Jensen-Shannon Divergence and norm-based measures of Discrimination and Variation

Flemming Topsøe*
Department of Mathematics
University of Copenhagen

060303, still incomplete first version - moved to sh.tex

Abstract

Jensen-Shannon divergence is a symmetrized and smoothed version of the all important divergence measure of informaton theory, *Kullback-Leibler divergence*. Recently it has been discovered that Jensen-Shannon divergence defines a true metric – precisely, it is the square of a metric.

In the first part of the paper we strengthen this result and discuss further properties of Jensen-Shannon divergence. Then, in the second part, we use considerations of classical norms to define a two-parameter family of measures of discrimination which contains the Jensen-Shannon divergence. The approach taken is based on a preliminary introduction of measures of variation, referred to as *variation breadth* which is believed to be of independent interest. These measures are defined for functions modelling variation in *features*, using terminology borrowed from pattern recognition. Specializing the set-up we obtain measures of divergence and can return to the study of metric properties.

Keywords Jensen-Shannon divergence, total variation, Kullback-Leibler divergence, information diagrams, variation breadth.

*Supported by the Danish Natural Science Resarch Council and by INTAS, project 00-738.

1 Introduction

An overall aim of the present paper is to point to some of the possibilities to use information theoretical principles in statistics. Based on previous research starting with Kullback [12], cf. also the survey article xx, the view taken is by now well established. However, certain basic elements related to coding deserves more focus as this leads to sound interpretations and guiding principles for statistical inference.

In particular, it is often fruitful to exploit the one-to-one relationship between (discrete) distributions and codes. In this way, problems formulated as problems of selecting some sensible distribution may be thought of, instead, as problems of choosing a good description defined in terms of the associated code. We shall only briefly indicate the necessary interpretations needed to make the above statements meaningful, and realize that acquaintance with background material as found in e.g. Cover and Thomas [4] may be useful. We can also point to the authors papers [24] or [25] for quick introductions.

The main part of the present text is taken up by a specialized study of some of the information theoretical measures of discrimination which find applications in statistics. In particular, we shall study the so-called *Jensen-Shannon divergence*. As point of departure we take the notion of *entropy* introduced by Shannon in his celebrated 1948 paper [22] for discrete distributions:

$$H(P) = -\sum_n p_n \log p_n \,.$$

Here, the p_n's are the point probabilities of P and we take "log" to denote the natural logarithm. We may take P to be supported by $\mathbb{N}$, in which case we write $P \in M_+^1(\mathbb{N})$, or by a finite subsection $\{1, \cdots, k\}$ of $\mathbb{N}$, in which case we write $P \in M_+^1(k)$.

The important concavity property of the entropy function is best expressed after the introduction of the *Kullback-Leibler divergence* $D(\cdot\|\cdot)$. This measure, introduced in [13], is given by

$$D(P\|Q) = \sum_n p_n \log \frac{p_n}{q_n}$$

in the discrete case (with Q a second distribution with point probabilities q_n and with usual conventions regarding $x \log \frac{x}{y}$ in extreme cases). For the continuous case the formula

$$D(P\|Q) = \int \log \frac{dP}{dQ} dP$$

applies. For the main part of our discussion we concentrate on the discrete case as divergence can be obtained therefrom by a simple limiting process, cf. [24, Theorem 5.2].

Therefore, results for the discrete case lead in a natural way to results for distributions on general Borel spaces. This only applies when divergence and related measures of discrepancy are concerned. Entropy is tied naturally to the discrete case.

Returning to concavity of the entropy function, consider a finite or infinite mixture $\sum_\nu \alpha_\nu P_\nu$ of probability distributions (α_ν's are all non negative and sum to 1). Put $\overline{P} = \sum_\nu \alpha_\nu P_\nu$. Then

$$H\left(\sum_\nu \alpha_\nu P_\nu\right) = \sum_\nu \alpha_\nu H(P_\nu) + \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) \tag{1}$$

and Jensen's inequality confirming concavity, viz. the inequality

$$H(\sum \alpha_\nu P_\nu) \geq \sum \alpha_\nu H(P_\nu)\,, \tag{2}$$

then follows from the most fundamental inequality of information theory: $D(P\|Q) \geq 0$ (with equality if and only if $P = Q$). In case $\sum_\nu \alpha_\nu H(P_\nu) < \infty$, which is surely the case if we consider distributions of fixed finite support, we may write (1) in the form

$$H\left(\sum_\nu \alpha_\nu P_\nu\right) - \sum_\nu \alpha_\nu H(P_\nu) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P})\,. \tag{3}$$

The left hand side of (3) we call the *general Jensen-Shannon Divergence* pertaining to the mixture $\sum_\nu \alpha_\nu P_\nu$. In order not to overcomplicate the notation we write, by abuse of notation, $\mathrm{JSD}(\sum \alpha_\nu P_\nu)$ for this quantity whenever confusion seems unlikely.

The right hand side of (3) has important advantages over the left hand side: On the technical side, it is always well defined, even for distributions over arbitrary Borel spaces. Therefore, as defining relation for the general JSD-divergence we take the equation

$$\mathrm{JSD}(\sum \alpha_\nu P_\nu) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) \tag{4}$$

(with $\overline{P} = \sum_\nu \alpha_\nu P_\nu$ as above).

Another advantage of the expression (4) over the left hand side of (3) is the interpretations it gives rise to. In fact, the quantity is the information transmission rate for a discrete memoryless channel with input letters indexed by ν, each sent with probability α_ν, and with the P_ν's as conditional distributions on the output side. This interpretation will not be pursued in the following.

A second interpretation of more interest to statisticians relates to the model which we shall refer to as the *switching model* where a source generates a string $x_1x_2\cdots$ of "letters", each letter selected independently of previous letters and according to a specific distribution among the P_ν's and in such a way that the probability that P_ν is used for the selection is α_ν. Consider an observer who knows the P_ν's as well as the α_ν's but does not know precisely which distribution is used at any particular time instant. The observer wants to design a code such that the expected *redundancy* is minimized. By "redundancy" we have the following in mind:

An *ideal observer* who knows at each time instant which distribution was used for the selection at the source of the letter sent, will use this insight to choose, at each instant, a code adapted to this distribution. Hence, the ideal observer will use in average $H(P_\nu)$ nits (bits transformed to natual units based on the natural logarithm function) for his observations in all instances where P_ν was used for the selection at the source. The actual observer cannot know when this or that distribution is used at the source and has to choose one and the same code, say κ, at each time instant. There is a one-to-one correspondance between distributions and codes (understood in a certain idealized sense, allowing codewords of any length, not necessarily tied to the natural numbers). If Q and κ correspond to each other under this correspondance, the code length used for the letter with index n is $-\log q_n$. If the code κ used by the actual observer corresponds to Q, then the average amount of nits used by the observer is the average $\sum_n p_{\nu,n} \log \frac{1}{q_n}$ in case $P_\nu = (p_{\nu,n})$ is the actual distribution. The redundancy in such cases is, therefore, the difference between this number and $H(P_\nu)$, i.e. the redundancy is $D(P_\nu \| Q)$. The average redundancy is then $\sum_\nu \alpha_\nu D(P_\nu \| Q)$.

