entropy

MDPI

Article

Tight Bounds Between the Jensen–Shannon Divergence and the Minmax Divergence

Arseniy Akopyan [1], Herbert Edelsbrunner [2,*], Žiga Virk [3,4] and Hubert Wagner [5,*]

[1] Fora Capital, Miami, FL 33131, USA
[2] ISTA (Institute of Science and Technology Austria), 3400 Klosterneuburg, Austria
[3] Faculty of Computer and Information Science, University of Ljubljana, 1000 Ljubljana, Slovenia
[4] Institute of Mathematics, Physics and Mechanics (IMFM), 1000 Ljubljana, Slovenia
[5] Department of Mathematics, University of Florida, Gainesville, FL 32611, USA
* Correspondence: edels@ist.ac.at (H.E.); hwagner@ufl.edu (H.W.)

Abstract

Motivated by questions arising at the intersection of information theory and geometry, we compare two dissimilarity measures between finite categorical distributions. One is the well-known Jensen–Shannon divergence, which is easy to compute and whose square root is a proper metric. The other is what we call the minmax divergence, which is harder to compute. Just like the Jensen–Shannon divergence, it arises naturally from the Kullback–Leibler divergence. The main contribution of this paper is a proof showing that the minmax divergence can be tightly approximated by the Jensen–Shannon divergence. The bounds suggest that the square root of the minmax divergence is a metric, and we prove that this is indeed true in the one-dimensional case. The general case remains open. Finally, we consider analogous questions in the context of another Bregman divergence and the corresponding Burbea–Rao (Jensen–Bregman) divergence.

Keywords: information theory; relative entropy; Kullback–Leibler divergence; Jensen–Shannon divergence; minmax divergence; Bregman divergence; Burbea–Rao divergence; Jensen-Bregman divergence; metric; bounds

Academic Editor: Nikolai Leonenko

Received: 24 May 2025
Revised: 31 July 2025
Accepted: 4 August 2025
Published: 11 August 2025

Citation: Akopyan, A.; Edelsbrunner, H.; Virk, Ž.; Wagner, H. Tight Bounds Between the Jensen–Shannon Divergence and the Minmax Divergence. *Entropy* **2025**, *27*, 854. https://doi.org/10.3390/e27080854

1. Introduction

The starting point of this work is the introduction of the minmax divergence between two finite categorical distributions. It is a special case of a natural measurement arising in certain constructions from computational geometry and topology. Specifically, it coincides with the smallest radius at which two appropriately defined balls intersect. However, since our results promise to be useful also outside the field of geometry and topology, we will tailor the exposition accordingly.

The main result are tight bounds between the minmax divergence and the standard Jensen–Shannon divergence (as well as its generalizations). Given these bounds and the well-known fact the square root of the Jensen–Shannon divergence is a metric, we ask if the same can be proven about the minmax divergence. A supplementary result is a proof that in one dimension its square root is a metric. The general case is left as an open question.

The basic definition of the minmax divergence is based on the classic notion of Shannon entropy and has an intuitive information theoretic interpretation, as will be explained shortly. The definition coincides with an interpretation of the Chernoff information recently provided by Nielsen in [1]. In Section 2, we generalize this definition slightly, allowing us to work with the entire positive orthant of $\mathbb{R}^n$. Later in Section 5, we further generalize

using the language of Bregman divergences, which allows us to work with other notions of entropy, such as the Burg entropy. Before we proceed, we mention that the following exposition is tailored towards this generalization. In particular, we prefer to work with negative Shannon entropy. Being a strictly convex function it allows for a smoother transition to the case of Bregman divergences, which are defined via such functions.

Shannon entropy and related concepts. We start from basic definitions and review their information-theoretical interpretations. Consider a random variable that takes one of n values. Letting $x = (x_1, x_2, \ldots, x_n)$ be the vector of probabilities, we can encode the outcome with *expected efficiency* $-E(x)$, in which

$$E(x) = \sum_{i=1}^{n} x_i \ln x_i \tag{1}$$

is the (negative) *Shannon entropy* of x. We remark that if a binary logarithm was used, this quantity would be expressed in bits. We use natural logarithms to simplify calculations. Suppose we mistakenly assume that the underlying probability vector is $y = (y_1, y_2, \ldots, y_n)$, and we encode the random variable based on this faulty assumption. The expected efficiency is then $-E(y) - \langle \nabla E(y), x - y \rangle$. Comparing this with $-E(x)$, we get

$$D(x\|y) = E(x) - E(y) - \langle \nabla E(y), x - y \rangle \tag{2}$$

$$= \sum_{i=1}^{n} x_i \ln \tfrac{x_i}{y_i} \tag{3}$$

as a measure of the efficiency loss due to the faulty assumption. This quantity is often referred to as the *relative entropy*, which we note is not symmetric. Next assume we know that the random variable is either chosen according to the distribution x or according to the distribution y, each with 50% likelihood. Our best bet is to encode the result as if the underlying probability distribution were the average, $\mu = \frac{1}{2}(x + y)$. The expected efficiency is $-E(\mu)$, which we compare with $-\frac{1}{2}E(x) - \frac{1}{2}E(y)$, the expected efficiency assuming we know from which distribution the random variable is chosen. The difference,

$$\mathrm{JS}(x, y) = \tfrac{1}{2}E(x) + \tfrac{1}{2}E(y) - E(\mu) \tag{4}$$

$$= \tfrac{1}{2}\sum_{i=1}^{n} [x_i \ln \tfrac{2x_i}{x_i + y_i} + y_i \ln \tfrac{2y_i}{x_i + y_i}] \tag{5}$$

is the *Jensen–Shannon divergence*. This can be generalized to non-negative likelihoods $\xi + \eta = 1$, for which our best bet is to encode using $\xi x + \eta y$, with expected divergence $\xi E(x) + \eta E(y) - E(\xi x + \eta y)$. There are unique likelihoods, $\xi_0 + \eta_0 = 1$, for which the expected divergence is maximized. Because E is convex, this is also the situation in which the maximum divergence is minimized. We therefore consider this solution our best bet if we do not know how x and y split the likelihood. Setting $z = \xi_0 x + \eta_0 y$, the expected efficiency is $-E(z)$, which we compare with $-\xi_0 E(x) - \eta_0 E(y)$. The difference,

$$\mathrm{Mx}(x, y) = \xi_0 E(x) + \eta_0 E(y) - E(z) \tag{6}$$

$$= \sum_{i=1}^{n} [\xi_0 x_i \ln \tfrac{x_i}{z_i} + \eta_0 y_i \ln \tfrac{y_i}{z_i}], \tag{7}$$

minimizes the maximum divergence over all possible choices of likelihoods $\xi + \eta = 1$. We therefore call $\mathrm{Mx}(x, y)$ the *minmax divergence* between the two probability distributions. In the next section, we will provide an easier to work with alternative definition.

Main result. Our main result is a comparison of the minmax divergence with the standard Jensen–Shannon divergence. Specifically, we prove

$$\tfrac{e \ln 2}{2} \mathrm{Mx}(x, y) \le \mathrm{JS}(x, y) \le \mathrm{Mx}(x, y) \tag{8}$$

as tight bounds. We remark that the final result uses a generalized form of the above concepts, as defined in the next section.

Given that $\frac{e \ln 2}{2} \approx 0.94208$, the two divergences are very close. (As we will argue shortly, it makes sense to consider the square roots of these divergences, in which case the constant is approximately 0.9706, which is even closer to 1.) Additionally, the Jensen–Shannon divergence and the minmax divergence also yield the same intrinsic metric, which is $\frac{1}{4}$ times the intrinsic metric defined by the relative entropy. In the literature, the relative entropy is also known as the *Kullback–Leibler divergence* [2], and the corresponding intrinsic metric is known as the *Fisher information metric* [3]. Both are however not *length metrics*, which means that integrating infinitesimal steps gives a corresponding *intrinsic metric*, which is different from the original metric.

In light of these similarities, we ask if they share more properties. In particular Endres and Schindelin proved that $\sqrt{\mathrm{JS}(x,y)}$ is a metric [4]. We were able to prove a similar result in dimension one (specifically, in a slightly generalized setting in which $x, y \in \mathbb{R}_+$). Namely, we prove that $\sqrt{\mathrm{Mx}(x,y)}$ is a metric. The result in higher dimensions turns out to be challenging, and remains open. We do believe that the geometric proof techniques we introduce are worth sharing, and are likely to be useful in the high-dimensional case (in combination with some other techniques).

Applications in computational geometry. We briefly explain how the above concepts can be used in computational geometry and topology and, in particular, why the main result is useful. Our main motivation comes from the field of topological data analysis, a subfield of computational geometry and topology. In short, the idea is to characterize the *shape of data* or, in other words, its geometric-topological structure. We briefly describe a simple case to which our main result is immediately applicable. For a comprehensive treatment of topological data analysis in the context of arbitrary Bregman divergences, which includes our setup with relative entropy, see [5].

In the simplest case—which is also most relevant here—the input data is a finite collection of points. We consider the union of the balls centered at these points and increase their common radius from zero to infinity. As the radius changes, so does the connectivity (or topology) of the union of balls. One basic topological property is the structure of the connected components. At radius zero, each point constitutes its own connected component, and these components may merge as the radius increases. Specifically, two components may merge when two balls develop a nonempty intersection for the first time. (We remark that higher-degree topological features can also be considered, but this requires tools from algebraic topology, which are beyond the scope of this paper.).

This setup can be easily implemented in the Euclidean space. However, in the more interesting situation in which each point is a finite categorical distribution, the balls are better defined using the relative entropy (and not the Euclidean distance), as it allows the outcome to have an information-theoretic interpretation, as outlined above. The radius at which two balls intersect coincides with the minmax divergence [6], whose direct computation (especially for many pairs of points) can be slow. The proposed inequality suggests we compute the approximating Jensen–Shannon divergence instead. It remains fairly accurate while being significantly faster, simpler, and more robust. In particular, this allows for a simple computation of a weighted undirected graph that describes the connected structure of data measured with relative entropy. In topological data analysis language, it would be called the 1-skeleton of the Čech complex (the full complex would be a weighted simplicial complex that encodes intersections between arbitrary tuples of balls and captures higher-degree topological information.).

