entropy

MDPI

Article

Some Order Preserving Inequalities for Cross Entropy and Kullback–Leibler Divergence

Mateu Sbert [1,2,*], Min Chen [3], Jordi Poch [2] and Anton Bardera [2]

1 College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
2 Graphics and Imaging Laboratory, University of Girona, Campus Montilivi, 17003 Girona, Spain; poch@imae.udg.edu (J.P.); anton.bardera@ima.udg.edu (A.B.)
3 Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK; min.chen@oerc.ox.ac.uk
* Correspondence: mateusbert@mac.com

Received: 8 November 2018; Accepted: 8 December 2018; Published: 12 December 2018

Abstract: Cross entropy and Kullback–Leibler (K-L) divergence are fundamental quantities of information theory, and they are widely used in many fields. Since cross entropy is the negated logarithm of likelihood, minimizing cross entropy is equivalent to maximizing likelihood, and thus, cross entropy is applied for optimization in machine learning. K-L divergence also stands independently as a commonly used metric for measuring the difference between two distributions. In this paper, we introduce new inequalities regarding cross entropy and K-L divergence by using the fact that cross entropy is the negated logarithm of the weighted geometric mean. We first apply the well-known rearrangement inequality, followed by a recent theorem on weighted Kolmogorov means, and, finally, we introduce a new theorem that directly applies to inequalities between K-L divergences. To illustrate our results, we show numerical examples of distributions.

Keywords: cross entropy; Kullback–Leibler divergence; likelihood; Kolmogorov mean; generalized mean; weighted mean; stochastic dominance; stochastic order

1. Introduction

Cross entropy and Kullback–Leibler (K-L) divergence are applied to and discussed in studies on machine learning [1], visualization (e.g., to interpret the pipeline of visualization using the cost–benefit model [2] and several other applications [3]), computer graphics [4], and many other fields. Cross entropy and K-L divergence are two fundamental quantities of information theory. Cross entropy gives the average code length needed to represent one distribution by another distribution, and the excess code needed over the optimal coding is given by the K-L divergence. As cross entropy is just the negated logarithm of likelihood, minimizing cross entropy means maximizing likelihood, and thus, cross entropy is widely used now for optimization in machine learning. K-L divergence also stands independently as a widely used metric for measuring the difference between two distributions, and it appears in many theoretical results. As examples, the definition of mutual information between two variables is a K-L divergence, and the absolute difference of K-L divergences with respect to a third distribution was recently used to define a family of metrics [5].

Recently, order invariance for inequalities between means was presented [6]. Since likelihood is the weighted geometric mean of the representing distribution weighted by the data distribution, likelihood and thus cross entropy will share the order invariance properties too. Order invariance properties are characterized by the stochastic order between weights or data distributions [7]. Here, we further develop this idea in order to compare representations of different data by the same distribution, and we extend it to a new order, K-L dominance, that allows us to define inequalities between representations of the same data by different distributions. Finally, we establish the connection between the two orders. We also give numerical examples to illustrate the theoretical results.

This paper is organized as follows. After this introduction, in Section 2, we recapitulate the meaning of cross entropy and its relationship to likelihood and K-L divergence, and we obtain the first theoretical results based on the rearrangement inequality. Next, in Section 3, we apply the invariance results for weighted means to cross entropy. In Section 4, we introduce a new order, K-L dominance, that allows us to establish inequalities between K-L divergences and give the relationship between stochastic and K-L order. Finally, in Section 5, we present our conclusions and future work.

2. Cross Entropy as a Measure of Goodness of Representation

Consider two probability distributions, $\{\alpha_k\}_{k=1}^n$, $\{x_k\}_{k=1}^n$, for all k $\alpha_k \geq 0$, $x_k > 0$, $\sum_k \alpha_k = \sum_k x_k = 1$ (when not explicit, the sum limits will be understood to be between 1 and n). Cross entropy $CE(\alpha_k, x_k)$ (to avoid cluttered notation we will write $CE(\alpha_k, x_k)$ instead of $CE(\{\alpha_k\}, \{x_k\})$) is defined as $CE(\alpha_k, x_k) = \sum_k \alpha_k \log \frac{1}{x_k} = -\sum_k \alpha_k \log x_k$ and can be written as $CE(\alpha_k, x_k) = H(\alpha_k) + D_{\mathrm{KL}}(\alpha_k, x_k)$ (similar notation to $CE(\alpha_k, x_k)$), where $H(\alpha_k) = -\sum_k \alpha_k \log \alpha_k$ is the entropy of distribution $\{\alpha_k\}$, and $D_{\mathrm{KL}}(\alpha_k, x_k) = \sum_k \alpha_k \log \frac{\alpha_k}{x_k} \geq 0$ is the Kullback–Leibler divergence of $\{\alpha_k\}$ and $\{x_k\}$ distributions. It represents the average length of code per symbol needed to represent $\{\alpha_k\}$ using $\{x_k\}$. In the context of coding, we do not let any component of $\{x_k\}$ be zero, as the corresponding component of $\{\alpha_k\}$ would not be coded; thus, $D_{\mathrm{KL}}(\alpha_k, x_k)$ is well defined. The minimum code length to code $\{\alpha_k\}$ is obtained by taking $\{\alpha_k\} \equiv \{x_k\}$, because $D_{\mathrm{KL}}(\alpha_k, \alpha_k) = 0$. Minimum code length will be between $H(\alpha_k)$ and $H(\alpha_k) + 1$ bit (Huffman coding, [8]).