To summarize, the observer should choose that distribution Q as the basis for coding which minimizes average redundancy given by

$$R(Q) = \sum_\nu \alpha_\nu D(P_\nu \| Q)\,. \tag{5}$$

In order to identify the associated argmin-distribution, we refer to an identity of great significance for optimization problems of information theory, viz. the socalled *compensation identity.* According to this identity, the equation

$$\sum_\nu \alpha_\nu D(P_\nu \| Q) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) + D(\overline{P} \| Q) \tag{6}$$

holds for any distribution Q.

The identity has numerous analogs and applications, cf. Topsøe [6].

For our discussion, we realize that it follows immediately from the identity that $Q = \overline{P}$ is the unique argmin-distribution sought and that $\mathrm{JSD}(\sum_\nu \alpha_\nu P_\nu)$ is the corresponding minimum value. Therefore, the general Jensen-Shannon divergence can also be interpreted as *minimum redundancy* for the switching model.

We also learn from (6) that for any fixed Q, divergence $D(\cdot\|Q)$ is a convex function:

$$D\left(\sum_\nu \alpha_\nu P_\nu \| Q\right) \le \sum_\nu \alpha_\nu D(P_\nu \| Q)\,. \tag{7}$$

Furthermore, we realize that if we apply the same strategy of definition as the one applied to the entropy function and consider the "Jensen-type" divergence, looking at the difference between the right hand and the left hand side in (7), then nothing new is obtained – we are back to the quantity (4), we started with and this even quite independently of the distribution Q we choose to take as our reference.

The general Jensen-Shannon divergence (4) then is a universal type of divergence with natural interpretations, either as an information transmission rate or as minimum redundancy in a switching model with changing statistics at unknown instances.

It seems that the importance of this general divergence measure has not previously been fully recognized. However, especially if we turn to the case of an even mixture of two distributions – which is the case we shall now look into – the quantity has been discussed before and appears to be gaining in popularity among statisticians. Somewhat implicitly the quantity was introduced in Wong and You [26, Definition 13]. Lin and Wong derives some simple properties in [15] and this information is essentially repeated in [14]. More recently, the author derived further identities and inequalities in [23] and for other recent references with a focus on statistical applications, see El-Yaniv et al [20] and Minka [17].

We introduce the notation $\mathrm{JSD}(P,Q)$ for $\mathrm{JSD}(\frac{1}{2}P + \frac{1}{2}Q)$, i.e.

$$JSD(P,Q) = \frac{1}{2}D(P\|M) + \frac{1}{2}D(Q\|M)\,, \tag{8}$$

where $M = \frac{1}{2}P + \frac{1}{2}Q$. As opposed to the general divergence, this is the *specific Jensen-Shannon divergence* between P and Q. It may be thought of as a symmetrized and smoothed variant of Kullback-Leibler divergence. Symmetrization is evident as the two terms in (8) are symmetric in a natural sense and smoothing is achieved by consideration of the midpoint distribution M. The specific Jensen-Shannon divergence may be expressed in terms of the midpoint probabilities $m_n = \frac{1}{2}(p_n + q_n)$ and the probability distributions $T_n \in M_+^1(2)$ given by $T_n = \frac{1}{2m_n}(p_n, q_n)$. It is convenient to introduce the function $x \curvearrowright \overline{x}$ on $\mathbb{R}_+$ defined by

$$\overline{x} = x \log x \,. \tag{9}$$

One then finds for distributions in $M_+^1(\mathbb{N})$ that

$$\begin{aligned} \mathrm{JSD}(P, Q) &= \sum_{n=1}^{\infty} m_n D(T_n \| (\tfrac{1}{2}, \tfrac{1}{2})) \\ &= \log 2 - \sum_{n=1}^{\infty} m_n H(T_n) \\ &= \log 2 - \frac{1}{2} \sum_{n=1}^{\infty} (\overline{p_n + q_n} - \bar{p}_n - \bar{q}_n) \,. \end{aligned} \tag{10}$$

For the remainder of this paper we shall work mainly with the specific divergence measure just introduced and, therefore, when we refer simply to *Jensen-Shannon divergence* it is the specific divergence measure we have in mind. In [23] the author suggested the terminology "capacitory discrimination", but in view of previous usage as in references quoted above and because of its greater suggestive flavour with the two good names of Jensen and Shannon, the Jensen-Shannon terminology should be promoted.

2 Metric properties of Jensen-Shannon Divergence

In Endres and Schindelin [7] and in Österreicher and Vajda [19] it is proved that the Jensen-Shannon divergence is the square of a metric. We shall prove this result by a method which is quite different from any of the methods of these authors and which proves a stronger result. The basis is the following result, due to Schoenberg, cf. [21] and the text book presentation in [3, Proposition 3.2]:

Theorem 1 (Schoenberg's theorem). *Let X be a set and $K : X \times X \to \mathbb{R}$ a mapping (a kernel) such that, for $(x, y) \in X \times X$,*

$$K(x, y) = 0 \iff x = y \,, \tag{11}$$

$$K(x, y) = K(y, x) \,, \tag{12}$$

$$K(x, y) \geq 0 \tag{13}$$

and such that, for all finite sets $(c_i)_{i\le n}$ of real numbers and all corresponding finite sets $(x_i)_{i\le n}$ of points in X, the implication

$$\sum_{i=1}^{n} c_i = 0 \implies \sum_{i,j} c_i c_j K(x_i, x_j) \le 0 \tag{14}$$

holds. Then $(X, \sqrt{K})$ is a metric space which can be embeded isometrically as a subspace of a real Hilbert space.

So, what we have to prove is that the kernel JSD is negative definite on $X \times X$ with $X = M_+^1(\mathbb{N})$. Since the kernel can be written as a sum:

$$\mathrm{JSD}(P,Q) = \frac{1}{2}\sum_n \left(p_n \log \frac{2p_n}{p_n + q_n} + q_n \log \frac{2q_n}{p_n + q_n} \right), \tag{15}$$

it suffices to show that each summand is negative definite on $\mathbb{R}_+ \times \mathbb{R}_+$. These summands are related to the kernel

$$(x,y) \curvearrowright x \log \frac{2x}{x+y} + y \log \frac{2y}{x+y}$$

on $\mathbb{R}_+ \times \mathbb{R}_+$. Clearly, it suffices to show that this kernel is negative definite. As a further simplification we note that

$$\sum_{i,j} c_i c_j f(x_i) = 0$$

for any real function f on $\mathbb{R}$ and any finite set of scalars (c_i) with $\sum c_i = 0$. Therefore, it suffices to prove that the kernel

$$(x,y) \curvearrowright -(x+y)\log(x+y)$$

on $\mathbb{R}_+$ is negative definite.