In summary, our results simplify a fundamental computation in topological data analysis for an important kind of inputs. This is often the first step towards computing a

topological descriptor, which has been proven useful in a variety of fields [7] ranging from biology, to astronomy, to materials science. Moreover, if the square root of the minmax divergence is indeed a metric (a fact we were unable to prove beyond dimension one), it would allow one to use existing theoretic in computational tools. Here we mention stability results in topological data analysis that exploit the metric structure of data [8], and classical nearest neighbour search tools that focus on metric spaces [9].

Outline. Section 2 introduces the main concepts in their generalized forms and proves some of their fundamental properties. Section 3 shows that the Jensen–Shannon divergence approximates the minmax divergence within a factor 1.061... Section 4 proves that in one dimension the square root of the minmax divergence is a metric. Section 5 extends the results beyond the Shannon entropy using the framework of Bregman divergences. Section 6 concludes the paper.

Related work. Many concepts used in this paper lead back to the seminal work of Claude Shannon [10] on information theory and, in particular, the notion of Shannon entropy. This notion was extended to a dissimilarity measure between two probability distributions by Kullback and Leibler [2], often referred to as the relative (Shannon) entropy or the Kullback–Leibler divergence. The best known metric derived from relative Shannon entropy is based on the Jensen–Shannon divergence, defined by Lin in 1991 [11]. Interestingly, its more general form was introduced a decade earlier by Burbea and Rao [12]. Lew Bregman introduced a notion of Bregman divergence [13], which generalizes the relative entropy. Various other metrics derived from Bregman divergences were studied in [14,15]. More recently, Bregman divergences were further generalized by Nielsen in various ways [16–18]. Our work is motivated by results at the intersection of Bregman geometry and computational geometry. The starting point for this direction is the work of Boissonnat, Nielsen and Nock [19–21] on computational geometry in the Bregman context.

2. Generalized Definitions

In the Introduction, we described basic concepts (such as the relative entropy) along with their information-theoretical interpretation. In this section, we introduce more general versions of these concepts, allowing us to work in the entire positive orthant $\mathbb{R}^n_+$. We remark that the definitions change in subtle ways, and that applying the usual definitions in this extended setup may on occasion be non-sensical.

For the reminder of this paper, we are exclusively concerned with two spaces: the n-dimensional *positive orthant*, denoted $\mathbb{R}^n_+$, which consists of all points $x = (x_1, x_2, \ldots, x_n)$ with $x_i > 0$ for $1 \leq i \leq n$, and the open $(n-1)$-dimensional *standard simplex*, denoted $\Delta = \Delta^{n-1}$, which consists of the points $x \in \mathbb{R}^n_+$ that satisfy $\sum_{i=1}^n x_i = 1$. A point $x \in \Delta$ is really a finite probability distribution, which leads us to believe that Δ is the more important setting. However, $\mathbb{R}^n_+$ is often easier to work with, and we can restrict results to $\Delta \subseteq \mathbb{R}^n_+$.

Shannon entropy and relative entropy. Writing $\ln t$ for the natural logarithm of $t > 0$, the (negative) *Shannon entropy* is $E\colon \mathbb{R}^n_+ \to \mathbb{R}$ defined by $E(x) = \sum_{i=1}^n [x_i \ln x_i - x_i]$. (Subtracting the extra term is a standard trick to simplify computations while not affecting the interpretation of the resulting relative entropy. However, the interpretation of the Shannon entropy mentioned in the Introduction holds only up to a constant.) We write $E|_\Delta\colon \Delta \to \mathbb{R}$ for its restriction to the standard simplex.

As mentioned in Section 1, $-E(x)$ pertains to the *expected efficiency* of encoding a random variable that distributes according to $x \in \Delta$. If we encode the values assuming the random variable distributes according to $y \in \Delta$, the expected efficiency is the negative of the best linear approximation of E at y evaluated at x, which is $-E(y) - \langle \nabla E(y), x - y \rangle$. The relative entropy can be viewed as the non-negative difference between the above approximation and the Shannon entropy of x; see Figure 1.

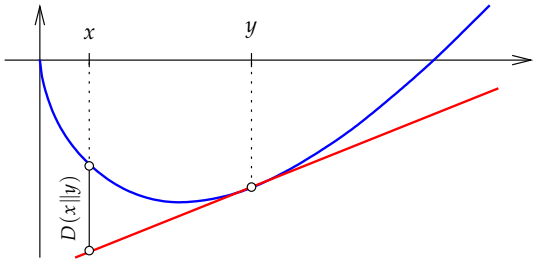

Figure 1. The graph of the Shannon entropy on $\mathbb{R}_+$, the graph of its best linear approximation at y, and the relative entropy from x to y.

Going back to $x, y \in \mathbb{R}^n_+$, we consider a generalized form of the *relative entropy* from x to y:

$$\begin{aligned} D(x\|y) &= E(x) - E(y) - \langle \nabla E(y), x - y \rangle && (9)\\ &= \textstyle\sum_{i=1}^n [x_i \ln \frac{x_i}{y_i} - x_i + y_i] && (10)\\ &= \textstyle\sum_{i=1}^n D(x_i\|y_i). && (11) \end{aligned}$$

Note the extra additive terms, which are absent in the usual form and ensure that the result is nonnegative. When restricted to the standard simplex, it simplifies to its more standard form, namely $\sum_{i=1}^n x_i \ln \frac{x_i}{y_i}$.

We say the relative entropy is *decomposable* because it satisfies (11). Observe that the notation separates the points x and y by a double bar as opposed to a comma to remind us that the measure is generally not symmetric. The relative entropy is also known as the *Kullback divergence*, the *Kullback–Leibler divergence*, or simply the *divergence*; see [3] (page 57). It measures the divergence in encoding efficiency due to assuming a different distribution.

Jensen–Shannon divergence. Similar to the relative entropy, the *Jensen–Shannon divergence* generalized to the positive orthant is a function $\mathrm{JS}\colon \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}$, but it is symmetric by taking the average relative entropy from x to $\mu = \frac{1}{2}(x + y)$ and from y to μ; see Figure 2:

$$\begin{aligned} \mathrm{JS}(x, y) &= \tfrac{1}{2}[D(x\|\mu) + D(y\|\mu)] && (12)\\ &= \tfrac{1}{2}E(x) + \tfrac{1}{2}E(y) - E(\mu) && (13)\\ &= \tfrac{1}{2}\textstyle\sum_{i=1}^n [x_i \ln \frac{2x_i}{x_i+y_i} + y_i \ln \frac{2y_i}{x_i+y_i}]. && (14)\\ &= \textstyle\sum_{i=1}^n \mathrm{JS}(x_i, y_i), && (15) \end{aligned}$$

in which we get (13) by noting $(x - \mu) + (y - \mu) = 0$. Similar to the relative entropy, the Jensen–Shannon divergence is decomposable (15). As pointed out in [4], $\mathrm{JS}(x, y)$ measures the divergence of expected efficiency when we encode a random variable that distributes half of the time according to $x \in \Delta$ and the other half of the time according to $y \in \Delta$ using the average of x and y.

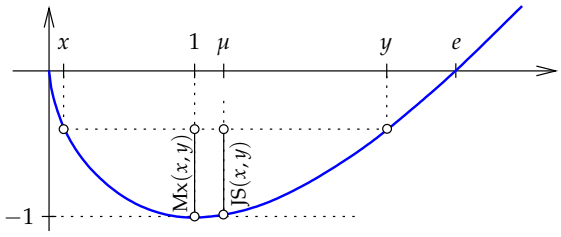

Figure 2. A pair $x, y \in \mathbb{R}_+$ that satisfies $E(x) = E(y)$, the midpoint, $\mu = (x + y)/2$, and the minmax divergence center $z = 1$. The Jensen–Shannon divergence and the minmax divergence of x and y are labeled.

It is not difficult to prove that if we substitute any other point for μ, then the average relative entropy and therefore the divergence in efficiency increases. For reasons that will become obvious shortly, we prove this optimality result for a weighted version of this divergence. Let $\xi + \eta = 1$ and set $w = \xi x + \eta y$. The corresponding *weighted Jensen–Shannon divergence* is

$$\begin{aligned} \mathrm{JS}_{\xi}(x\|y) &= \xi D(x\|w) + \eta D(y\|w) && (16)\\ &= \xi E(x) + \eta E(y) - E(w) && (17)\\ &= \textstyle\sum_{i=1}^{n} [\xi x_i \ln \frac{x_i}{w_i} + \eta y_i \ln \frac{y_i}{w_i}] && (18)\\ &= \textstyle\sum_{i=1}^{n} \mathrm{JS}_{\xi}(x_i\|y_i). && (19) \end{aligned}$$

To prove optimality, we set $f_{\xi}(u) = \xi D(x\|u) + \eta D(y\|u)$, noting that $f_{\xi}(w) = \mathrm{JS}_{\xi}(x\|y)$. The following lemma and proof can also be found in [22].