Cross Entropy and Likelihood

Observe that cross entropy is the negated logarithm of likelihood. Let us suppose a distribution with n states, which are guessed to be $\{x_1, x_2, \ldots, x_n\}$, and a relative frequency of realizations $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ of the different states. That is, we guess that the distribution $\{x_k\}$ represents the collected data $\{\alpha_k\}$. Likelihood is then used as a measure of how well the distribution represents the data. The likelihood $\mathcal{L}(\alpha_k, x_k)$ of $\{x_k\}$ is the true distribution of the data with relative frequency $\{\alpha_k\}$, and $\mathcal{L} = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n} = \Pi_k x_k^{\alpha_k}$. Observe that $\mathcal{L}(\alpha_k, x_k)$ is the weighted geometric mean of $\{x_1, x_2, \ldots, x_n\}$ with weights $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$. The likelihood represents the probability of $\{x_k\}$ being the true distribution of the known data $\{\alpha_k\}$. As $CE(\alpha_k, x_k) = -\log \mathcal{L}(\alpha_k, x_k)$, maximizing the likelihood is equivalent to minimizing the cross entropy.

Now, let us suppose we take an arbitrary $\{x_k\}$ distribution to code $\{\alpha_k\}$. Without loss of generality, we will consider two or more sequences to be equally ordered, or in the same order, when, by the same permutation of indexes, we can make it such that all the sequences are increasing. We can state the following theorem.

Theorem 1. *Consider the distribution $\{\alpha_k\}$ and the sequence of strictly positive numbers $\{x_1, x_2, \ldots, x_n\}$, $\sum_k x_k = 1$. When $\{x_1, x_2, \ldots, x_n\}$ is permuted so that $\{x_k\}$ is ordered in the same order as $\{\alpha_k\}$, then the cross entropy $CE(\alpha_k, x_k)$ takes a global minimum, and, correspondingly, the likelihood $\mathcal{L}(\alpha_k, x_k)$ takes a global maximum.*

Proof. Using the rearrangement inequality [6,9] and the definition of cross entropy, the minimum of the sum of the product $\Pi_k \alpha_k \log \frac{1}{x_k}$ is realized when the two sequences are ordered inversely to each other. That means the $\{\log x_k\}$ sequence will be ordered as the $\{\alpha_k\}$ sequence, and, by the monotonicity of the logarithmic function, the $\{x_k\}$ sequence will be ordered as the $\{\alpha_k\}$ sequence. □

As $\{\alpha_k\}$, and thus $H(\alpha_k)$, is fixed, $D_{\mathrm{KL}}(\alpha_k, x_k)$ also takes the global minimum when $\{x_k\}$ is ordered as $\{\alpha_k\}$. Also, when there are elements repeated in $\{\alpha_k\}$, there can be non-ordered $\{x_k\}$ distributions with the same minimum cross entropy value. The extreme case is when $\alpha_k = 1/n$ for all k, and all orderings of $\{x_k\}$ will give the same result.

Observe that by Theorem 1, if we have a series of distributions $\{\alpha_{t,k}\}_{t=1}^m$ that are all ordered in the same order, which we can consider to be increasing, and want to represent them with a

single distribution $\{x_k\}$—to be chosen by permuting the elements in $\{x_1, x_2, \ldots, x_n\}$—then the best distribution will be the one for which we select the elements in $\{x_1, x_2, \ldots, x_n\}$ in increasing order too.

Example 1. *For any increasing sequence* $\{\alpha_k\}$*, we have that*

1. $CE(\{\alpha_k\}, \{2/9, 3/9, 4/9\}) \leq CE(\{\alpha_k\}, \{4/9, 3/9, 2/9\})$,
2. $CE(\{\alpha_k\}, \{2/9, 3/9, 4/9\}) \leq CE(\{\alpha_k\}, \{4/9, 2/9, 3/9\})$, *and so on.*

However, what happens when we have to choose between two distributions, $\{x_k\}$ and $\{x'_k\}$, to represent a set of increasing distributions, $\{\alpha_{t,k}\}_{t=1}^{m}$? Can we find which one is better? Section 4 provides some answers to this. Before answering this question, we introduce the concept of stochastic dominance in Section 3.

3. First Stochastic Order Dominance

When for all increasing distributions $\{x_k\} : \sum_k \alpha'_k x_k \leq \sum_k \alpha_k x_k$, we say that $\{\alpha_k\}$ stochastically dominates $\{\alpha'_k\}$, and we denote this relation as $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$ [7]. This is a partial order on the set of distributions.