By definition, a function $f : \mathbb{R}_+ \to \mathbb{R}$ is *negative definite* if the associated kernel $(x,y) \curvearrowright f(x+y)$ is negative definite, cf. [3]. Therefore, what we have to prove is that $\varphi : x \curvearrowright -x\log x$ is negative definite on $\mathbb{R}_+$. This fact is stated explicitly in [3, Example 6.5.16]. For the convenience of the reader we include a proof of this fact.

What we need is some tricks from harmonic analysis. Firstly we note that for all $x \ge 0$,

$$-x\frac{x^\alpha - 1}{\alpha} \uparrow -x\log x \text{ as } \alpha \downarrow 0\,.$$

Clearly the identity function $x \curvearrowright x$ is negative definite. By the above limit relation it follows that if we can show that the function $x \curvearrowright -x^{\alpha+1}$ is negative definite, say for all $0 \leq \alpha \leq 1$, then the desired negative definiteness of the function φ will follow. The cases $\alpha = 0$ and $\alpha = 1$ are easy to handle and the desired result follows by a kind of interpolation. For this to run smoothly, we use the following representation formula for the power functions involved:

$$-x^{\alpha+1} = \frac{\alpha(1+\alpha)}{\Gamma(1-\alpha)} \int_0^\infty \frac{1 - e^{-tx} - tx}{t^{\alpha+2}} dt \,, \tag{16}$$

cf. [3, proof of Corollary 3.2.11].[1]

What (16) shows is that, for $0 < \alpha < 1$ fixed, $x \curvearrowright -x^{\alpha+1}$ can be obtained as the limit of linear combinations with positive coefficients of functions of the type $x \curvearrowright 1 - e^{-tx} - tx$. Each such function is negative definite since $x \curvearrowright 1 - tx$ is so and since, for real numbers $c_1, \cdots, c_n$ (with $\sum_1^n c_i = 0$) and non-negative real numbers $x_1, \cdots, x_n$,

$$\sum_{i,j} c_i c_j \cdot \left(-e^{-t(x_i + x_j)}\right) = -\left(\sum_i c_i e^{-tx_i}\right)^2 \leq 0 \,.$$

Backtracing our steps and appealing to Schoenberg's theorem we have proved the following main result:

Theorem 2. *The kernel* JSD *is negative definite on* $M_+^1(\mathbb{N}) \times M_+^1(\mathbb{N})$. *Therefore, denoting by* $\mathcal{H}$ *a real separable Hilbert space, there exists a subset* $\mathcal{H}_{\mathrm{JSD}} \subseteq \mathcal{H}$ *and a one-to-one bijection* Φ *between* $M_+^1(\mathbb{N})$ *and* $\mathcal{H}_{\mathrm{JSD}}$ *such that, for all* $(P, Q) \in M_+^1(\mathbb{N}) \times M_+^1(\mathbb{N})$,

$$\mathrm{JSD}(P, Q) = \|\Phi(P) - \Phi(Q)\|^2$$

with $\|\cdot\|$ *denoting the norm in* $\mathcal{H}$.

The theorem indicates that interesting geometrical properties are associated with the Jensen-Shannon divergence. However, in order to study this further, it seems that one would need more information on the Hilbert space embedding Φ. We shall not pursue this topic further here. Instead, we focus on the fact that $(M_+^1(\mathbb{N}), \sqrt{\mathrm{JSD}})$ is a metric space and answer the natural question concerning identification of the induced topology:

[1] This follows by integration in the formula

$$x^\alpha = \frac{\alpha}{\Gamma(1-\alpha)} \int_0^\infty \frac{1 - e^{-tx}}{t^{\alpha+1}} dt$$

which, in turn, can be easily checked by differentiation. As usual, Γ denotes the *gamma function*: $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$; $a > 0$.

Theorem 3. *The metric space* $(M_+^1(\mathbb{N}), \sqrt{\mathrm{JSD}})$ *is complete and bounded. The induced topology is that of convergence in total variation. Indeed, denoting total variation by* V*, the following inequalities hold for any pair of distributions in* $M_+^1(\mathbb{N})$*:*

$$\frac{1}{8}V^2(P,Q) \leq \mathrm{JSD}(P,Q) \leq \frac{1}{2}\log 2 \cdot V(P,Q)\,. \tag{17}$$

In fact, the inequalities (17) were proved in [23, Theorem 2 and (11)]. In the next section we shall give independent proofs of these inequalities together with more precise results. We emphasize that the two theorems of this section carry over to the continuous case without difficulty.

We end this section by pointing out that without recourse to the rather deep theorem of Schoenberg, one can establish the central metric property directly from negative definiteness:

Proposition 4. *Assume that* $K : X \times X \to \mathbb{R}$ *is a negative definite kernel in the sense that* (11)-(14) *hold. Then* K *is a squared metric. Regarding the condition* (14) *it is enough to assume that this holds for* $n = 3$ *(corresponding to triples* x_1, x_2, x_3 *and* c_1, c_2, c_3 *with* $c_1 + c_2 + c_3 = 0$*).*

Proof. First, let us agree that a triple (α, β, γ) of non-negative numbers are *squared distances* if $\sqrt{\alpha} \leq \sqrt{\beta} + \sqrt{\delta}$, $\sqrt{\beta} \leq \sqrt{\delta} + \sqrt{\alpha}$ and $\sqrt{\delta} \leq \sqrt{\alpha} + \sqrt{\beta}$. Assuming that $\gamma = \max\{\alpha, \beta, \gamma\}$ this amounts to the condition $\sqrt{\gamma} \leq \sqrt{\alpha} + \sqrt{\beta}$ which, after some simple algebra can be transformed to the equivalent inequality

$$\alpha^2 + \beta^2 + \gamma^2 \leq 2\alpha\beta + 2\beta\gamma + 2\gamma\alpha\,. \tag{18}$$

Assume that (14) holds for triples. Let x, y and z be elements of X and put $\alpha = K(x,y)$, $\beta = K(y,z)$ and $\gamma = K(z,x)$. We realize that what we have to prove is that (α, β, γ) are squared distances. To prove this we exploit (14) with $c_1 = 1$, $c_2 = t$, $c_3 = -t-1$ where t is a real parameter and with $x_1 = x$, $x_2 = y$, $x_3 = z$. This gives us the inequality

$$\alpha t + \beta t(-t-1) + \gamma(-t-1) \leq 0\,.$$

It is then a matter of simple algebra to deduce (18) from the non-positivity of this second order polynomial. $\square$

3 The V/JSD-diagrams

Let A and B be two quantities which have our interest in a given context and assume that A and B both depend on one or more parameters, say $A =$

$A(\alpha)$, $B = B(\alpha)$, possibly with $\alpha = (\alpha_{\mathbf{1}}, \cdots, \alpha_\nu)$ running over some multi-dimensional parameter set. By the *A/B-diagram* we understand the range of the map $\alpha \curvearrowright (\mathbf{A}(\alpha), B(\alpha))$. In [10] and [8] examples and techniques for the determination of such diagrams can be found. It turns out that when the A/B-diagram can be determined this often leads to more precise inequalities relating A and B than one would obtain by a more direct approach.