Lemma 1 (Optimality of Weighted JS)**.** *Let $x, y \in \mathbb{R}^n_+$, $\xi + \eta = 1$, and $w = \xi x + \eta y$. Then $f_{\xi}(w) \leq f_{\xi}(u)$ for every $u \in \mathbb{R}^n_+$, with equality iff $u = w$.*

Proof. Computing the difference, $X = f_{\xi}(u) - f_{\xi}(w)$, most terms cancel and we get

$$\begin{aligned} X &= \xi[D(x\|u) - D(x\|w)] + \eta[D(y\|u) - D(y\|w)] && (20)\\ &= E(w) - E(u) - \langle \nabla E(u), w - u \rangle, && (21) \end{aligned}$$

in which we use $\xi(x - w) + \eta(y - w) = 0$. In other words, the difference is equal to $D(w\|u)$, which is non-negative and zero iff $u = w$ by the strict convexity of E. □

Remark 1. *To get an information theoretic interpretation of the result, we assume $x, y \in \Delta$ and suppose a random variable that distributes according to x with likelihood $0 \leq \xi \leq 1$ and according to y with likelihood $\eta = 1 - \xi$. Lemma 1 says that our best bet is to encode with $w = \xi x + \eta y$. In words, w minimizes the weighted Jensen–Shannon divergence.*

Minmax divergence. Similar to the Jensen–Shannon divergence, the *minmax divergence* is a symmetric function $\mathrm{Mx}\colon \mathbb{R}^n_+ \times \mathbb{R}^n_+ \to \mathbb{R}$. It is defined by mapping x, y to the larger relative entropy to a third point, $z \in \mathbb{R}^n_+$, in which z is selected so as to minimize this maximum:

$$\mathrm{Mx}(x, y) = \inf_{z \in \mathbb{R}^n_+} \max\{D(x\|z), D(y\|z)\}. \quad (22)$$

We call the point z that gives the infimum the *minmax divergence center* of the pair. We remark that the general form of minmax divergence introduced in Section 5 generalizes the Chernoff information [1] popular in statistics.

As proved for example in [6], z is a convex combination of x and y. We strengthen this result by proving that it is the particular convex combination that maximizes the weighted Jensen–Shannon divergence, and that this weighted Jensen–Shannon divergence equals the minmax divergence.

Lemma 2 (Minmax-Maxweight)**.** *Let $x, y \in \mathbb{R}^n_+$ and $\xi_0 + \eta_0 = 1$ such that $z_0 = \xi_0 x + \eta_0 y$ is the minmax divergence center. Then $\mathrm{Mx}(x, y) = \mathrm{JS}_{\xi_0}(x\|y) \geq \mathrm{JS}_{\xi}(x\|y)$ for all ξ.*

Proof. Let $\xi + \eta = 1$ and write $z = \xi x + \eta y$ for a general affine combination of x and y. The restriction of E to the line of such points is a strictly convex function. It follows that there is a unique affine combination, $z_0 = \xi_0 x + \eta_0 y$, such that $D(x\|z_0) = D(y\|z_0)$. The

weighted Jensen–Shannon divergence at z_0 is $\mathrm{JS}_{\xi_0}(x\|y) = \xi_0 D(x\|z_0) + \eta_0 D(y\|z_0)$, which is equal to $D(x\|z_0) = D(y\|z_0)$, as claimed.

By convexity of the restriction of E, the maximum relative entropy is larger than this shared value for every affine combination $z \neq z_0$. Similarly, the weighted Jensen–Shannon divergence is smaller than $\mathrm{JS}_{\xi_0}(x\|y)$ at every affine combination $z \neq z_0$. □

Remark 2. *In contrast to the other measures discussed so far, the minmax divergence is not decomposable.*

Remark 3. *Since the minmax divergence center of x and y lies between these two points, it is constrained to a compact set so we can replace the infimum in (22) by a minimum. This justifies the name of corresponding divergence. Lemma 2 says that the minimum of the maximum relative entropy is equal to the maximum weighted Jensen–Shannon divergence, which justifies the name of the lemma.*

Explicit formula. While the minmax divergence is defined as an infimum, it is possible to compute it explicitly. To state the formula, we introduce $G\colon \mathbb{R}_+ \to \mathbb{R}$ defined by

$$G(t) = \frac{t^{\frac{t}{t-1}}}{e} - \frac{t \ln t}{t-1}. \tag{23}$$

It is well defined at all positive $t \neq 1$, and we get $G(1) = 0$ in the limit because $\lim_{t\to 1} t^{\frac{t}{t-1}} = e$ and $\lim_{t\to 1} \frac{t\ln t}{t-1} = 1$ by the rule de l'Hôpital.

Lemma 3 (Minmax divergence Formula)**.** *For $x, y \in \mathbb{R}_+$, we have* $\mathrm{Mx}(x,y) = xG(\frac{y}{x})$.

Proof. For $x = y$ both sides vanish, so the relation holds. We therefore assume $x \neq y$ for the remainder of the proof. Letting z be the minmax divergence center of x and y, we recall that the derivative at z is the slope of the line that passes through $(x, E(x))$ and $(y, E(y))$: $E'(z) = \ln z = [E(y) - E(x)]/[y - x]$. We express this equation in terms of $\frac{y}{x}$:

$$\ln z = \frac{y \ln y - x \ln x - (y - x)}{y - x} \tag{24}$$

$$= \frac{\frac{y}{x}(\ln \frac{y}{x} + \ln x) - \ln x}{\frac{y}{x} - 1} - 1 \tag{25}$$

$$= \frac{\frac{y}{x} \ln \frac{y}{x}}{\frac{y}{x} - 1} + \ln x - 1. \tag{26}$$

Write $A = (\frac{y}{x} \ln \frac{y}{x})/(\frac{y}{x} - 1)$ for the first term on the right-hand side of (26). Recall that the minmax divergence of x and y is the vertical distance between the point $(z, E(z))$ and the line that passes through $(x, E(x))$ and $(y, E(y))$. By construction, the slope of this line is $\ln z$. We express the vertical distance as the sum of two vertical distances, which we then express in terms of A and $\frac{y}{x}$:

$$\mathrm{Mx}(x,y) = [(z - x)\ln z] + [E(x) - E(z)] \tag{27}$$

$$= z - x + x \ln x - x \ln z \tag{28}$$

$$= z - x + x \ln x - x[A + \ln x - 1] \tag{29}$$

$$= x[e^{A-1} - A] \tag{30}$$

$$= xG\left(\tfrac{y}{x}\right), \tag{31}$$

in which we use (26) to get (29), we cancel $-x + x \ln x$ and replace z to get (30), and finally use $\frac{y}{x} = e^{\ln \frac{y}{x}}$ to get (31). □

Bregman divergences. The relative entropy can be viewed from the perspective of Bregman divergences [13]. Indeed, it is an instance of a Bregman divergence generated by the negative Shannon entropy. We briefly introduce the setup for general Bregman divergences, generated by arbitrary convex functions, or more technically functions of Legendre type.

Given an open convex set $\Omega \subseteq \mathbb{R}^d$, a function $F : \Omega \to \mathbb{R}$ is of Legendre type if it is (1) differentiable, (2) strictly convex and (3) the magnitude of its gradient diverges to positive infinity when evaluated at points converging to the boundary of the domain. Given a function F of Legendre type, the Bregman divergence [13] generated by F is defined as

$$D_F : \Omega \times \Omega \to \mathbb{R}, \qquad D_F(x\|y) = F(x) - (F(y) + \langle \nabla F(y), x - y \rangle).$$

We will use this concept to generalize our main result in Section 5.

3. Comparison of Divergences

We think of the Jensen–Shannon divergence as a readily computed approximation of the minmax divergence. The approximation is very close, and we prove in this section that the Jensen–Shannon divergence is always between $\frac{e \ln 2}{2} = 0.942\ldots$ and 1 times the minmax divergence. We prove this first in one dimension and then generalize the result to n dimensions.

Approximation with ellipse. The main tool in proving the relation between the minmax divergence and the Jensen–Shannon divergence is the—surprisingly close—approximation of the graph of the Shannon entropy defined over $\mathbb{R}_+$ with an arc of an ellipse. We are interested in the interval $[0, e]$, so we choose the ellipse to

- pass through the points $(0, 0)$ and $(e, 0)$;
- have its minimum at the point $(1, -1)$;
- have the same curvature at $(1, -1)$ as the graph of the Shannon entropy.

Writing the ellipse as the zero-set of a function, we introduce $\Gamma \colon \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\begin{aligned} \Gamma(x_1, x_2) = {} & x_1^2 + (2 - e)x_1 x_2 + (3 - e)x_2^2 \\ & - e x_1 + (2 - e)x_2. \end{aligned} \tag{32}$$

It is not difficult to check that $\Gamma^{-1}(0)$ satisfies the above three properties. We also note that $\Gamma(x_1, x_2)$ is negative for points inside the ellipse and positive for points outside the ellipse. Within $[0, e]$, the approximation of E by the lower portion of the ellipse is astonishingly close, and we exaggerate the difference in Figure 3 to make it visible. We prove that from 0 to 1 the graph of E is below the ellipse, and from 1 to e it is above the ellipse.

Lemma 4 (Below-Above)**.** *Letting $E \colon \mathbb{R}_+ \to \mathbb{R}$ be the 1-dimensional Shannon entropy and $\Gamma \colon \mathbb{R}^2 \to \mathbb{R}$ the function defined in (32). Then*

$$\Gamma(x, E(x)) \begin{cases} > 0 & \text{for } 0 < x < 1, \\ < 0 & \text{for } 1 < x < e. \end{cases} \tag{33}$$

Proof. Write $\Gamma(x, E(x)) = x f(x)$, in which

$$\begin{aligned} f(x) = {} & (2x - 2) + (2 - e + ex - 4x) \ln x \\ & + (3 - e)x \ln^2 x. \end{aligned} \tag{34}$$

We have $f(1) = f(e) = 0$ by construction of Γ, but not necessarily $f(0) = 0$ because we divided by x. It suffices to prove that $f(x)$ is positive for $0 < x < 1$ and negative for $1 < x < e$. Next we compute the first two derivatives, again after removing a monotonic factor to simplify the computations. Specifically, we write $f'(x) = \frac{1}{x}g(x)$, in which

$$g(x) = (e-2)(x-1-x\ln x) + (3-e)x\ln^2 x, \tag{35}$$

$$g'(x) = (3-e)\ln^2 x - (3e-8)\ln x. \tag{36}$$

The derivative of g is quadratic in $\ln x$, with zeros at $x = 1$ and $x = u_0 = \exp(\frac{3e-8}{3-e}) = 1.732\ldots$ Hence, $g'(x)$ is negative for $1 < x < u_0$ and positive outside the corresponding closed interval. Returning to g, we note that g is zero, negative, positive at $1 < u_0 < e$:

$$g(1) = 0,\ g(u_0) = -0.010\ldots,\ g(e) = 0.047\ldots \tag{37}$$

Our analysis of g' implies that g increases from 0 to 1, it decreases from 1 to u_0, and finally it increases again from u_0 to e, with a zero at some value u_1 between u_0 and e. Hence, f decreases from 0 to u_1, with $1 < u_1 < e$, and it increases from u_1 to e. Since $f(1) = f(e) = 0$, this implies that f is positive from 0 to 1 and negative from 1 to e. The claimed inequalities for Γ follow. □

Midpoint lines. We are interested in the relative position of two midpoint lines. The first is defined by the Shannon entropy, E, and the second by the function $G\colon [0,e] \to \mathbb{R}$ whose graph is the portion of the ellipse on and below the horizontal coordinate axis. Both lines consist of points $(\frac{1}{2}(x+y), t)$ with $0 \le x \le 1 \le y \le e$, in which the points of the first line satisfy $E(x) = E(y) = t$ and the points of the second line satisfy $G(x) = G(y) = t$.