The necessary and sufficient conditions for a sequence $\{\alpha_k\}$ to stochastically dominate another one $\{\alpha'_k\}$ are given in the following theorem [6].

Theorem 2. *Given that the sequences* $\{\alpha_k\}$ *and* $\{\alpha'_k\}$ *are positive and add up to 1, the following conditions are equivalent:*
(1) The sequence $\{\alpha_k\}$ *stochastically dominates the sequence* $\{\alpha'_k\}$
(2)

$$\begin{aligned} \alpha'_1 &\geq \alpha_1 \\ \alpha'_1 + \alpha'_2 &\geq \alpha_1 + \alpha_2 \\ \ldots &\geq \ldots \\ \alpha'_1 + \alpha'_2 + \ldots + \alpha'_n &= \alpha_1 + \alpha_2 + \ldots + \alpha_n \end{aligned} \tag{1}$$

(3)

$$\begin{aligned} \alpha'_n &\leq \alpha_n \\ \alpha'_n + \alpha'_{n-1} &\leq \alpha_n + \alpha_{n-1} \\ \ldots &\leq \ldots \\ \alpha'_n + \alpha'_{n-1} + \ldots + \alpha'_1 &= \alpha_n + \alpha_{n-1} + \ldots + \alpha_1 \end{aligned} \tag{2}$$

(4) For all increasing sequences $\{x_k\}$ *and for any* $f(x)$ *strictly monotonous function,* $f^{-1}\left(\sum_k \alpha'_k f(x_k)\right) \leq f^{-1}\left(\sum_k \alpha_k f(x_k)\right)$ *(quasi-arithmetic, or Kolmogorov, mean), whenever* $f(x)$ *is applicable.*
(5) There exists a strictly monotonous function $f(x)$ *such that for all increasing sequences* $\{x_k\}$*:*

$$f^{-1}\left(\sum_k \alpha'_k f(x_k)\right) \leq f^{-1}\left(\sum_k \alpha_k f(x_k)\right) \tag{3}$$

(quasi-arithmetic, or Kolmogorov, mean), whenever $f(x)$ *is applicable.*

A sufficient condition for $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$ is given by the following theorem [6,10].

Theorem 3. *Given the distributions* $\{\alpha_k\}, \{\alpha'_k\}$*, a sufficient condition for* $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$ *is that whenever* $i \leq j$*, then* $\frac{\alpha'_i}{\alpha'_j} \geq \frac{\alpha_i}{\alpha_j}$.

By Theorem 3, if $\{\alpha_k\}$ is increasing, then $\{\alpha_k\} \succ_{st} \{1/n\}$, where $\{1/n\}$ is the uniform distribution; if $\{\alpha_k\}$ is decreasing, then $\{1/n\} \succ_{st} \{\alpha_k\}$; if $\{\alpha_k\}$ is increasing and $\{\alpha'_k\}$ is decreasing, then $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$.

Observe now that, by condition (5) of Theorem 2, if Equation (3) applies for one strictly monotonous function, then, for the equivalence between conditions (4) and (5), it applies for any other strictly monotonous function. Then, by the definition of cross-entropy,

Theorem 4. *Given two distributions $\{\alpha_k\}$ and $\{\alpha'_k\}$, $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$ if and only if for all increasing distributions $\{x_k\}$, $CE(\alpha_k, x_k) \leq CE(\alpha'_k, x_k)$ (or $\mathcal{L}(\alpha_k, x_k) \geq \mathcal{L}(\alpha'_k, x_k)$).*

Theorem 4 means that, if $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$, then, taking all possible increasing representations $\{x_k\}$, the code length for $\{\alpha_k\}$ will be shorter than for $\{\alpha'_k\}$, and the likelihood for $\{\alpha_k\}$ will be bigger than the likelihood for $\{\alpha'_k\}$. Observe also that if $\{\alpha_k\}$ is an increasing sequence, then $\{\alpha_k\} \succ_{st} \{1/n\}$; thus, the representation of $\{\alpha_k\}$ by an increasing distribution $\{x_k\}$ will be shorter than the representation of uniform distribution $\{1/n\}$, and the likelihood of $\{\alpha_k\}$ will be bigger than that of $\{1/n\}$.

