{R(s_1|s - \{s_1\})R(s - \{s_1\})} \right) \right\} \right|
\end{aligned}$$

where s_1 denotes a state of a single variable X_1 and $s - \{s_1\}$ denotes the state of all other variables of ν.

Therefore, $\mathfrak{D}(R(X_1) \parallel R(X_1|\nu - \{X_1\})) = \mathfrak{D}(R(X_1)R(\nu - \{X_1\}) \parallel R(\nu))$. Generalizing from a single variable to a subset ν' we have:

$$\mathfrak{D}\big(R(\nu') \parallel R(\nu'|\nu - \nu')\big) = \mathfrak{D}\big(R(\nu')R(\nu - \nu') \parallel R(\nu)\big) \tag{23}$$

which is, of course, dependent on the reference function, P, except in the limit where the distance goes to zero.

9. Comparing Approximations from Different Truncated Series

An application of these metric spaces lies in the area of statistical physics: for example, an application that considers reduced probability distribution functions to approximate the true distribution functions. High degree of interactions, highly multivariable, and non-equilibrium problems defined by trajectories could be directly approached with this apparatus. These approximations have often involved physically motivated simplifying truncation relationships, such as those discussed above. The formalism developed here can be used to calculate the distance between probability functions that

are truncated at different levels of approximation. This allows an assessment of the convergence of higher level truncations, in terms of the distance converging to zero.

The approximate functions that result from truncations of the variable number expansion at different numbers of variables can now be directly compared with a quantitative metric. The truncations described above define the forms of density functions as factors. The actual pdf's are determined by the true, or reference pdf.

Comparing distributions truncated at the first and second order, the probability functions P_1' and P_2' are determined by the factorizations of Equations (6c) and (7a). The distance between these two truncation approximations, relative to the reference function, is then

$$\mathfrak{D}\left(P_2' \parallel P_1'\right) = \left|D\left(P_2' \parallel P\right) - D\left(P_1' \parallel P\right)\right| = |A_2 - A_1| \tag{24a}$$

Referring to Equations (6a) and (8c) this expression simplifies to

$$\mathfrak{D}\left(P_2' \parallel P_1'\right) \ = |A_2 - A_1| = \sum_{i>j}\left(H'(X_i) + H'(X_j) - H'(X_i, X_j)\right) = \sum_{i>j} I'(X_i, X_j)$$
$$\mathfrak{D}\left(P_2' \parallel P_1'\right) = \sum_{\tau_2 \subset \nu} I'(\tau_2) \tag{24b}$$

Recall that this is the sum of "cross" mutual information between all pairs of variables defined by the reference function. The distance of Equation (24b) represents the distance between functions of pairwise dependence and independence. In general, the distance between two different truncation approximations can be seen easily from the expansion of Equation (4). The distance is simply the absolute value of the sum of the terms present in only one of the truncated series.

10. Application to Networks

Metric spaces can be used to devise a simple and direct way of estimating the distance between two networks, which is a problem that has attracted attention for many years. Begin by considering networks in terms of subsets of dependent variables, where the variables are nodes, so that if there are only pairwise dependencies we have a graph. Furthermore, the measures of dependence can be considered as weights for the edges, so that the mutual information between variable pairs provides these weights. For higher-degree dependencies the corresponding network is a hypergraph. Consider here how our metric spaces apply to graphs. The general formalism described here can be used for hypergraphs by direct analogy. Extending the analysis to hypergraphs adds some additional consideration that we do not address here, but the parallel is clear. A graph describing the dependencies present in a dataset, for example, would result from truncation of the divergence at the $m = 2$ level. Following the previous discussion, the divergence can be used then to quantitate the approximation represented by the graph. Moreover, the metrics previously defined now present a simple way to calculate real metric distances between graphs.