The ellipse can be obtained by shearing a circle, and since this operation takes straight lines to straight lines, it follows that the midpoint line of G is a straight line segment. Its endpoints are $(1,-1)$ and $(\frac{e}{2},0)$. By Lemma 4, the midpoint line of E is a curve that connects the same two endpoints but lies otherwise to the left of the line segment; see Figure 3.

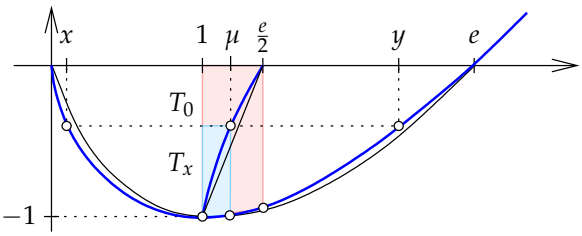

Figure 3. The (blue) midpoint line of E lies to the left of the (black) midpoint line of G. The two shaded trapezoids visualize the ratios between the minmax divergence and the Jensen–Shannon divergence for the pairs $0 < e$ and $0 < x < y < e$ with $E(x) = E(y)$. Since the upper right corner of the smaller trapezoid lies on the midpoint line of E, the second ratio is smaller than the first ratio.

Inequalities. Recall the definition of the weighted Jensen–Shannon divergence for a real parameter and points:

$$\mathrm{JS}_\xi(x\|y) = \xi E(x) + \eta E(y) - E(z), \tag{38}$$

in which $\eta = 1-\xi$ and $z = \xi x + \eta y$. The (unweighted) Jensen–Shannon divergence is $\mathrm{JS}(x,y) = \mathrm{JS}_{1/2}(x\|y)$, and from Lemma 2 we know that the minmax divergence can be written as $\mathrm{Mx}(x,y) = \max_\xi \mathrm{JS}_\xi(x\|y)$. We prove that the two measures of information divergence are good approximations of each other.

Theorem 1 (Loss Comparison). *Letting* $x, y \in \mathbb{R}^n_+$, *we have* $\frac{e \ln 2}{2} \mathrm{Mx}(x,y) \leq \mathrm{JS}(x,y) \leq \mathrm{Mx}(x,y)$.

Proof. The upper bound on $\mathrm{JS}(x,y)$ follows from Lemma 2, so we focus on proving the lower bound. We begin with the 1-dimensional case, when $n = 1$. Recall that $D(Cx\|Cy) = C\,D(x\|y)$ for every $C \geq 0$. Hence,

$$\mathrm{JS}(Cx, Cy) = C\,\mathrm{JS}(x,y), \tag{39}$$

$$\mathrm{Mx}(Cx, Cy) = C\,\mathrm{Mx}(x,y), \tag{40}$$

which implies that the ratio is independent of C. Given $x < y$, we can find $C > 0$ such that $E(Cx) = E(Cy)$, so we assume $E(x) = E(y)$ for the remainder of the 1-dimensional argument. This implies that the minmax divergence center is $z = 1$. Setting $x = 0$ and $y = e$, we have $E(0) = E(e) = 0$, and the ratio is

$$\frac{\mathrm{Mx}(0,e)}{\mathrm{JS}(0,e)} = \frac{-1}{E(\frac{e}{2})} = \frac{-1}{\frac{e}{2}\ln\frac{e}{2} - \frac{e}{2}} = \frac{2}{e \ln 2}. \tag{41}$$

Observe that this ratio is the length of the left edge divided by the length of the right edge of the trapezoid T_0 in Figure 3. For a general pair $0 < x < y$ with $E(x) = E(y)$, we represent the ratio by the left and right edges of a similar trapezoid, T_x. Importantly, the two trapezoids share $(1, -1)$ as their common lower left corner, and the bottom edge of T_x has smaller positive slope than the bottom edge of T_0. The height of T_0 is 1 and that of T_x is $t < 1$. To compare the two ratios, we thus consider $T_x/t + (1-t, t-1)$, which is the scaled version of T_x whose lower left corner is $(1, -1)$ and whose height is 1. The upper right corner of T_x lies on the midpoint line of E, which by Lemma 4 implies that the upper right corner of the scaled trapezoid lies to the left of $\frac{e}{2}$ on the horizontal coordinate axis. It follows that the width of the scaled trapezoid is smaller than the width of T_0. Hence,

$$\frac{\mathrm{Mx}(x,y)}{\mathrm{JS}(x,y)} < \frac{\mathrm{Mx}(0,e)}{\mathrm{JS}(0,e)} = \frac{2}{e \ln 2}. \tag{42}$$

We get equality for $x = 0$ and for $x = 1$, which implies the claimed lower bound for $n = 1$ dimension. Moving on to $n \geq 1$ dimensions, we recall that the minmax divergence center is a convex combination of the two points [6]. Specifically, the center satisfies $z = \xi x + \eta y$ with $\xi, \eta \geq 0$ and $\xi + \eta = 1$. Using Lemma 2, the decomposability of the weighted Jensen–Shannon divergence (19), and the claimed inequality in $n = 1$ dimension—in this sequence—we get

$$\mathrm{Mx}(x,y) = \mathrm{JS}_\xi(x\|y) \tag{43}$$

$$= \textstyle\sum_{i=1}^n \mathrm{JS}_\xi(x_i\|y_i) \tag{44}$$

$$\leq \frac{2}{e \ln 2} \textstyle\sum_{i=1}^n \mathrm{JS}(x_i, y_i) \tag{45}$$

$$= \frac{2}{e \ln 2} \mathrm{JS}(x,y), \tag{46}$$

as claimed. □

Remark 4. *The bounds in Theorem 1 are tight. To see this for the lower bound, we note that* $\mathrm{Mx}(0,e) = 1$ *and* $\mathrm{JS}(0,e) = -E(\frac{e}{2}) = \frac{e \ln 2}{2}$. *While* 0 *is formally not part of the domain, we can take points arbitrarily close to* 0 *and thus get the bound in the limit. To see that the upper bound is tight, we let* $\varepsilon > 0$ *be small and set* $x = 1 - \varepsilon$ *and* $y = 1 + \varepsilon$. *We can see the two entropies geometrically as the vertical distance of two points on the graph of* E *below the line that passes*

through $(x, E(x))$ *and* $(y, E(y))$*; see Figure 2. For* $\mathrm{JS}(x, y)$ *this point is* $(\mu, E(\mu)) = (1, -1)$*, and for* $\mathrm{Mx}(x, y)$ *this point is* $(z, E(z))$*, in which* z *is the minmax divergence center of* x *and* y*. To determine* z*, we note that* $D(x\|z) = D(y\|z)$*. After some computations, including the Taylor expansions of* $\ln s$ *around* $s = 1$ *and of* e^s *around* $s = 0$*, we find that* z *is* $1 - \frac{1}{3}\varepsilon^2$ *plus a fourth-order term in* ε*. In words,* z *approaches* $\mu = 1$ *much faster than* a *and* b*. It follows that in the limit, the two entropies are the same, as required.*

4. Proof of Metric in Dimension One

As proved in [4], the square root of the Jensen–Shannon divergence is a metric in $\mathbb{R}_+$. Using the decomposability of the Jensen–Shannon divergence together with the Minkowski inequality, it is then easy to prove that this square root is also a metric in $\mathbb{R}_+^n$. We prove that the square root of the minmax divergence is a metric in $\mathbb{R}_+$. Since the minmax divergence is not decomposable, we do not know yet whether its square root is a metric in $\mathbb{R}_+^n$.

The ratio method. Suppose $A\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ satisfies the triangle inequality and $B\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is another function on the product. To prove that B also satisfies the triangle inequality, we may consider the ratio, $f(x, y) = B(x, y)/A(x, y)$ and prove its *monotonicity*, that is:

$$f(a, b) \le f(x, y) \tag{47}$$

whenever $a \le x \le y \le b$.

Lemma 5 (Triangle Inequality)**.** *Let* $A, B, f\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *in which* A *satisfies the triangle inequality and* $f = B/A$ *is monotonic. Then* B *satisfies the triangle inequality.*

Proof. Let $x \le y \le z$. Then

$$\begin{aligned} B(x, y) + B(y, z) &= f(x, y)A(x, y) + f(y, z)A(y, z) && (48)\\ &\ge f(x, z)A(x, z) && (49)\\ &= B(x, z), && (50) \end{aligned}$$

in which we get (49) using the monotonicity of f and the triangle inequality for A. □

To apply the lemma, we set $B(x, y) = \sqrt{\mathrm{Mx}(x, y)}$ and $A(x, y) = |\sqrt{y} - \sqrt{x}|$. Clearly, A is a metric in $\mathbb{R}_+$, so it will suffice to show that the ratio is monotonic.