Corollary 1. *For any increasing distributions $\{x_k\}$ and $\{\alpha_k\}$, we have that $CE(\alpha_k, x_k) \leq CE(1/n, x_k)$ (or $\mathcal{L}(\alpha_k, x_k) \geq \mathcal{L}(1/n, x_k)$).*

Indeed, the above result is valid too if, by an index permutation, we can bring $\{x_k\}$, $\{\alpha_k\}$ to be in the same order. Thus,

Corollary 2. *For any distributions $\{x_k\}$ and $\{\alpha_k\}$ that are ordered equally, we have that $CE(\alpha_k, x_k) \leq CE(1/n, x_k)$.*

Now, suppose $\{\alpha_k\}$ is decreasing; then, $\{1/n\} \succ_{st} \{\alpha_k\}$. Also, for any distribution $\{x_k\}$, $CE(1/n, x_k) \geq CE(1/n, 1/n) = \log n$, and thus, the following corollary,

Corollary 3. *For any increasing distribution $\{x_k\}$ and decreasing distribution $\{\alpha_k\}$, we have that $\log n \leq CE(1/n, x_k) \leq CE(\alpha_k, x_k)$.*

Also, by reordering the indexes, Corollary 3 can be extended to

Corollary 4. *For any distributions $\{x_k\}$ and $\{\alpha_k\}$ that are inversely ordered to each other, we have that $\log n \leq CE(1/n, x_k) \leq CE(\alpha_k, x_k)$.*

Observe now that if $\{\alpha_k\}$ is increasing and $\{\alpha'_k\}$ is decreasing, then, $\{\alpha_k\} \succ_{st} \{\alpha'_k\}$, and thus,

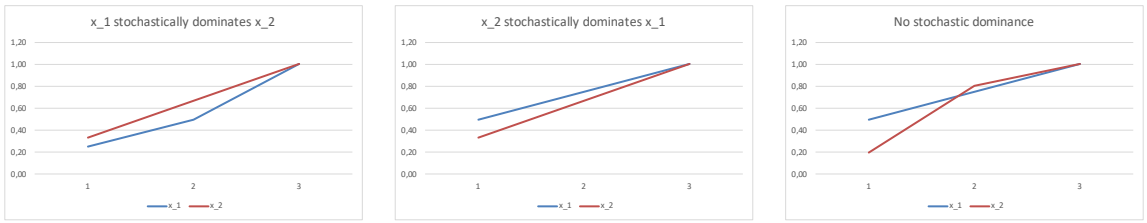
Corollary 5. *For any increasing distribution $\{\alpha_k\}$ and decreasing distribution $\{\alpha'_k\}$, we have that for any increasing distribution $\{x_k\}$, $CE(\alpha_k, x_k) \leq CE(\alpha'_k, x_k)$.*

Corollary 5 is also a direct consequence of Corollaries 1 and 3.

Example 2. *We have:*

1. $\{1/4, 1/4, 1/2\} \succ_{st} \{1/3, 1/3, 1/3\}$
2. $\{1/2, 1/4, 1/4\} \not\succ_{st} \{1/3, 1/3, 1/3\}$
3. $\{1/3, 1/3, 1/3\} \succ_{st} \{1/2, 1/4, 1/4\}$
4. $\{1/2, 1/4, 1/4\} \not\succ_{st} \{1/5, 3/5, 1/5\}$
5. $\{1/5, 3/5, 1/5\} \not\succ_{st} \{1/2, 1/4, 1/4\}$

In Figure 1, we plot the sums in Equation (1) corresponding to relation 1 and relations 2–3, 4–5 in Example 2. Observe that the sums in Equation (1) correspond to the cumulative distribution function (cdf). A first-order stochastic dominance implies that there is no crossing of the respective cdf's.



Relation 1 Relations 2 and 3 Relations 4 and 5

Figure 1. Plotting the sums in Equation (1) for relations 1–5 in Example 2.

Discussion

We show above how the interpretation of cross entropy as a weighted mean allows for obtaining interesting order invariance results for the cross entropy of two data distributions represented by the same distribution (correspondingly for likelihood). Since, in our results, the distribution of the data is variable while the representative distribution is fixed, the resulting cross entropy depends on both entropy and K-L divergence. In the next section, we consider the data fixed and the representative distribution variable; thus, the comparison of cross entropies of a given data distribution for two representative distributions will be equivalent to the comparison of their K-L divergences.

4. A New Partial Order: K-L Dominance

We extend here the results of Section 2. We first define a new partial order between distributions and call it K-L dominance.

Definition 1. *K-L dominance. We say that distribution* $\{x_k\}$ *K-L-dominates distribution* $\{x'_k\}$*, written symbolically as* $\{x_k\} \succ_{\mathrm{KL}} \{x'_k\}$*, when, for all increasing distributions* $\{\alpha_k\}$*,* $D_{\mathrm{KL}}(\alpha_k, x_k) \leq D_{\mathrm{KL}}(\alpha_k, x'_k)$*.*

Observe that $\{x_k\} \succ_{\mathrm{KL}} \{x'_k\} \not\Leftrightarrow \{x_k\} \succ_{st} \{x'_k\}$, because, in general, $D_{\mathrm{KL}}(\alpha_k, x_k) \leq D_{\mathrm{KL}}(\alpha_k, x'_k) \not\Leftrightarrow D_{\mathrm{KL}}(x_k, \alpha_k) \leq D_{\mathrm{KL}}(x'_k, \alpha_k)$.

From the definition of cross entropy, $\{x_k\} \succ_{\mathrm{KL}} \{x'_k\}$ is equivalent to—for all increasing $\{\alpha_k\}$—$CE(\alpha_k, x_k) \leq CE(\alpha_k, x'_k)$, and thus, for all increasing $\{\alpha_k\}$ sequences, coding them with $\{x_k\}$ will always generate a shorter codification than coding them with $\{x'_k\}$.