In this section we shall determine the V/JSD-diagrams Δ_n for $n = (1), 2, 3, \cdots$, where

$$\Delta_n = \{(V(P,Q), \mathrm{JSD}(P,Q)) | (P,Q) \in M_+^1(n) \times M_+^1(n)\} . \tag{19}$$

In Section x we shall determine more general information diagrams. This will be done by applying the same methods as here, but involving more complicated computations. It is believed that the clarification gained by dealing first with a special case justifies the slight duplication.

Returning to (19), we first note that Δ_1 is trivial ($= \{(0,0)\}$). In the following we assume that $n \geq 2$. The technique we shall apply is a direct variational analysis, which is simpler than the variational principles which were applied in the sources quoted.

We characterize Δ_n by its *upper*- and *lower functions* given, for $V \in [0,2]$, by

$$U_n(V) = \sup\{\mathrm{JSD}(P,Q) | P, Q \in M_+^1(n), V(P,Q) = V\} , \tag{20}$$
$$L_n(V) = \inf\{\mathrm{JSD}(P,Q) | P, Q \in M_+^1(n), V(P,Q) = V\} . \tag{21}$$

Clearly, $U_n(0) = L_n(0) = 0$ and $U_n(2) = L_n(2) = 2 \log 2$. So we now assume that $0 < V < 2$.

By a natural compactness argument, it is seen that the extrema in (20) and (21) are attained. By the variational argument we have in mind, we shall determine at the same time the upper as well at the lower function. The point of departure is an *extremal pair* (P,Q) by which we mean that $P, Q \in M_+^1(n)$, that $V(P,Q) = V$ and that either $\mathrm{JSD}(P,Q) = U_n(V)$ or else $\mathrm{JSD}(P,Q) = L_n(V)$ hold.

The point probabilities in P and Q are denoted as follows: $P = (p_1, \cdots, p_n)$, $Q = (q_1, \cdots, q_n)$. We may and do assume that the number of *zero-columns*, i.e. the number of $i \leq n$ such that $\binom{p_i}{q_i} = \binom{0}{0}$ is maximal. In other words, no other pair (P', Q') with $V(P', Q') = V$ and $\mathrm{JSD}(P', Q') = \mathrm{JSD}(P, Q)$ has a higher number of zero-columns.

For the variational arguments to be used below, we will introduce a parameter ε and consider distributions $P_\varepsilon, Q_\varepsilon$ with point probabilities $p_i(\varepsilon)$ and

$q_i(\varepsilon)$. From (10) one then finds that

$$\frac{d}{d\varepsilon}\,\mathrm{JSD}(P_\varepsilon, Q_\varepsilon) = \sum_{i=1}^{n} (\log p_i \cdot p_i' + \log q_i \cdot q_i' - \log(p_i + q_i) \cdot (p_i' + q_i'))\,. \quad (22)$$

Here, p_i, p_i', q_i and q_i' are short for $p_i(\varepsilon)$, $\frac{d}{d\varepsilon}p_i(\varepsilon), q_i(\varepsilon)$ and $\frac{d}{d\varepsilon}q_i(\varepsilon)$,

1.st observation No non-zero column is proportional to another column.

In more detail, if $\binom{p_j}{q_j} \neq \binom{0}{0}$ then for no $i \neq j$ and no scalar α is $\binom{p_j}{q_j} = \alpha\binom{p_i}{q_i}$. To see this, assume the contrary and replace the two columns concerned by the columns $(1+\alpha)\binom{p_i}{q_i}$ and $\binom{0}{0}$ to obtain another extremal pair (P', Q') with one more zero column than for (P, Q) and with V and JSD unchanged. This contradiction gives the result.

2.nd observation For no $i \neq j$ can both $p_i > q_i$ and $p_j > q_j$ hold.

To see this, consider $i \neq j$ and assume that $p_i > q_i$ and $p_j > q_j$ hold. Then, for $|\varepsilon|$ sufficiently small we consider $(P_\varepsilon, Q_\varepsilon)$ defined by $Q_\varepsilon = Q$ and by changing p_i to $p_i + \varepsilon$ and p_j to $p_j - \varepsilon$. As $V(P_\varepsilon, Q_\varepsilon) = V$ for $|\varepsilon|$ sufficiently small, we must have $\frac{d}{d\varepsilon}\,\mathrm{JSD}(P_\varepsilon, Q_\varepsilon) = 0$ for $\varepsilon = 0$. By (22) this gives the equation

$$(\log p_i - \log(p_i + q_i)) + (-\log p_j + \log(p_j + q_j)) = 0\,.$$

It follows that

$$\frac{q_i}{p_i} = \frac{q_j}{p_j}\,,$$

thus $\binom{p_i}{q_i}$ and $\binom{p_j}{q_j}$ are proportional. This contradicts the first observation and we conclude that the second observation is correct.

As P and Q appear symmetrically in our problem, the second observation also shows that for no $i \neq j$ does $q_i > p_i$ and $q_j > p_j$ both hold. It now follows that the number of columns $\binom{p_i}{q_i}$ with $p_i \neq q_i$ is at most 2. As this number cannot be 1 and not either 0 (as $P \neq Q$), it must be exactly 2. We can, therefore, assume that $\binom{p_1}{q_1}$ and $\binom{p_2}{q_2}$ are the two columns in question and further, we may assume that $p_1 > q_1$ and $p_2 < q_2$. By the first observation, there can be at most one column $\binom{p_i}{q_i}$ with $p_i = q_i > 0$. We have now shown that P and Q must have the following form, assuming that the zero-columns come last:

$$\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} p_1, & p_2, & p_3, & 0, \cdots, 0 \\ q_1, & q_2, & p_3, & 0, \cdots, 0 \end{pmatrix}\,.$$

It is convenient with one more general observation:

3.rd observation If, for $i \neq j$, all numbers p_i, q_i, p_j and q_j lie in the open interval $]0,1[$, then $p_i p_j = q_i q_j$.