A first application. As a warm up exercise, we use the ratio method expressed in Lemma 5 to re-prove the main result of [4].

Theorem 2 (JS Revisited)**.** *Let* $E\colon \mathbb{R}_+^n \to \mathbb{R}$ *be the Shannon entropy and* $\mathrm{JS}\colon \mathbb{R}_+^n \times \mathbb{R}_+^n \to \mathbb{R}$ *map every pair to the Jensen–Shannon divergence. Then* $\sqrt{\mathrm{JS}}$ *is a metric in* $\mathbb{R}_+^n$*.*

Proof. We begin with the one-dimensional case, $n = 1$. It is clear that $\mathrm{JS}(x, y)$ is non-negative, zero iff $x = y$, and symmetric. It remains to prove that its square root satisfies the triangle inequality. Setting

$$f(x, y) = \frac{\sqrt{\mathrm{JS}(x, y)}}{|\sqrt{y} - \sqrt{x}|}, \tag{51}$$

we will prove that f is monotonic as defined in (47). Recall that $\mathrm{JS}(Cx, Cy) = C\,\mathrm{JS}(x, y)$ for every $C > 0$. It follows that $f(Cx, Cy) = f(x, y)$, which allows us to assume $x = 1$. To simplify the notation, we set $t^2 = y$ assuming $t^2 \ge 1$. Writing

$g(t) = \mathrm{JS}(1, t^2)/(t-1)^2 = f(1,y)^2$, we aim at proving that g is monotonically decreasing, that is: $g'(t) < 0$ for all $t \geq 1$. Recall from (5) that

$$\mathrm{JS}(1, t^2) = \tfrac{1}{2}\left[\ln \tfrac{2}{1+t^2} + t^2 \ln \tfrac{2t^2}{1+t^2}\right], \tag{52}$$

$$\frac{\partial \mathrm{JS}(1, t^2)}{\partial t} = t \ln \tfrac{2t^2}{t^2+1}. \tag{53}$$

The derivative of g is $g'(t) = N(t)/(t-1)^3$, in which

$$N(t) = (t-1)\frac{\partial \mathrm{JS}(1, t^2)}{\partial t} - 2\mathrm{JS}(1, t^2) \tag{54}$$

$$= -t \ln t^2 + (t^2+1) \ln \tfrac{t^2+1}{2}. \tag{55}$$

For $t = 1$, both the numerator and the denominator vanish: $N(1) = 0$ and $D(1) = 0$ in which $D(t) = (t-1)^3$. Applying the rule de l'Hôpital three times, we get

$$N'(t) = \ln \tfrac{t^2+1}{2} - \ln t^2 + \tfrac{2t-2}{t^2+1}, \tag{56}$$

$$N''(t) = -\tfrac{2(t+1)(t-1)^2}{(t^2+1)^2 t}, \tag{57}$$

$$N'''(t) = \tfrac{4t^7-6t^6-8t^5+6t^4-12t^3+14t^2+1}{(t^5+2t^3+t)^2}, \tag{58}$$

with $N'(1) = N''(1) = N''(1) = 0$. However, $D'(1) = D''(1) = 0$ and $D'''(1) = 6$. Hence, $g'(1) = 0$ and it suffices to prove $g''(t) < 0$ for $t > 1$. Since the denominator of g' is positive, this is equivalent to $N'(t) < 0$ for $t > 1$. But this follows from $N'(1) = 0$ and $N''(t) < 0$ for $t > 1$, which can be seen from (57). Hence f is monotonic and since the denominator in (51) satisfies the triangle inequality, Lemma 5 implies that $\sqrt{\mathrm{JS}}$ also satisfies the triangle inequality and therefore is a metric.

Having established the claim in one dimension, we get the n-dimensional result using the Minkowski inequality, which for non-negative real numbers a_i and b_i implies

$$\sqrt{\textstyle\sum_{i=1}^n (a_i + b_i)^2} \leq \sqrt{\textstyle\sum_{i=1}^n a_i^2} + \sqrt{\textstyle\sum_{i=1}^n b_i^2}. \tag{59}$$

Recall that the Jensen–Shannon divergence is *decomposable*: $\mathrm{JS}(x,y) = \sum_{i=1}^n \mathrm{JS}(x_i, y_i)$ for $x, y \in \mathbb{R}_+^n$. Letting $z \in \mathbb{R}_+^n$ be a third point, we set $a_i^2 = \mathrm{JS}(x_i, y_i)$ and $b_i^2 = \mathrm{JS}(y_i, z_i)$ for all i. Since $\sqrt{\mathrm{JS}}$ satisfies the triangle inequality in one dimension, we have $(a_i + b_i)^2 \geq \mathrm{JS}(x_i, z_i)$ for $1 \leq i \leq n$. It follows that the left-hand side of (59) is larger than or equal to $\sqrt{\mathrm{JS}(x,z)}$. The right-hand side of (59) is equal to $\sqrt{\mathrm{JS}(x,y)} + \sqrt{\mathrm{JS}(y,z)}$, which implies the triangle inequality for the square root of the Jensen–Shannon divergence in n dimensions. □

Further preparations. Recall that the Shannon entropy is defined by $E(t) = t \ln t - t$. The related function, $F\colon \mathbb{R}_+ \to \mathbb{R}$, defined by

$$F(t) = \sqrt{E(t^2) + 1} \tag{60}$$

$$= \sqrt{2t^2 \ln t - t^2 + 1}, \tag{61}$$

will play a crucial role in the proof of our next theorem; see Figure 4. The derivative has a discontinuity at $t = 1$, but if we reflect the preceding branch to get $\bar{F}\colon \mathbb{R}_+ \to \mathbb{R}$ defined by $\bar{F}(t) = -F(t)$ for $0 < t \leq 1$ and $\bar{F}(t) = F(t)$ for $1 \leq t$, we obtain a convex function; see again Figure 4 on the left. Appendix A proves that $\bar{F}$ is convex and everywhere differentiable, and that its derivative, $\bar{F}'$, is concave; see again Figure 4.

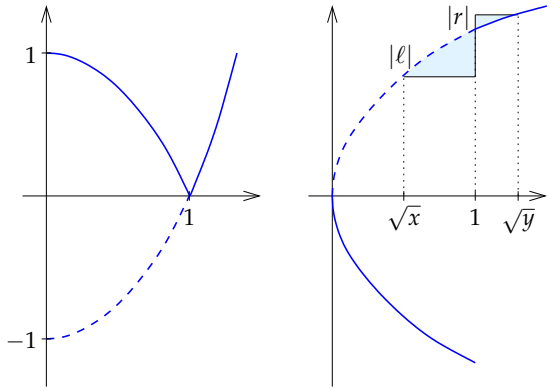

Figure 4. Left: the (solid) graph of F and the (partially dotted) graph of $\bar{F}$. **Right**: the (solid) graph of F' and the (partially dotted) graph of $\bar{F}'$. The shaded regions have area $|\ell|$ and $|r|$, as discussed in the proof of Theorem 3.

One dimension. We are now ready to prove that the square root of the minmax divergence is a metric in one dimension.

Theorem 3 (1D Metric). *Let* $E\colon \mathbb{R}_+ \to \mathbb{R}$ *be the Shannon entropy and* $\mathrm{Mx}\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ *map every pair to its minmax divergence. Then* $\sqrt{\mathrm{Mx}}$ *is a metric in* $\mathbb{R}_+$.

Proof. It is clear that $\sqrt{\mathrm{Mx}(x,y)}$ is non-negative, zero iff $x = y$, and symmetric. It thus remains to prove that it satisfies the triangle inequality. Setting

$$f(x,y) = \frac{\sqrt{\mathrm{Mx}(x,y)}}{|\sqrt{y} - \sqrt{x}|}, \tag{62}$$

we will prove shortly that f is monotonic as defined in (47). Lemma 5 then implies that $\sqrt{\mathrm{Mx}}$ satisfies the triangle inequality. It thus remains to prove $f(a,b) \leq f(x,y)$ whenever $a \leq x \leq y \leq b$. We begin by noting that we may assume these intervals are *canonical*, by which we mean that $E(a) = E(b)$ and $E(x) = E(y)$. Indeed, if $E(x) \neq E(y)$, then we can find $C > 0$ such that $E(Cx) = E(Cy)$, which then implies $f(Cx, Cy) = f(x,y)$.

We now use the function $f\colon \mathbb{R}_+ \to \mathbb{R}$ defined by $F(t) = \sqrt{E(t^2) + 1}$ to draw a geometric picture of the situation; see Figure 5.

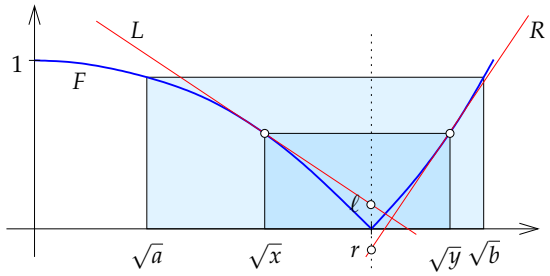

Figure 5. To improve visibility, we draw the graph of F stretched in the horizontal direction. The aspect ratios of the shaded rectangles are the values of f at x, y and at a, b. The tangent lines L at $t = \sqrt{x}$ and R at $t = \sqrt{y}$ intersect at a point above the zero line.