The following theorem holds:

Theorem 5. *A necessary and sufficient condition for distributions* $\{x_k\}$*,* $\{x'_k\}$ *to hold* $\{x_k\} \succ_{\mathrm{KL}} \{x'_k\}$ *(equivalent to—for all increasing distributions* $\{\alpha_k\}$*—*$D_{\mathrm{KL}}(\alpha_k, x_k) \leq D_{\mathrm{KL}}(\alpha_k, x'_k)$*, or* $CE(\alpha_k, x_k) \leq CE(\alpha_k, x'_k)$*) or* $\mathcal{L}(\alpha_k, x_k) \geq \mathcal{L}(\alpha_k, x'_k)$ *is the following:*

$$\begin{aligned} x'_n &\leq x_n \\ x'_n x'_{n-1} &\leq x_n x_{n-1} \\ \cdots &\leq \cdots \\ x'_n x'_{n-1} \ldots x'_1 &\leq x_n x_{n-1} \ldots x_1 \end{aligned} \tag{4}$$

Observe that Theorem 1 is a particular case of Theorem 5, as a sequence rearranged in increasing order will always fill the second members of inequalities in Equation (4) with respect to all other rearrangements. Thus, an increasing distribution $\{x_k\}$ K-L-dominates all the distributions obtained by permuting the values of $\{x_k\}$. Observe also that Equation (4) condition makes the $\succ_{\mathrm{KL}}$ order reflexive,

antisymmetric, and transitive, and thus, it is a partial order. To prove Theorem 5, let us consider first the following lemma:

Lemma 1. *Consider the sequences of n numbers $\{w_k\}$ and $\{y_k\}$. Then, conditions (1) and (2) are equivalent: (1) for any sequence of n positive numbers in increasing order $\{z_k\}$, the following inequality holds:*

$$\sum_{k=1}^{n} w_k z_k \leq \sum_{k=1}^{n} y_k z_k, \tag{5}$$

(2) the following inequalities hold:

$$\begin{aligned} w_n &\leq y_n \\ w_n + w_{n-1} &\leq y_n + y_{n-1} \\ \cdots &\leq \cdots \\ w_n + \ldots + w_2 &\leq y_n + \ldots + y_2 \\ w_1 + \ldots + w_{n-1} + w_n &\leq y_1 + \ldots + y_{n-1} + y_n. \end{aligned} \tag{6}$$

(7)

Proof. That (1) $\Rightarrow$ (2) is immediate, considering, respectively, the z_k sequences $\{0, 0, \ldots, 1\}, \{0, 0, \ldots, 1, 1\}, \ldots, \{0, 1, \ldots, 1\}, \{1, 1, \ldots, 1\}$. Let us see now that (2) $\Rightarrow$ (1): Define for $1 \leq n$, $A_k = \sum_{j=1}^{k} y_{n-j+1}$, $A'_k = \sum_{j=1}^{k} w_{n-j+1}$, and $A_0 = A'_0 = 0$, $z_0 = 0$. Observe that condition (2) is equivalent to—for all k—$A_k - A'_k \geq 0$. Observe also that the $\{z_k\}$ sequence augmented with z_0 is still increasing. We have

$$\begin{aligned} & \sum_{k=1}^{n} y_{n-k+1} z_{n-k+1} - \sum_{k=1}^{n} w_{n-k+1} z_{n-k+1} \\ = & \sum_{k=1}^{n} (y_{n-k+1} - w_{n-k+1}) z_{n-k+1} \\ = & \sum_{k=1}^{n} (A_k - A_{k-1} - A'_k + A'_{k-1}) z_{n-k+1} \\ = & \sum_{k=1}^{n} (A_k - A'_k) z_{n-k+1} - \sum_{k=1}^{n} (A_{k-1} - A'_{k-1}) z_{n-k+1} \\ = & \sum_{k=1}^{n} (A_k - A'_k) z_{n-k+1} - \sum_{k=0}^{n-1} (A_k - A'_k) z_{n-k} \\ = & \sum_{k=1}^{n} (A_k - A'_k) z_{n-k+1} - \sum_{k=1}^{n} (A_k - A'_k) z_{n-k} \\ = & \sum_{k=1}^{n} (A_k - A'_k)(z_{n-k+1} - z_{n-k}) \geq 0, \end{aligned} \tag{8}$$

as, by hypothesis for all k, $A_k - A'_k \geq 0$, and $\{z_k\}$ is an increasing sequence.
□

The proof of Theorem 5 follows immediately from applying Lemma 1 to the inequality $\sum_k \alpha_k \log x'_k \leq \sum_k \alpha_k \log x_k$.

The following corollary follows from observing that, by taking $\{\alpha_k\} \equiv \{x_k\}$, then, $D_{\mathrm{KL}}(x_k, x_k) = 0 < D_{\mathrm{KL}}(x_k, x'_k)$, $\{x'_k\} \not\equiv \{x_k\}$.