To prove this, consider P_ε, Q_ε with $|\varepsilon|$ sufficiently small and with point probabilities as for P and Q except for the following 4 changes:

$$p_i(\varepsilon) = p_i + \varepsilon\,,\; p_j(\varepsilon) = p_j - \varepsilon\,,\; q_i(\varepsilon) = q_i + \varepsilon\,,\; q_j(\varepsilon) = q_j - \varepsilon\,.$$

Again, as for the second observation, the total variation is unchanged and $\frac{d}{d\varepsilon}\,\mathrm{JSD}(P_\varepsilon, Q_\varepsilon)$ must vanish for $\varepsilon = 0$, hence, by (22),

$$(\log p_i + \log q_i - 2\log(p_i + q_i)) + (-\log p_j - \log q_j + 2\log(p_j + q_j)) = 0\,.$$

From this we conclude that

$$(1 + x_i)(1 + \frac{1}{x_i}) = (1 + x_j)(1 + \frac{1}{x_j})$$

where $x_i = \frac{p_i}{q_i}$, $x_j = \frac{p_j}{q_j}$. As $x_i = x_j$ is not feasible in view of observation 1, we see that $x_i = \frac{1}{x_j}$ which gives the desired conclusion.

It is now easy to conclude the analysis. In case $p_\varepsilon = 0$, P and Q must be of the form

$$\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} p & , & 1-p & , & 0\cdots 0 \\ p - \frac{V}{2} & , & 1 - p + \frac{V}{2} & , & 0\cdots 0 \end{pmatrix}$$

with $\frac{V}{2} \leq p \leq 1$. If $\frac{V}{2} < p < 1$ we conclude from the third observation that $p(1-p) = \left(p - \frac{V}{2}\right)\left(1 - p + \frac{V}{2}\right)$ or

$$\frac{1}{4} - \left(p - \frac{1}{2}\right)^2 = \frac{1}{4} - \left(p - \frac{V}{2} - \frac{1}{2}\right)^2\,,$$

hence $|p - \frac{1}{2}| = |p - \frac{1}{2} - \frac{V}{2}|$ and we see that $p = \frac{1}{2} + \frac{V}{4}$. Together with the two "end point possibilities" $\left(p = \frac{V}{2} \text{ or } p = 1\right)$ this gives three possibilities:

$$\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \frac{1}{2} + \frac{V}{4}, & \frac{1}{2} - \frac{V}{4}, & 0\cdots 0 \\ \frac{1}{2} - \frac{V}{4}, & \frac{1}{2} + \frac{V}{4}, & 0\cdots 0 \end{pmatrix} \tag{23}$$

$$\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \frac{V}{2}, & 1 - \frac{V}{2}, & 0, \cdots, 0 \\ 0, & 1, & 0\cdots 0 \end{pmatrix} \tag{24}$$

and then a further possibility (corresponding to $p = 1$) which is, however, symmetric to (24) and may, therefore, be discarded.

Finally, we must consider the possibility that $p_3 > 0$. Note that this requires $n \geq 3$. If, in this case $q_1 > 0$, then the third observation applies to column 1 and 3 and we conclude that $p_1 p_3 = q_1 p_3$, i.e. $p_1 = q_1$. However, in view of the first observation, this is not possible. Therefore, $q_1 = 0$ must hold. Similarly, $p_2 = 0$ must hold. This leaves us with P and Q of the form

$$\begin{pmatrix} P \\ Q \end{pmatrix} = \begin{pmatrix} \frac{V}{2}, & 0, & 1-\frac{V}{2} \\ 0, & \frac{V}{2}, & 1-\frac{V}{2} \end{pmatrix} . \tag{25}$$

It remains to calculate the Jensen-Shannon divergence in the three extremal cases isolated:

$$\text{Case (22):} \qquad \mathrm{JSD}(P,Q) = \frac{1}{2}\,\overline{(1+\frac{V}{2})} + \frac{1}{2}\,\overline{(1-\frac{V}{2})} \tag{26}$$

$$\text{Case (23):} \qquad \mathrm{JSD}(P,Q) = \log 2 - \frac{1}{2}\,\overline{(2-\frac{V}{2})} + \frac{1}{2}\,\overline{(1-\frac{V}{2})} \tag{27}$$

$$\text{Case (24):} \qquad \mathrm{JSD}(P,Q) = \frac{1}{2}\,\log 2 \cdot V\,. \tag{28}$$

The functions in (26)-(28) are related as follows:

$$\frac{1}{2}\,\overline{(1+\frac{V}{2})} + \frac{1}{2}\,\overline{(1-\frac{V}{2})} < \log 2 - \frac{1}{2}\,\overline{(2-\frac{V}{2})} + \frac{1}{2}\,\overline{(1-\frac{V}{2})} < \frac{1}{2}\,\log 2 \cdot V\,. \tag{29}$$

This follows by standard considerations. Of course, for $V = 0$ or $V = 2$, equality holds in (29). Recalling that (28) requires $n \geq 3$ we can now conclude as follows:

Theorem 5. *For all $n \geq 2$, the lower function L_n coincides with the common lower function L given by*

$$L(V) = \frac{1}{2}\,(1+\frac{V}{2})\log(1+\frac{V}{2}) + \frac{1}{2}\,(1-\frac{V}{2})\log(1-\frac{V}{2})\,. \tag{30}$$

For all $n \geq 3$, the upper function U_n coincides with the common upper function U given by

$$U(V) = \frac{1}{2}\,\log 2 \cdot V \tag{31}$$

and for $n = 2$, one has

$$U_2(V) = \log 2 - \frac{1}{2}\,(2-\frac{V}{2})\log(2-\frac{V}{2}) + \frac{1}{2}\,(1-\frac{V}{2})\log(1-\frac{V}{2})\,. \tag{32}$$

With just a little more effort, we can determine the exact form of the V/JSD-diagrams Δ_n.

Theorem 6. *For $n = 2$, the V/JSD-diagram Δ_n is the compact region in the plane determined by the Jordan curve composed by the two curves given by (31), run from $V = 0$ to $V = 2$, followed by (30), run from $V = 2$ to $V = 1$.*

For $n \geq 3$, the diagrams Δ_n coincide with the common V/JSD-diagram constructed as above but with the curve given by (32) as the upper curve.