Consider an example: choose for simplicity a uniform probability density over the n variable set, ν, as a reference function, resulting in a very simple expression for a metric, $\mathfrak{D}_0$, as shown in Equation (18). If $\mathbb{N}$ is the number of values that the set of variables can take (consider it a vector) and two acyclic graph, R and S, are defined by density functions, as in Equation (8a):

$$R(X_1, X_2, X_3, \ldots, X_n) = \prod_{i>j} R\left(X_i | X_j\right)$$
$$S(X_1, X_2, X_3, \ldots, X_n) = \prod_{i>j} S\left(X_i | X_j\right) \tag{25}$$

Using the uniform density metric, $\mathfrak{D}_0$, and the relationship between the mutual informations for these distributions, we have then

$$\mathfrak{D}_0(R \parallel S) = \frac{1}{\mathbb{N}} \sum_s \left| \sum_{i>j} log \frac{R(X_i|X_j)}{S(X_i|X_j)} \right| = \frac{1}{\mathbb{N}} \left| \sum_{i>j} I_R(X_i, X_j) - I_S(X_i, X_j) \right| \tag{26}$$

where the sums are over the weights of the edges. Note that using graphs with the weights as mutual information between nodes to describe a dataset is exactly like the Chow-Liu approach. In this example, the distance is simply a sum of differences for the same sets of nodes in the two graphs. This is a simple result, but with an interesting subtlety. These differences may be positive or negative, and the $\mathfrak{D}_0$ is the absolute value of the overall sum. The absence of an edge (zero mutual information) in one graph may be compensated for by a different absence in the other in order to leave the distance the same. This formalism guarantees that the distance is a true metric distance. Other reference functions, which also produce metrics, lead to more complex results that are not so easy to visualize, and the extension to hypergraphs, which is a natural extension of the above, leads to results that are increasingly difficult to see. In any case, this demonstrates the use of our formalism in network comparison based on information functions. The further applications of these network comparison results will be explored in a later paper.

11. Conclusions

The Kullback-Leibler divergence as a means of comparing probability density functions has played a central role information theory and has been used for several practical purposes in data analysis, machine learning, and model inference. It has provided ways to explore some key ideas in fields from information theory to thermodynamics. This paper shows that it can continue to yield new results. The divergence can be expanded in the number of interacting variables, yielding a systematic hierarchy of truncations, approximations to the probability density function, which is effectively a hierarchy of factorizations. Factorizations, since they focus on the kinds and degrees of independence, are central and can be thought of as hierarchies of spaces of ever more complex functions, with evermore complex dependencies. The relationship between the set of entropies and the set of interaction informations through the Möbius inversion relation is a fundamental symmetry that is manifest here, but the full symmetry spectrum is yet deeper. It reflects a number of relationships with other information-related measures that are based on this symmetry [15,16]. Since these relations can also be used to express the cross entropy differently, they should generate different expansions of the divergence with different structures. This intriguing area may itself yield additional, useful applications, and has yet to be explored.

As noted in the introduction, there are several areas of potential application of these ideas. The relation of one level of the truncation hierarchy to the Chow-Liu approximation [12] previously noted immediately suggests the extension of the Chow-Liu algorithm to higher levels—Chow-Liu-like hypergraphs. This remains to be explored fully and will be addressed in a future publication. Other applications to networks and network inference are suggested by the notion of the metric classes based on specific reference functions. It is interesting that the divergence provides the basis for a new finite difference metric that gives measures of distances between pdf's, real or estimated, continuous or discrete. An example is given of the application of these metrics to a graph comparison problem. This example shows that the entire formalism can be brought to bear on a wide range of network problems. It should be possible to simplify graph distance measures, given a set of specific constraints, by optimizing the choice of a reference function. Tailoring the metric to specific classes of graphs, for example, should enable simplification of model inference in some cases. These ideas will be addressed in future work.

If the constraint of specific function forms for the pdfs—the exponential family of functions like Gaussians, for example—were added, the natural extension leads to a number of specific

approximations and metric form. The considerations here raise the question of the strategy that should be used to select a reference function. There are at least two considerations. If it is important that a true metric, rather than a pseudometric, be provided then the reference function and the function space should be selected to provide that property. It could be as simple as picking the right parameter range for the pdf, as illustrated for the Poisson and Gaussian pdfs. Another consideration is the accuracy of the calculated distances. This will depend on what the relevant functions are expected to be, so that a reference function might be chosen near this region in function space to avoid having to subtract two large divergences to find the distance. These issues are important and practical, and somewhat problem specific, for the effective use of these concepts. They will be systematically considered in future work.