Corollary 6. *An increasing distribution $\{x_k\}$ cannot be K-L dominated by any other distribution (except trivially by itself). In particular, the uniform distribution $\{x_k\} \equiv \{1/n\}$ does not dominate any other increasing distribution.*

Observe now that the product $1/n^n$ is always bigger than or equal $\Pi_k x_k$, as the maximum value for $\Pi_k x_k$ subjected to the condition $\sum_k x_k = 1$ is, for all k, $x_k = 1/n$. Thus, from the last inequality in Equation (4), we have the following corollary,

Corollary 7. *For all distributions* $\{x_k\}$, $\{x_k\} \not\equiv \{1/n\}$, *we have that* $\{x_k\} \not\succ_{\mathrm{KL}} \{1/n\}$.

Observe that Corollary 7 is also a particular case of Corollary 6, considering $\{1/n\}$ as an increasing distribution.

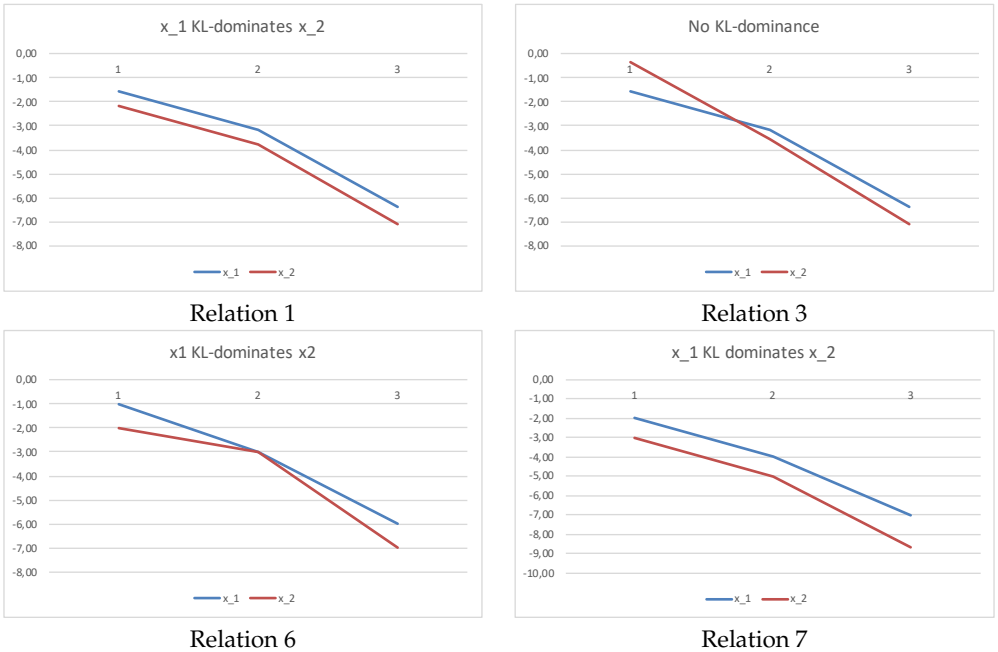
Corollary 8. *The uniform distribution* $\{1/n\}$ *K-L-dominates all decreasing distributions.*

Proof. For all $\{\alpha_k\}$ increasing and $\{x_k\}$ decreasing, we apply Corollary 4, and we have $CE(\alpha_k, 1/n) = \log n \leq CE(\alpha_k, x_k)$. □

Example 3. *Some examples are:*

1. $\{1/3, 1/3, 1/3\} \succ_{\mathrm{KL}} \{4/9, 3/9, 2/9\}$
2. $\{1/3, 1/3, 1/3\} \succ_{\mathrm{KL}} \{7/9, 1/9, 1/9\}$
3. $\{1/3, 1/3, 1/3\} \not\succ_{\mathrm{KL}} \{1/9, 1/9, 7/9\}$
4. $\{1/3, 1/3, 1/3\} \succ_{\mathrm{KL}} \{4/9, 2/9, 3/9\}$
5. $\{1/3, 1/3, 1/3\} \not\succ_{\mathrm{KL}} \{3/9, 2/9, 4/9\}$
6. $\{1/4, 1/4, 1/2\} \succ_{\mathrm{KL}} \{1/4, 1/2, 1/4\}$
7. $\{1/2, 1/4, 1/4\} \succ_{\mathrm{KL}} \{5/8, 2/8, 1/8\}$

The first and second relation are examples of Corollary 8. The third is an example of Corollary 6. The sixth is an example of Theorem 1. The fourth and fifth relations mean that the uniform distribution can both K-L dominate and non-dominate non-monotonous sequences, and we have to look for each case at the Condition (4) of Theorem 5. The seventh relation tells us that a decreasing sequence can dominate another decreasing sequence. In Figure 2, we plot the logarithm of products in Equation (4) for relations 1, 3, 6, and 7, respectively. If a sequence K-L-dominates another one, the plots do not cross, and its plot appears over that of the second sequence. Inversely, no dominance means that the plots will cross.