Proof. Assume first that $n \geq 3$. By Theorem 5 we know that Δ_n is contained in the compact domain described. That there are no "holes" in Δ_n follows by considering, for fixed $0 < V < 2$, P_ε, Q_ε given by

$$\begin{pmatrix} P_\varepsilon \\ Q_\varepsilon \end{pmatrix} = \begin{pmatrix} \frac{1}{2} + \frac{V}{4} - \varepsilon, & \frac{1}{2} - \frac{V}{4} - \varepsilon, & 2\varepsilon, & 0, \cdots, 0 \\ \frac{1}{2} - \frac{V}{4} - \varepsilon, & \frac{1}{2} + \frac{V}{4} - \varepsilon, & 2\varepsilon, & 0, \cdots, 0 \end{pmatrix}$$

for $0 \leq \varepsilon \leq \frac{1}{2} - \frac{V}{4}$. Without changing the total variation, this leads in a continuous way from the appropriate point on the graph of the lower function to the corresponding point on the upper function. The desired result follows.

The case $n = 2$ is handled in a similar way. □

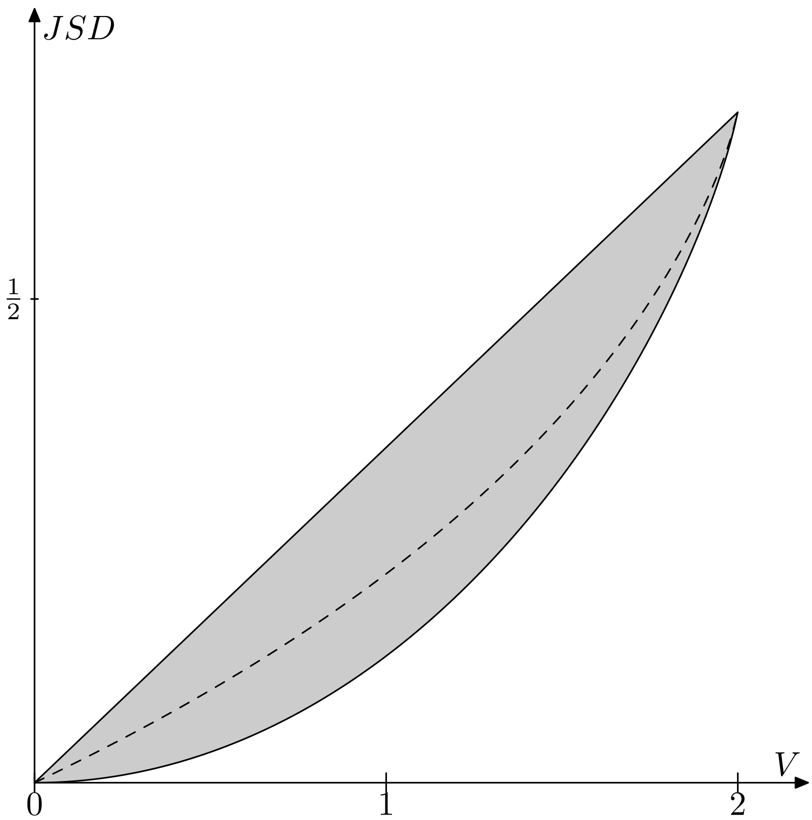

Fig. 1. The V/JSD-diagram.

In Figure 1 we have depicted the V/JSD-diagram for $n \geq 3$ (the shaded region) and for $n = 2$ (the region obtained by replacing the upper bounding curve by the dotted curve).

By expansion of $L(V)$ given by (30) one obtains the inequality

$$\mathrm{JSD}(P, Q) \geq \sum_{n=1}^{\infty} \frac{1}{2n(2n-1)} \left(\frac{V}{2}\right)^{2n}, \tag{33}$$

which was proved by a different method in [23, Theorem 5].

4 Norm-based measures of variation

We shall embed Jensen-Shannon divergence in a two-parameter family of discrimination measures. This will build on and expand related studies of one-parameter families as detailed in several publications of which we only point to (obs: udvides!) [2] and to [19] where references to the literature can be found. Instead of referring to f-divergences as defined by Csiszár [5], and by Ali and Silvey [1] as a guide to the sought measures of discrimination we shall take another route and base the definitions on considerations of classical p-norms.

Discrimination measures as those we shall define are of significance for hypothesis testing and for feature selection problems in pattern recognition. We choose a terminology taken from the latter field. As point of departure we take a *feature space* $(X, \mathcal{A}, \mu)$ which is nothing but an abstract probability space. Elements in X are called *features* and μ is the *(Bayesian) prior*. To satisfy technical requirements, we have also shown $\mathcal{A}$, the σ-field serving as domain of definition for μ. By a *feature function* we shall understand a measurable and non-negative function defined on A. To avoid technicalities, the reader may assume that feature functions are bounded.

Let f be a feature function and consider the *p-norm* of f defined as is classical by

$$\|f\|_p = \left(\int_B f(x)^p d\mu(x) \right)^{\frac{1}{p}} . \tag{34}$$

Here we allow any extended real value for the parameter p. Regarding this definition, we follow standard conventions as can be found e.g. in [9] by Hardy, Littlewood and Pólya. There, one will also find standard properties. In particular, the following result is pointed out:

Lemma 1. *Let f be a feature function and denote by $f_{\min}$ the essential minimum and by $f_{\max}$ the essential maximum of f. Then the function ϕ defined on $\mathbb{R}$ (or even on the extended real line) by*

$$\phi(p) = \|f\|_{1/p} \tag{35}$$

is log-convex and decreasing on $\mathbb{R}_+$ and log-concave and, likewise, decreasing on $\mathbb{R}_-$. Furthermore, $\phi(0^-) = f_{\min}$, $\phi(-1) = H$, the harmonic mean of f, $\phi(-\infty) = \phi(+\infty) = G$, the geometric mean of f, $\phi(1) = A$, the arithmetic mean of f, and $\phi(0^+) = f_{\max}$.

As to the log-convexity/concavity assertion we point to [9, Theorems 17, 18, 87 and 197]. The result goes back to Lyapunov, [16], and the inequality expressing this fact (in one form or another) is called *Lyapunov's inequality*, cf. Mitrinović [18]. In [9, Theorem 197], the log-convexity of $\psi : p \curvearrowright \|f\|_p^p$ is asserted. As ϕ and ψ are duals of each other in the sense that $\psi(p) = \phi(1/p)^p$, it is easy to check that if one of these functions is log-convex in $\mathbb{R}_+$, then so is the other. For instance, one may exploit a notion of dual convex combinations (combinations $m = \alpha x + \beta y$, $m^* = \alpha^* x^* + \beta^* y^*$ with $xx^* = yy^* = mm^* = 1$). A recent reference to the result is [11, Corollary 1.2.9]. For the convenience of the reader we briefly indicate a direct proof: Assume say, that p and q are positive. Then note that for a convex combination $m = \alpha p + \beta q$, the inequality $(\log \phi)(m) \le \alpha(\log \phi)(p) + \beta(\log \phi)(q)$m is equlivalent with the inequality

$$\int f^{1/m} \le \left(\int (f^{\frac{1}{ps}})^s\right)^{\frac{1}{s}} \left(\int (f^{\frac{1}{qt}})^t\right)^{\frac{1}{t}} \tag{36}$$

with $s = (\alpha p + \beta q)/(\alpha p)$ and $t = (\alpha p + \beta q)/(\beta q)$. Clearly, (36) follows from Hölder's inequality.