To prove monotonicity, that is: $f(a,b) \leq f(x,y)$, we consider the linear functions $L, R\colon \mathbb{R} \to \mathbb{R}$ that satisfy $L(t) = F(t)$, for $t = \sqrt{a}, \sqrt{x}$, and $R(t) = F(t)$, for $t = \sqrt{y}, \sqrt{b}$. The goal is to show that the aspect ratio of the rectangle defined by a, b is less than that defined by x, y. Equivalently, we show that the point at which the lines L and R meet has positive second coordinate. By the differentiability of F and the transitivity of order along the real line, it suffices to show this in the limit case, when $a = x$ and $y = b$. In

this case, L and R are the tangent lines of F at $t = \sqrt{x}$ and $t = \sqrt{y}$, as shown in Figure 5. Let $\ell = L(1)$ and $r = R(1)$ be the coordinates of the points at which L and R intersect the vertical line $t = 1$, and note that $r \leq 0 \leq \ell$ by convexity of $\bar{F}$. The convexity of this function furthermore implies

$$1 - \sqrt{x} > \sqrt{y} - 1, \tag{63}$$
$$|L'(\sqrt{x})| < |R'(\sqrt{y})|. \tag{64}$$

To compare the absolute values of ℓ and r, we express both as integrals:

$$\ell = \int_{t=\sqrt{x}}^{1} [F'(\sqrt{x}) - F'(t)]\,\mathrm{d}t, \tag{65}$$
$$r = \int_{t=1}^{\sqrt{y}} [F'(t) - F'(\sqrt{y})]\,\mathrm{d}t; \tag{66}$$

see Figure 4 on the right. The concavity of $\bar{F}'$ together with (63) implies $|\ell| \geq |r|$, and since L has smaller absolute slope than R (64), it follows that the two lines intersect above the zero line, as required. □

This concludes the proof in dimension one. The main obstacle to extending the proof to arbitrary dimension is the lack of decomposability (separability) of the minmax loss. Still, we decided to present the partial results and techniques, as they may help other researchers complete the proof in the future. In particular, techniques for extending results from one to arbitrary dimensions are present in the information theory literature; see for example the work on Pinsker's inequality [23], which compares the relative entropy with another distance. This gives us hope that researchers in information theory may be well-equipped to extend the result to arbitrary dimension.

5. Extensions to Other Bregman Divergences

In this section, we provide a perspective for our results by extending them beyond the Shannon entropy. Specifically, we weaken the bounds while keeping them tight to generalize Theorem 1 to Bregman divergences, and we prove that the square root of the minmax divergence for the Burg entropy is a metric in $\mathbb{R}_+$.

This way the inequality can be used in other applied contexts. For example, the Burg entropy (and the Itakura–Saito divergence it induces) are used in speech recognition [24].

Burbea–Rao divergence. The *Burbea–Rao divergence*, also called the Jensen–Bregman [25], is a straightforward generalization of the Jensen–Shannon divergence. Specifically, the underlying divergence is generalized from the relative entropy to a Bregman divergence. Letting $F\colon \Omega \to \mathbb{R}$ be a function of Legendre type generating a Bregman divergence and $x, y \in \Omega$, we define

$$\mathrm{BR}_F(x, y) = \tfrac{1}{2}[D_F(x\|\mu) + D_F(y\|\mu)] \tag{67}$$
$$= \tfrac{1}{2}[F(x) + F(y) - 2F(\mu)], \tag{68}$$

in which $\mu = (x + y)/2$. Lemma 1 generalizes with a verbatim proof in which we substitute F for E. As done in [1], we also generalize the minmax divergence from the Kullback–Leibler divergence to a general underlying Bregman divergence D_F:

$$\mathrm{Mx}_F(x, y) = \min_{z \in \Omega} \max\{D_F(x\|z), D_F(y\|z)\}. \tag{69}$$

When we compare the two, we get bounds that are considerably worse than for the special case of the Shannon entropy.

Theorem 4 (Burbea–Rao divergence Comparison)**.** *Let $F\colon \Omega \to \mathbb{R}$ be a Legendre type function, and let x, y be points in Ω. Then $\frac{1}{2}\mathrm{Mx}_F(x,y) \leq \mathrm{BR}_F(x,y) \leq \mathrm{Mx}_F(x,y)$.*

Proof. Let $z \in \Omega$ be the point that minimizes the right-hand side of (69). Using the generalization of Lemma 1 to the Burbea–Rao divergence, we get

$$\mathrm{BR}_F(x,y) \leq \tfrac{1}{2}[D_F(x\|z) + D_F(y\|z)]. \tag{70}$$

Since $D_F(x\|z) = D_F(y\|z)$, the right-hand side of (70) is equal to $\mathrm{Mx}_F(x,y)$, which implies the claimed upper bound on the Burbea–Rao divergence. To prove the lower bound, we note that the larger of the two divergences to μ is at least as large as $\mathrm{Mx}_F(x,y)$. Hence, $\mathrm{Mx}_F(x,y)$ is at most the sum:

$$\mathrm{Mx}_F(x,y) \leq D_F(x\|\mu) + D_F(y\|\mu), \tag{71}$$

and the claimed lower bound follows because the right-hand side of (71) evaluates to $2\mathrm{BR}_F(x,y)$. □

Remark 5. *The bounds in Theorem 4 are tight. To see this for the lower bound, we consider $F(t) = t + \frac{1}{t}$, which is strictly convex, differentiable, and with minimum at $t = 1$. Setting $x = \varepsilon$, $y = \frac{1}{\varepsilon}$ for $0 < \varepsilon < 1$, we have $F(x) = F(y) = \varepsilon + \frac{1}{\varepsilon}$, which implies that $t = 1$ minimizes the maximum divergence from x and y. Some computations show that the divergences are*

$$\mathrm{Mx}_F(x,y) = F(x) - F(1) = \varepsilon - 2 + \tfrac{1}{\varepsilon}, \tag{72}$$

$$\mathrm{BR}_F(x,y) = \tfrac{F(x)+F(y)}{2} - F(\mu) = \tfrac{\varepsilon^2+1}{2\varepsilon} - \tfrac{2\varepsilon}{\varepsilon^2+1}. \tag{73}$$

For $\varepsilon \to 0$, the ratio of $\mathrm{BR}_F(x,y)$ over $\mathrm{Mx}_F(x,y)$ goes to $\frac{1}{2}$, as required. To see the upper bound, we choose $F(t) = t^2$ for which $\mathrm{BR}_F(x,y) = \mathrm{Mx}_F(x,y)$ for all $x, y \in \mathbb{R}$.

Minmax divergence for Burg entropy. Another significant entropy appearing in this context is the *Burg entropy*: $B(x) = x - \ln(x) + 1$, for $x > 0$. Let us denote the corresponding minmax divergence my M_B.

Theorem 5 (Metric for Burg Entropy)**.** *Let $F\colon \mathbb{R}_+ \to \mathbb{R}$ be the Burg entropy. Then $\sqrt{M_B}$ is a metric in $\mathbb{R}_+$.*

Proof. The proof follows a similar structure as the proof of Theorem 3. Setting up the ratio method, we define

$$f(x,y) = \frac{\sqrt{M_B(x,y)}}{|\ln x - \ln y|}, \tag{74}$$

with the denominator being the induced path metric. By Lemma 5, we have to prove that f is monotonic.

Next we simplify. Specifically, because $D_B(x\|y) = D_B(kx\|ky)$, we have $f(kx,ky) = f(x,y)$ for all $k > 0$. We may therefore assume all intervals considered by the Ratio Method to be canonical, i.e., $B(x) = B(y)$. In such case, we have $M_B(x,y) = B(x) = B(y)$.

To prepare the geometric picture sketched in Figure 6, we define $H(x) = \sqrt{e^x - x - 1}$ for $x \in \mathbb{R}$. Note that $f(x,y)$ is the aspect ratio of the rectangle in this picture. As in the proof of Theorem 3, the monotonicity of f is established by proving that the tangent lines

to H at $\ln x$ and $\ln y$ intersect above the horizontal axis for all canonical intervals $[x, y]$. To this end, it suffices to show that $\lambda > \rho$, in which

- λ is the coordinate of the intersection of the horizontal axis with the tangent on H at $\ln x$;
- ρ is the coordinate of the intersection of the horizontal axis with the tangent on H at $\ln y$.

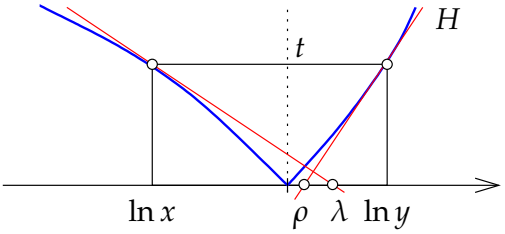

Figure 6. The two branches of the function H, two tangent lines touching the branches at points of equal height, and their intersections with the horizontal coordinate axis.

Reflecting the left part of H, we obtain an injective function $\bar{H}\colon \mathbb{R} \to \mathbb{R}$ defined by

$$\bar{H}(x) = \begin{cases} -H(x) & \text{for } x < 0, \\ H(x) & \text{for } 0 \leq x. \end{cases} \tag{75}$$

Note that this does not change the intersections mentioned above. Letting K be the inverse of $\bar{H}$, We consider the graph of K; see Figure 7.

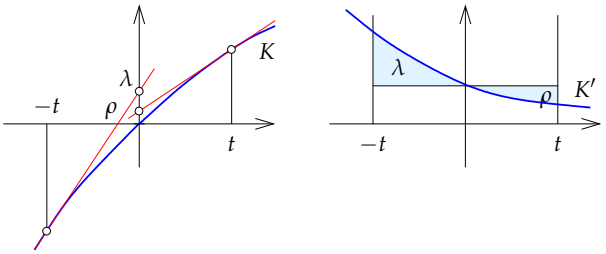

Figure 7. Left: the graph of K and the intersections of the two tangent lines with the vertical coordinate axis. **Right**: the derivative, K', and the two intersections represented as areas above and below the curve.