Relation 1 Relation 3

Relation 6 Relation 7

Figure 2. Plotting the logarithm of products in Equation (4) for relations 1, 3, 6, and 7 in Example 3.

4.1. Discussion

Remember that the equivalence between cross entropy and likelihood means that the likelihood that $\{x_k\}$ is the true distribution, given the data distribution $\{\alpha_k\}$, and the average length of code using $\{x_k\}$ to represent $\{\alpha_k\}$ are related by the second quantity being the negated logarithm of the first. If you have one distribution $\{\alpha_k\}$ and have the choice to use $\{x_{1,k}\}$ or $\{x_{2,k}\}$ to represent it, you want to choose the one with the minimum length of code, or, alternatively, the one with maximum likelihood. In both cases, the best is the one that gives minimum Kullback–Leibler divergence. Suppose now you want to represent data $\{\alpha_k\}$ and have two alternative distributions to represent it, either $\{x_{1,k}\} \equiv \{1/n\}$, the uniform distribution, or $\{x_{2,k}\}$, which is ordered in the same order as $\{\alpha_k\}$. To fix ideas and without loss of generality, let us consider that both $\{\alpha_k\}$ and $\{x_{2,k}\}$ are in increasing order. It would seem reasonable to represent the data $\{\alpha_k\}$, which is increasing, as distribution $\{x_{2,k}\}$, which is increasing too; thus, the order is preserved. However, that would work as long as $D_{\mathrm{KL}}(\alpha_k, x_{2,k}) < D_{\mathrm{KL}}(\alpha_k, 1/n)$; otherwise, we should choose $\{x_{1,k}\}$, the uniform distribution. This is clear from corollary 7, which tells us that, for any sequence, particularly $\{x_{2,k}\}$, there will always be increasing $\{\alpha_k\}$ sequences such that $D_{\mathrm{KL}}(\alpha_k, 1/n) < D_{\mathrm{KL}}(\alpha_k, x_{2,k})$. Intuitively, if $\{\alpha_k\}$ has a high increasing gradient, then we would choose $\{x_{2,k}\}$, the increasing distribution to represent it, but if $\{\alpha_k\}$ becomes smoother, then we would choose $\{x_{1,k}\}$, the uniform distribution.

As an example, suppose the two distributions $\{0.3, 0.7\}, \{0.4, 0.6\}$ have to be represented by one of the distributions $\{0.33, 0.67\}, \{0.5, 0.5\}$. To preserve the order, we would choose $\{0.33, 0.67\}$ for both. It is intuitively a clear choice of $\{0.3, 0.7\}$ (and a correct one, as it has a K-L divergence, half the one for $\{0.5, 0.5\}$), but we might have some doubts for $\{0.4, 0.6\}$, where, actually, the K-L divergence is only slightly smaller for $\{0.5, 0.5\}$. To represent distributions $\{0.43, 0.57\}, \{0.45, 0.55\}$, it is intuitively more clear to select the uniform distribution $\{0.5, 0.5\}$ rather than $\{0.33, 0.67\}$, even if order is lost. In fact, the K-L divergence from both distributions to $\{0.33, 0.67\}$ is one and two orders of magnitude greater, respectively, than to the uniform distribution $\{0.5, 0.5\}$.

4.2. Relationship between K-L Dominance and First Stochastic Dominance Orders

Summarizing K-L dominance and first stochastic dominance orders, we have

1. $\{\alpha_k\} \succ_{st} \{\alpha'_k\} \leftrightarrow \forall x_k$ increasing $CE(\alpha_k, x_k) \leq CE(\alpha'_k, x_k)$
2. $\{x_k\} \succ_{\mathrm{KL}} \{x'_k\} \leftrightarrow \forall \alpha_k$ increasing $CE(\alpha_k, x_k) \leq CE(\alpha_k, x'_k)$

Is there a relationship between the two orders? When $\Pi_k x_k = \Pi_k x'_k$ or, equivalently, $\sum_k \log x_k = \sum_k \log x'_k$ (observe that Theorem 1 is a particular case), we can find a necessary and sufficient condition for $\{x_k\} \succ_{KL} \{x'_k\}$.