Corollary 7. *The function ϕ defined by* (35) *is convex and decreasing in* $\mathbb{R}_+$.

This follows as a log-convex function is convex. In passing we note that a log-concave function need not be concave and, in fact, the function ϕ is, typically, *not* concave for negative arguments close to zero.

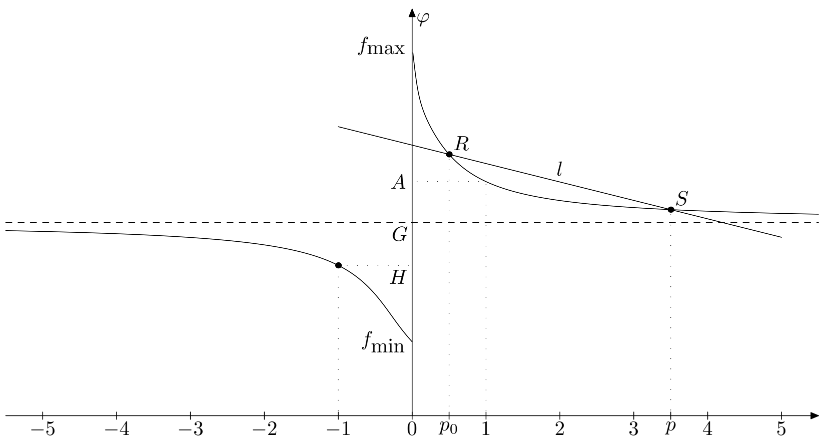

Fig. 2. The function $\phi :\ p \curvearrowright \|f\|_{1/p}$.

Figure 2 shows the graph of a typical ϕ-function. It mirrors the variation of f as f varies from $f_{\min}$ to $f_{\max}$. One can take $f_{\max} - f_{\min} = \phi(0^+) - \phi(0^-)$ as a measure for the variation of f. The idea now is to define other measures of variation by fixing two arguments of ϕ and considering the difference. More precisely, for β and β_0 two positive numbers we define the $(\beta|\beta_0)$-*variation breadth* of f as the absolute value of the slope of the chord on ϕ corresponding to the two arguments $p = 1/\beta$ and $p_0 = 1/\beta_0$ (the chord in question is shown in Figure 2 as the line RS). Denoting the resulting quantity by $\mathrm{VB}_{\beta|\beta_0}$, we therefore have:

$$\mathrm{VB}_{\beta|\beta_0}(f) = -\frac{\phi(p)-\phi(p_0)}{p-p_0} = \beta\,\beta_0\frac{\|f\|_\beta - \|f\|_{\beta_0}}{\beta-\beta_0}. \tag{37}$$

If need be, we write $\mathrm{VB}_{\beta|\beta_0}(f|\mu)$ to stress the dependence on the probability measure μ.

Thus, variation breadth is proportional to $|\phi(p)-\phi(p_0)| = |\|f\|_\beta - \|f\|_{\beta_0}|$ with a factor of proportionality chosen appropriately so that the family of measures introduced varies in a natural and smooth way with the parameters.

In (37) we allow $\beta = \beta_0$ with the understanding that then $\mathrm{VB}_{\beta_0|\beta_0}(f) = -\phi'(p_0)$, corresponding to the slope of the tangent to ϕ in $(p_0, \phi(p_0))$. From the second expression in (37) we also see that $\mathrm{VB}_{\beta_0|\beta_0}(f) = \beta_0^2\,\psi'(\beta_0)$ where ψ is the function $\beta \curvearrowright \|f\|_\beta$. A direct calculation shows that

$$\mathrm{VB}_{\beta|\beta}(f) = \|f\|_\beta^{1-\beta}\left(\int f^\beta(x)\log f^\beta(x)\,d\mu(x) - \|f\|_\beta^\beta \log\|f\|_\beta^\beta\right) \tag{38}$$

$$= \beta\|f\|_\beta^{1-\beta}\int f(x)^\beta\,\log\frac{f(x)}{\|f\|_\beta}\,d\mu(x) \tag{39}$$

$$= \|f\|_\beta\int\left(\frac{f(x)}{\|f\|_\beta}\right)^\beta\log\left(\frac{f(x)}{\|f\|_\beta}\right)^\beta\,d\mu(x). \tag{40}$$

The important special case $\beta = 1$ gives the expression

$$\mathrm{VB}_{1|1}(f) = \int f(x)\log\frac{f(x)}{\|f\|_1}d\mu(x). \tag{41}$$

The formulas above may also be expressed using the function $x \curvearrowright \overline{x}$ introduced in (9). This may be convenient if one takes (40) as a definition. Then, from the elementary inequality $\overline{x} \geq x - 1$ one immediately deduces the fact that $\mathrm{VB}_{\beta|\beta}(f) \geq 0$ and also one sees that the condition for equality is that f is constant (a.e.).

The facts just pointed out may also be deduced by referring to standard properties of the usual (Kullback-Leibler) divergence. Indeed, we realize that

$$\mathrm{VB}_{\beta|\beta}(f|\mu) = \|f\|_\beta\,\mathrm{D}(\mu_{f,\beta}\|\mu) \tag{42}$$

where the probability measure $\mu_{f,\beta}$ is defined by normalization of the measure with density f^β w.r.t. μ, i.e., for $A \subseteq X$, we put

$$\mu_{f,\beta}(A) = \int_A \left(\frac{f}{\|f\|_\beta}\right)^\beta d\mu. \tag{43}$$

Qualitatively speaking, it is not surprising that variation breadth and divergence are related in this way since what we express is that large divergence between $\mu_{f,\beta}$ and μ corresponds to large deviation for the density function for the former measure when compared to the density of the latter measure (which is the constant function 1). We may exploit the relation (42) by appealing to *Pinskers inequality*, cf. [4], to conclude that

$$\mathrm{VB}_{\beta|\beta}(f) \geq \frac{1}{2}\|f\|_\beta \left(\int \left|\left(\frac{f}{\|f\|_\beta}\right)^\beta - 1\right| d\mu\right)^2, \tag{44}$$

an inequality which also sheds some light on the significance of variation breadth. In particular, it shows that variation breadth vanishes if and only if f is constant a.e. – this observation for parameters β, β_0 with $\beta \neq \beta_0$ is easily deduced from the case of equal parameter values (by definition (37) and by Corollary 7).