Expressing this with integrals, and setting $t = H(\ln x) = H(\ln y)$, we can compare the two coordinates:

$$\lambda = K(-t) + tK'(-t) = \int_0^{-t} [K'(s) - K'(-t)]\,\mathrm{d}s, \tag{76}$$

$$\rho = K(t) - tK'(t) \quad = \int_0^{t} [K'(s) - K'(t)]\,\mathrm{d}s. \tag{77}$$

By Lemmas A3 and A4, K' is convex and decreasing. Using the integral representation above, we interpret λ and ρ as areas; see Figure 3. It follows that $\lambda > \rho$ and consequently $\sqrt{M_B}$ is a metric. □

6. Discussion

The main result of our work are the tight bounds between the minmax information divergence and the Jensen–Shannon divergence (as well as analogous results for their

generalizations). The former arises naturally in the context of computational geometry and topology. Specifically, it coincides with the smallest radius of the nonempty intersection between two balls in the geometry induced by the relative entropy (which generalizes to other Bregman divergences). In this setting, this quantity is calculated repeatedly, for example to compute the one-skeleton of the Čech complex, which is one of the standard constructions in topological data analysis. In this case, our bounds are best presented as

$$\sqrt{\mathrm{JS}(x,y)} \leq \sqrt{\mathrm{Mx}(x,y)} \leq \sqrt{\tfrac{2}{e \ln 2}\mathrm{JS}(x,y)} \approx 1.030 \sqrt{\mathrm{JS}(x,y)}. \tag{78}$$

One can therefore closely approximate the relatively costly computations of the minmax divergence, with the straightforward and efficient computations of the Jensen–Shannon divergence. Indeed, computing the minmax information requires performing a binary search or another numerical algorithm. On the other hand the formula for the Jensen–Shannon divergences is not more complex then computing the Euclidean distance.

The tightness of the bounds makes it believable that the square root of the minmax divergence may be a metric, similar to the Jensen–Shannon divergence. This turned out to be true in the one dimensional case, but the general case of n-dimensional spaces appears to be significantly more difficult. One key reason is that—unlike the Jensen–Shannon divergence—the minmax divergence is not decomposable (separable). We leave this case open and hope that a combinations of the proposed geometric proof techniques with additional techniques may eventually lead to a successful resolution.

Author Contributions: Conceptualization, A.A., H.E., Ž.V. and H.W.; methodology, A.A., H.E., Ž.V. and H.W.; investigation, A.A., H.E., Ž.V. and H.W.; writing—original draft preparation, A.A., H.E., Ž.V. and H.W.; writing—review and editing, H.E. and H.W.; funding acquisition, H.E. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received partial funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, grant no. 788183, the Wittgenstein Prize, Austrian Science Fund (FWF), grant no. Z 342-N31, the DFG Collaborative Research Center TRR 109, 'Discretization in Geometry and Dynamics', Austrian Science Fund (FWF), grant no. I 02979-N35, and the 2022 Google Research Scholar Award for project 'Algorithms for Topological Analysis of Neural Networks'. The APC was waived.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: Author Arseniy Akopyan was employed by the company Fora Capital. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. First Curve Discussion

In this appendix, we prove that the function $\bar{F}$ used in the proof of Theorem 3 is convex and that its derivative is concave.

Convexity of $\bar{F}$. Recall that the functions $F, \bar{F}\colon \mathbb{R}_+ \to \mathbb{R}$ used in the proof of Theorem 3 are defined by

$$F(t) = \sqrt{E(t^2)+1}, \tag{A1}$$

$$\bar{F}(t) = \begin{cases} -F(t) & \text{for } 0 < t \leq 1, \\ F(t) & \text{for } 1 \leq t, \end{cases} \tag{A2}$$

in which $E(t) = t \ln t - t$ is the Shannon entropy; see Figure 4, left.

Lemma A1 (Convexity of Function)**.** *The function* $\bar{F}\colon \mathbb{R}_+ \to \mathbb{R}$ *is convex and differentiable.*

Proof. We aim at proving that $\bar{F}''(t)$ is positive for all $t > 0$. This is equivalent to $\bar{F}''(t)^2 = F''(t)^2$ being positive for all $t > 0$, provided the second derivative is continuous and positive at some $t > 0$. Indeed, for $\bar{F}''$ to change sign, its square must vanish. We compute

$$F'(t) = \frac{2t \ln t}{F(t)}, \tag{A3}$$

$$F''(t) = \frac{(2 \ln t + 2)F(t) - 2t \ln t F'(t)}{F(t)^2} \tag{A4}$$

$$= \frac{2[(t^2+1)\ln t - (t^2-1)]}{F(t)^3}, \tag{A5}$$

noting that $F''(e) = 4/(e^2+1)^{3/2}$, which is positive as required. The numerator is twice $a(t) = (t^2+1)\ln t - (t^2-1)$, which vanishes iff $\ln t = (t^2-1)/(t^2+1)$. The derivatives on the two sides satisfy $1/t < 4t/(t^2+1)^2$ for all positive $t \neq 1$ because $(t^2-1)^2 > 0$ for all positive $t \neq 1$. Since $a(1) = 0$, this implies that $t = 1$ is the only positive root of the numerator. It follows that F''^2 is positive everywhere, except possibly at $t = 1$, where we get $F''(1) = \frac{0}{0}$. We settle this case with the rule de l'Hôpital, computing derivatives of the numerator and the denominator. We begin with the numerator:

$$a'(t) = 2t \ln t + \tfrac{t^2+1}{t} - 2t, \tag{A6}$$

$$a''(t) = 2 \ln t + \tfrac{t^2-1}{t^2}, \tag{A7}$$

$$a'''(t) = \tfrac{2}{t} + \tfrac{2}{t^3}, \tag{A8}$$

with $a(1) = a'(1) = a''(1) = 0$ and $a'''(1) = 4$. Deriving $N = a^2$, we get a structure similar to the Pascal triangle with a non-zero term after six steps:

$$N' = 2aa', \tag{A9}$$

$$N'' = 2a'^2 + 2aa'', \tag{A10}$$

$$N''' = 6a'a'' + 2aa''', \tag{A11}$$

$$N'''' = 6a''^2 + 8a'a''' + 2aa'''', \tag{A12}$$

$$N''''' = 20a''a''' + 10a'a'''' + 2aa''''', \tag{A13}$$

$$N'''''' = 20a'''^2 + 30a''a'''' + 12a'a''''' + 2aa''''''. \tag{A14}$$

All derivatives of a are well defined and take on finite values at $t = 1$. Since $a(1) = a'(1) = a''(1) = 0$ and $a'''(1) \neq 0$, all derivatives of N vanish at $t = 1$, except for the last, which is $N''''''(1) = 20a'''(t)^2 = 320$. To do the same for the denominator in (A5), we set $c(t) = F(t)^2 = 2t^2 \ln t - t^2 + 1$ and compute two derivatives,

$$c'(t) = 4t \ln t, \tag{A15}$$

$$c''(t) = 4 \ln t + 4, \tag{A16}$$

with $c(1) = c'(1) = 0$ and $c''(1) = 4$. Starting with $D = c^3$, we thus get

$$D' = 3c^2c', \tag{A17}$$
$$D'' = 6cc'^2 + 3c^2c'', \tag{A18}$$
$$D''' = 6c'^3 + 18cc'c'' + 3c^2c''', \tag{A19}$$
$$D'''' = 36c'^2c'' + 18cc''^2 + 24cc'c''' + 3c^2c'''', \tag{A20}$$
$$D''''' = 90c'c''^2 + 60c'^2c''' + 60cc''c''' + 30cc'c'''' \tag{A21}$$
$$+ 3c^2c''''', \tag{A22}$$
$$D'''''' = 90c''^3 + 360c'c''c''' + 90c'^2c'''' + 60cc'''^2 \tag{A23}$$
$$+ 90cc''c'''' + 36cc'c''''' + 3c^2c''''''. \tag{A24}$$

All derivatives of c are well defined and take on finite values at $t = 1$. Since $c(1) = c'(1) = 0$ and $c''(1) \neq 0$, all derivatives of D vanish at $t = 1$, except for the last, which is $D''''''(1) = 90c''(1)^3 = 5{,}760$. We conclude that F''^2 is positive at $t = 1$, namely $F''(1)^2 = \frac{4 \cdot 320}{5{,}760} = 0.222\ldots$ This implies that $\bar{F}$ is convex as well as differentiable. □

Concavity of $\bar{F}'$. Since $\bar{F}$ is differentiable, its derivative $\bar{F}' \colon \mathbb{R}_+ \to \mathbb{R}$ exists; see Figure 4, right.

Lemma A2 (Concavity of Derivative)**.** *The function $\bar{F}' \colon \mathbb{R}_+ \to \mathbb{R}$ is concave.*

Proof. By definition, $\bar{F}'$ is concave iff $\bar{F}''$ is monotonically decreasing, which is implied if $\bar{F}''^2 = F''^2$ is monotonically decreasing. Recall from (A5) that $F''(t)^2 = 4a(t)^2/F(t)^6$. The derivative of the squared second derivative is therefore

$$(F''(t)^2)' = \frac{8a(t)a'(t)F(t)^6 - 24a(t)^2F(t)^5F'(t)}{F(t)^{12}} \tag{A25}$$
$$= \frac{a(t)}{t} \cdot \frac{8F(t)^4 - 48a(t)t^2 \ln t}{F(t)^8}, \tag{A26}$$

in which we get (A26) using $a'(t) = F(t)^2/t$ and $F(t)F'(t) = 2t \ln t$. To continue, we write $b(t)$ for the numerator of the second factor in (A26). Since $tF(t)^8$ is non-negative and the sign of $a(t)$ is the same as the sign of $t - 1$, it suffices to show that the sign of $b(t)$ is that same as the sign of $1 - t$. But $b(1) = 0$, so it is enough to show that b is monotonically decreasing, which it is iff $b_1(t) = b(\sqrt{t})/4$ is monotonically decreasing. We get

$$b_1(t) = 2F(\sqrt{t})^4 - 6a(\sqrt{t})t \ln t \tag{A27}$$
$$= 2[t \ln t - t + 1]^2$$
$$- 3[(t+1) \ln t - 2(t-1)]t \ln t \tag{A28}$$
$$= (2t^2 - 2t) \ln t - (t^2 + 3t) \ln^2 t + 2(t-1)^2, \tag{A29}$$

in which we use $F(\sqrt{t})^2 = E(t) + 1 = t \ln t - t + 1$ as well as $a(\sqrt{t}) = \frac{1}{2}(t+1) \ln t - (t-1)$ to get (A28). Computing the derivative, we get

$$b_1'(t) = (2t - 8) \ln t - (2t + 3) \ln^2 t + 6t - 6, \tag{A30}$$

which we need to be non-positive. Since $b_1'(1) = 0$, it suffices to show that the sign of $b_1''(t)$ agrees with the sign of $1 - t$. Computing the second derivative, we get

$$b_1''(t) = -\tfrac{2}{t}[(t+3) \ln t + t \ln^2 t - 4t + 4]. \tag{A31}$$

The first factor, $-\frac{2}{t}$, is always negative, so we just need to show that the second factor, which we write as $b_2(t)$, has the same sign as $t-1$. This is indeed the case, which we see because $b_2(1)=0$ and

$$b_2'(t) = \ln^2 t + 3\ln t - 3 + \tfrac{3}{t} \tag{A32}$$

$$= \ln^2 t + \tfrac{3}{t}[E(t)+1] \tag{A33}$$

is non-negative for all $t \in \mathbb{R}_+$. We summarize by recalling the chain of implications from back to front:

- $b_2'(t)$ is non-negative and $\operatorname{sign}(b_2(t)) = \operatorname{sign}(t-1)$,
- $\operatorname{sign}(b_1''(t)) = \operatorname{sign}(1-t)$ and $b_1'(t)$ is non-positive,
- $\operatorname{sign}(b(t)) = \operatorname{sign}(1-t)$ and $(F''(t)^2)'$ is non-positive,
- F''^2 and $\bar{F}''$ are monotonically decreasing,
- and finally, $\bar{F}'$ is concave.

The only remaining uncertainty is at $t=1$, where both the numerator and the denominator of F'' vanishes. But as shown in the proof of Lemma A1, $F''(1)^2 = 0.222\ldots$ is well defined, which implies that $\bar{F}'$ is differentiable at $t=1$ and thus concave over all of $\mathbb{R}_+$. □

Appendix B. Second Curve Discussion

Recall that K is the inverse function of $\bar{H}$ used in the proof of Theorem 5. In this appendix, we prove that the derivative of K is convex and decreasing.

Function K**.** Recall that the functions $H, \bar{H}\colon \mathbb{R} \to \mathbb{R}$ used in the proof of Theorem 5 are defined by

$$H(x) = \sqrt{e^x - x - 1}, \tag{A34}$$

$$\bar{H}(x) = \begin{cases} -H(x) & \text{for } x < 0, \\ H(x) & \text{for } 0 \le x. \end{cases} \tag{A35}$$

It is easy to check that $\bar{H}$ is continuously differentiable with a positive derivative. By the Inverse Function Theorem, its inverse, K, exists and is continuously differentiable.

Lemma A3 (Convexity of Derivative)**.** *The derivative of the function* $K\colon \mathbb{R} \to \mathbb{R}$ *is convex.*

Proof. Consider first the case $x > 0$, introduce $y(x) = \sqrt{e^x - x - 1}$, and note that $y'(x) = \frac{\mathrm{d}y}{\mathrm{d}x} = \frac{e^x-1}{2y(x)}$. Setting $x'(y) = \frac{\mathrm{d}x}{\mathrm{d}y} = \frac{2y}{e^x-1}$, we prove that $x'(y)$ is convex or, equivalently, that $x''(y)$ is increasing. Using the definition of $y(x)$, we get

$$x''(y(x)) = 2\frac{(e^x-1) - ye^x x'(y)}{(e^x-1)^2} \tag{A36}$$

$$= 2\frac{-e^{2x} + 2xe^x + 1}{(e^x-1)^3}. \tag{A37}$$

Since both $x(y)$ and $y(x)$ are increasing functions, it suffices to show that $g_1(x) = x''(y(x))$ is increasing. Note that

$$g_1'(x) = \tfrac{e^x}{(e^x-1)^4}\, g_2(x), \tag{A38}$$

in which $g_2(x) = e^x - 4e^x(x-1) - 2x - 5$. At this point, we need to prove that $g_2(x)$ is positive. Observe that $g_2''(x) = 4e^x(e^x - x - 1) > 0$ and $g_2(0) = g_2'(0) = 0$, thus $g_2(x) > 0$ for $x > 0$. Tracing back the chain of derivations, we conclude that $K' = x'(y)$ is convex.

Consider second the case $x < 0$, introduce $y(x) = -\sqrt{e^x - x - 1}$, and note that $y'(x) = \frac{dy}{dx} = \frac{e^x - 1}{2y(x)}$, as above. Hence, we can repeat the steps using the same functions but now for $x < 0$. We conclude that $x'(y)$ is convex for $x > 0$ and for $x < 0$. We know that it is continuous everywhere, and continuously differentiable for $x \neq 0$ by the calculations above. Using (A36) and (A37), we may also verify that $x''(y)$ is continuous at $y = 0$, hence everywhere. This implies that $K = x'(y)$ is convex on $\mathbb{R}$. □

Lemma A4 (Decreasing Derivative)**.** *The derivative of the function* $K\colon \mathbb{R} \to \mathbb{R}$ *is decreasing.*

Proof. Consider first $x > 0$. By (A36) and (A37), it suffices to prove that

$$g_3(x) = -e^{2x} + 2xe^x + 1 < 0. \tag{A39}$$

Indeed, this holds as $g_3'(x) = -2e^x(e^x - (x+1)) < 0$ and $g_3'(0) = 0$. Since the derivative of K is convex, for all x, and decreasing, for $x > 0$, it must also be decreasing for $x \leq 0$. □

References

1. Nielsen, F. Revisiting Chernoff information with likelihood ratio exponential families. *Entropy* **2022**, *24*, 1400. [CrossRef] [PubMed]
2. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
3. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000.
4. Endres, D.M.; Schindelin, J.E. A new metric for probability distributions. *IEEE Trans. Inf. Theory* **2003**, *49*, 1858–1860. [CrossRef]
5. Edelsbrunner, H.; Wagner, H. Topological data analysis with Bregman divergences. In Proceedings of the 33rd International Symposium on Computational Geometry (SoCG 2017), Brisbane, Australia, 4–7 July 2017; Leibniz-Zentrum für Informatik: Wadern, Germany, 2017; pp. 391–3916.
6. Edelsbrunner, H.; Virk, Z.; Wagner, H. Smallest enclosing spheres and Chernoff points in Bregman geometry. In Proceedings of the 34th International Symposium on Computational Geometry (SoCG 2018), Budapest, Hungary, 11–14 June 2018; Leibniz-Zentrum für Informatik: Wadern, Germany, 2018; pp. 35:1–35:13.
7. Carlsson, G.; Vejdemo-Johansson, M. *Topological Data Analysis with Applications*; Topological Data Analysis with Applications; Cambridge University Press: Cambridge, UK, 2021.
8. Chazal, F.; De Silva, V.; Oudot, S. Persistence stability for geometric complexes. *Geom. Dedicata* **2014**, *173*, 193–214. [CrossRef]
9. Uhlmann, J.K. Satisfying general proximity/similarity queries with metric trees. *Inf. Process. Lett.* **1991**, *40*, 175–179. [CrossRef]
10. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
11. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
12. Burbea, J.; Rao, C.R. On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Inf. Theory* **1982**, *28*, 489–495. [CrossRef]
13. Bregman, L. The relaxation method of finding the common point of convex sets and its applications to the solution of problems in convex programming. *USSR Comput. Math. Math. Phys.* **1967**, *7*, 200–217. [CrossRef]
14. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences. *Commun. Math. Sci.* **2008**, *6*, 915–926. [CrossRef]
15. Chen, P.; Chen, Y.; Rao, M. Metrics defined by Bregman divergences: Part 2. *Commun. Math. Sci.* **2008**, *6*, 927–948. [CrossRef]
16. Nielsen, F. The Bregman chord divergence. In Proceedings of the Geometric Science of Information (GSI 2019), Toulouse, France, 27–29 August 2019; Springer: Cham, Switzerland, 2019; Volume 11712, pp. 269–276. [CrossRef]
17. Nielsen, F. Symplectic Bregman divergences. *Entropy* **2024**, *26*, 1101. [CrossRef] [PubMed]
18. Nielsen, F. Curved representational Bregman divergences and their applications. *arXiv* **2025**, arXiv:2504.05654. [CrossRef]
19. Nielsen, F.; Nock, R. On the smallest enclosing information disk. In Proceedings of the 18th Canadian Conference on Computational Geometry, Kingston, ON, Canada, 14–16 August 2006.
20. Nielsen, F.; Nock, R. The dual Voronoi diagrams with respect to representational Bregman divergences. In Proceedings of the 2009 Sixth International Symposium on Voronoi Diagrams, Copenhagen, Denmark, 23–26 June 2009; pp. 71–78.
21. Nock, R.; Nielsen, F. Fitting the smallest enclosing Bregman ball. In Proceedings of the Machine Learning: ECML 2005, Porto, Portugal, 3–7 October 2005; pp. 649–656. [CrossRef]
22. Banerjee, A.; Merugu, S.; Dhillon, I.S.; Ghosh, J. Clustering with Bregman divergences. *J. Mach. Learn. Res.* **2005**, *6*, 1705–1749.
23. Rioul, O. A historical perspective on Schützenberger-Pinsker inequalities. In Proceedings of the International Conference on Geometric Science of Information, St. Malo, France, 30 August–1 September 2023; Springer: Cham, Switzerland, 2023; pp. 291–306.

24. Juang, B.H. On using the Itakura-Saito measures for speech coder performance evaluation. *AT&T Bell Lab. Tech. J.* **1984**, *63*, 1477–1498. [CrossRef]
25. Nielsen, F.; Boltz, S. The Burbea-Rao and Bhattacharyya centroids. *IEEE Trans. Inf. Theory* **2011**, *57*, 5455–5466. [CrossRef]