Theorem 6. *Given distributions* $\{x_k\}$, $\{x'_k\}$, *and* $\Pi_k x_k = \Pi_k x'_k$, *then the following conditions are equivalent:*
(1)

$$\{x_k\} \succ_{\mathrm{KL}} \{x'_k\} \tag{9}$$

(2)

$$\{\frac{-\log x'_k}{-\sum_k \log x'_k}\} \succ_{st} \{\frac{-\log x_k}{-\sum_k \log x_k}\} \tag{10}$$

Proof. Observe first that $\{\frac{-\log x'_k}{-\sum_k \log x'_k}\}, \{\frac{-\log x_k}{-\sum_k \log x_k}\}$ are also distributions (we can indeed use the simpler expressions $\{\frac{\log x'_k}{\sum_k \log x'_k}\}, \{\frac{\log x_k}{\sum_k \log x_k}\}$, but, here, we have left the negated ones to remind the reader that both numerator and denominator are negative), because they are positive sequences adding to 1. Observe also that when $\Pi_k x_k = \Pi_k x'_k$, taking the logarithms in Equation (4) and changing the sign (which changes the direction of inequalities), we obtain condition (3) in Theorem 2, which is necessary and sufficient for $\{\frac{-\log x'_k}{-\sum_k \log x'_k}\} \succ_{st} \{\frac{-\log x_k}{-\sum_k \log x_k}\}$. □

Example 4. *Consider the distributions* $\{1/4, 1/4, 1/2\}, \{1/4, 1/2, 1/4\}$ *and their normalized negated logarithms* $\{2/5, 2/5, 1/5\}, \{2/5, 1/5, 2/5\}$, *respectively. We have that* $\{1/4, 1/4, 1/2\} \succ_{KL} \{1/4, 1/2, 1/4\}$ *and* $\{2/5, 1/5, 2/5\} \succ_{st} \{2/5, 2/5, 1/5\}$.

5. Conclusions and Future Work

In this paper, we present new inequalities for cross entropy and Kullback–Leibler divergence. As cross entropy is the negated logarithm of likelihood, the inequalities also hold for likelihood, changing the sense of the inequality. First, we applied to cross entropy the rearrangement inequality and recent stochastic order invariance results for Kolmogorov weighted means, as likelihood is a weighted geometric mean. Then, we introduced another partial order, K-L dominance, that applies directly to K-L divergences, and we give the relationship between both orders.

In data retrieval, exploration, analysis, and visualization, sorting some data objects into an ordered list or display is a common task performed by users. In the past, the benefit of having a sorted list has been typically articulated qualitatively. We plan to use the theorems presented in this paper to help establish an information-theoretic explanation about the cost–benefit of ordering in data retrieval, exploration, analysis, and visualization. We will also investigate the application to machine learning. We will study whether K-L-dominance order supports invariance properties, in the same way as first stochastic order supports the order invariance for weighted means, and we will look for sufficient conditions for K-L-dominance order, similar to the one in Theorem 3 for first stochastic order. Finally, we will investigate extensions of K-L dominance order: for instance, when $\{\alpha_k\} \equiv \{\frac{f(k)}{\sum_k f(k)}\}$, $f(x)$ increasing and concave/convex, in the same way as the extensions of first stochastic order.

Author Contributions: Conceptualization, M.S. and M.C.; validation, J.P. and A.B.; writing, original draft preparation, M.S.; review and editing, all authors.

Funding: Mateu Sbert acknowledges the funding of National Natural Science Foundation of China under grants No.61471261 and No.61771335, and by grant TIN2016-75866-C3-3-R from Spanish Government, Jordi Poch and Anton Bardera acknowledge the funding of TIN2016-75866-C3-3-R from Spanish Government.

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

References

1. Hu, B.G.; He, R.; Yuan, X.T. Information-Theoretic Measures for Objective Evaluation of Classifications. *Acta Autom. Sin.* **2012**, *38*, 1170–1182. [CrossRef]
2. Chen, M.; Golan, A. What may visualization processes optimize? *IEEE Trans. Vis. Comput. Graph.* **2016**, *22*, 2619–2632. [CrossRef] [PubMed]
3. Chen, M.; Feixas, M.; Viola, I.; Bardera, A.; Sbert, M.; Shen, H.W. *Information Theory Tools for Visualization*; CRC Press: Boca Raton, FL, USA, 2016.
4. Sbert, M.; Feixas, M.; Rigau, J.; Chover, M.; Viola, I. *Information Theory Tools for Computer Graphics*; Synthesis Lectures on Computer Graphics and Animation; Morgan & Claypool Publishers: San Rafael, CA, USA, 2009; Volume 4, pp. 1–153.
5. Galas, D.; Dewey, G.; Kunert-Graf, J.; Sakhanenko, N. Expansion of the Kullback-Leibler Divergence, and a New Class of Information Metrics. *Axioms* **2017**, *6*, *8*. [CrossRef]
6. Sbert, M.; Poch, J. A necessary and sufficient condition for the inequality of generalized weighted means. *J. Inequal. Appl.* **2016**, *2016*, 292. [CrossRef]
7. Shaked, M.; Shanthikumar, G. *Stochastic Orders*; Springer: New York, NY, USA, 2007.
8. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley–Interscience: New York, NY, USA, 1991.

9. Hardy, G.; Littlewood, J.; Pólya, G. *Inequalities*; Cambridge University Press: Cambridge, UK, 1952.
10. Belzunce, F.; Martinez-Riquelme, C.; Mulero, J. *An Introduction to Stochastic Orders*; Academic Press: Cambridge, MA, USA, 2016. doi:10.1016/B978-0-12-803768-3.09977-4.