The relationship of the metric presented herein (Equation (18)) to the Fisher information metric can be obtained from the convergence to zero of this distance. There are a wide variety of metrics that can be derived by symmetrizing the divergence in various ways. The Jensen-Shannon divergence is one of these, but there are several others that use variations on the theme of averaging the cross entropy terms in various ways. The proposed metric here is the first to our knowledge to use a third probability density function to define the character of the metric space, though there has been some consideration of the information-geometric interpretation of the difference between K-L divergences [17]. The connection of K-L divergence differences using a third distribution to expected log-likelihood ratios and to their use in building Riemannian metrics has also been discussed previously [17,18]. These approaches are quite distinct from that presented in this paper but may be connected by future work.

In general, this approach has the advantage, as indicated by the simple examples shown here, that the metric space can be tailored to the character of any function space. It is suggested that Equation (18) defines what could be interpreted in a general sense as a finite difference form of the Fisher metric. The metric can also be used directly to compare Bayesian estimators as the pdf is iteratively updated, to measure convergence.

The application of the general approach described here to a wide range of multivariable problems, including data analysis, model inference, multivariable physical problems, and problems involving complex biological systems, should be useful in providing new analysis methods and new insights.

Author Contributions: David Galas initiated the research, and ideas and mathematical contributions to the paper were made by all authors. Gregory Dewey initiated the idea of the metric and all contributed to the proofs and examples. David Galas led the writing of the paper, and all authors edited and corrected the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Singer, A. Maximum entropy formulation of the Kirkwood superposition approximation. *J. Chem. Phys.* **2004**, *121*, 3657–3666. [CrossRef] [PubMed]
2. Kirkwood, J.G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300. [CrossRef]
3. Watanabe, S. Information theoretic analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82. [CrossRef]
4. Nemenman, I.; Shafee, F.; Bialek, W. Entropy and Inference, Revisted. In *Advances in Neural Information Processing Systems 14*; The MIT Press: Cambridge, MA, USA, 2002.
5. Lin, H.; Tegmark, M. Why does deep and cheap learning work so well? *arXiv* **2016**, arXiv:1608.08225.
6. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
7. McGill, W.J. Multivariate information transmission. *Psychometrika* **1954**, *19*, 97–116. [CrossRef]

8. Bell, A.J. *The Co-Information Lattice*; ICA: Nara, Japan, 2003.

9. Jakulin, A.; Bratko, I. Quantifying and visualizing attribute interactions: An approach based on entropy. *arXiv* **2004**, arXiv preprint cs/0308002.

10. Sakhanenko, N.A.; Galas, D.J. Biological Data Analysis as an Information Theory Problem: Multivariable Dependence Measures and the Shadows Algorithm. *J. Comput. Biol.* **2015**, *22*, 1–20. [CrossRef] [PubMed]

11. Cochran, R.V.; Lund, L.H. On the Kirkwood Superposition Approximation. *J. Chem. Phys.* **1964**, *41*, 3499–3504. [CrossRef]

12. Chow, C.K.; Liu, C.N. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory* **1968**. [CrossRef]

13. Nielsen, F. A family of statistical symmetric divergences based on Jensen's inequality. *arXiv* **2010**, arXiv:1009.4004.

14. Amari, S.; Nagaoka, H. *Methods of Information Geometry, Volume 191 of Translations of Mathematical Monographs*; American Mathematical Society: Providence, RI, USA, 2000.

15. Galas, D.J.; Sakhanenko, N.A. Multivariate information measures: A unification using Möbius operators on subset lattices. *arXiv* **2016**, arXiv:1601.06780v2.

16. Galas, D.J.; Sakhanenko, N.A.; Skupin, A.; Ignac, T. Describing the Complexity of Systems: Multivariable "Set Complexity" and the Information Basis of Systems Biology. *J. Comput. Biol.* **2014**, *21*, 118–140. [CrossRef] [PubMed]

17. Cheng, Y.; Hua, X.; Wang, H.; Qin, Y.; Li, X. The geometry of signal detection with applications to radar signal processing. *Entropy* **2016**, *18*, 381. [CrossRef]

18. Eguchi, S.; Copas, J. Interpreting Kullback-Leibler divergence with the Neyman-Pearson Lemma. *J. Multivar. Anal.* **2006**, *97*, 2034–2040. [CrossRef]