Regarding special parameter values, we note that (with the natural definition), $\mathrm{VB}_{0|0}(f) = 0$. In the other exceptional case, $\beta = \infty$, we simplify and assume that μ is discrete with finite support. Then it is easy to see that

$$\mathrm{VB}_{\infty|\infty}(f) = -f_{\max} \log \mu(f = f_{\max}). \tag{45}$$

One may apply the definitions in cases where the value $+\infty$ is obtained, but of course such cases will often be undesirable exceptional cases. And (42) may be meaningless as the measures (43) may not make sense.

Regarding the more precise technical conditions for $\mathrm{VB}_{\beta|\beta}(f)$ to be finite, note that the involved calculations depend on a differentiation under the integral sign. Recalling the standard conditions for the validity of this operation we realize that what we need is the condition that $f^\beta \log f$ be integrable (in (41) with $\beta = 1$). Strictly speaking one should then only base the definition on one-sided limits (i.e. on half-tangents). This can be avoided if one assumes the integrability of $f^{\beta+}$ by which we mean that f^γ is integrable for some $\gamma > \beta$. Of course, if, as we urged the reader to assume, f is bounded, all technical difficulties disappear.

One may also define variation breadth for negative values of the parameters. One could base the definition on the function $\phi : p \curvearrowright \|f\|_{1/p}$ for $p < 0$

but this is slightly awquard in view of the non-concavity of ϕ for negative p close to 0. Instead we suggest a definition based on the formula

$$\mathrm{VB}_{\beta|\beta_0}(f) = \mathrm{VB}_{-\beta|-\beta_0}(\tilde{f}) \tag{46}$$

for negative values of β and β_0 and with $\tilde{f}$ defined by $\tilde{f}(x) = 1/f(x)$. For fixed f and $\beta_0 < 0$, $\mathrm{VB}_{\beta|\beta_0}(f)$ is then decreasing in β for $\beta < 0$. We shall not comment much on variation breadth for negative parameter values.

Let us collect some basic properties of variation breadth:

Proposition 8. *Variation breadth as defined by* (35) *and its limiting cases has the following general properties*

(i) $\mathrm{VB}_{\beta|\beta_0} = \mathrm{VB}_{\beta_0|\beta}$,

(ii) $\mathrm{VB}_{\beta|\beta_0}(c\,f) = c\,\mathrm{VB}_{\beta|\beta_0}(f)$,

(iii) $(\frac{1}{\beta_0} - \frac{1}{\beta_2})\,\mathrm{VB}_{\beta_2|\beta_0} = (\frac{1}{\beta_1} - \frac{1}{\beta_2})\,\mathrm{VB}_{\beta_2|\beta_1} + (\frac{1}{\beta_0} - \frac{1}{\beta_1})\,\mathrm{VB}_{\beta_1|\beta_0}$,

(iv) $\beta \curvearrowright \mathrm{VB}_{\beta|\beta_0}(f)$ *is increasing in* $\mathbb{R}_+$ *and varies from* 0 *to* $\beta_0\,(f_{\max} - \|f\|_{\beta_0})$*,*

(v) $\mathrm{VB}_{\beta|\beta_0}(f) = 0$ *if and only if* f *is constant a.e.*

We only have non-negative parameter values in mind for this result and functions occuring are assumed to satisfy suitable boundedness conditions (or the associated VB_{-} values are allowed to be infinite). In (ii), c denotes a positive constant. The proof follows from the preceeding discussion. In particular, (iv) follows from Lemma 1 and the proof of (v) for $\beta \neq \beta_0$ is readily reduced to the case $\beta = \beta_0$ which we already discussed.

A further property concerns behaviour of variation breadth under datareduction:

Theorem 9. *Let* $\partial\mathcal{A}$ *be a sub* σ*-algebra of* $\mathcal{A}$ *and denote by* ∂f *the conditional expectation of* f *given* $\partial\mathcal{A}$*. Then the datareduction inequality*

$$\mathrm{VB}_{\beta|\beta_0}(\partial f) \leq \mathrm{VB}_{\beta|\beta_0}(f) \tag{47}$$

holds for every pair (β, β_0) *with* $\beta_0 \leq 1 \leq \beta$ *(or* $\beta \leq 1 \leq \beta_0$*).*

Proof. Assume that $\beta_0 \leq 1 \leq \beta$. Then $\|\partial f\|_\beta \leq \|f\|_\beta$ and $\|\partial f\|_{\beta_0} \geq \|f\|_{\beta_0}$ (these well known facts follow by convexity considerations). The result for $\beta_0 < \beta$ then follows from the definition of variation breadth. If $\beta = 1$, the result may be obtained by a limit consideration (or by applying a well known datareduction inequality for divergence). □

Possibly, a more general and also natural result holds by letting the datareduction operation depend on the parameter values, but as stated, (47) fails without the restriction regarding the parameter values (it is easy to find examples with μ a uniform distribution on a three-element set and with $\beta = \beta_0 > 1$ for which (47) fails).

5 Integrated variation breadth

The notion of variation breadth may be extended to situations where, in addition to the feature space $(X, \mathcal{A}, \mu)$ we have also given a σ-finite measure space $(Y, \mathcal{B}, \eta)$. This could be interpretated as a sample space or, as we prefer here, as a space of *loci*. An individual element $y \in Y$ is referred to as a *locus*. If $f : (x, y) \curvearrowright f(y|x)$ is a non-negative bi-measurable function on the product space $X \times Y$, we introduce an integrated version of variation breadth by simply considering the quantity

$$\mathrm{VB}_{\beta|\beta_0}(f) = \int \mathrm{VB}_{\beta|\beta_0}(f(y|\cdot))\, d\eta(y). \tag{48}$$

When need be we speak of *integrated variation breadth* but there seems to be no need for introduction of a new notation.

In particular, we consider the case of a *Markov kernel* on $X \times Y$. Then, for each $x \in X$, there is given a probability measure P_x on Y which we assume is absolutely continuous w.r.t. the measure η. Denote by $y \curvearrowright f(y|x)$ the corresponding density and assume that the versions chosen render f bi-measurable. Then the notion of integrated variation breadth applies. In this case, we denote by $\overline{P}$ the natural mixture of the P_x's, i.e., for a measurable subset of Y, we put

$$\overline{P}(A) = \int P_x(A) d\mu(x). \tag{49}$$

In short we may write $\overline{P} = \int P_x d\mu(x)$. Clearly, $\overline{f}$ given by

$$\overline{f}(y) = \int f(y|x) d\mu(x) \tag{50}$$

is the density of $\overline{P}$ w.r.t. η.

From (41) we readily see that the special case of variation breadth with parameters $\beta = \beta_0 = 1$ is given by

$$\mathrm{VB}_{1|1}(f) = \int_Y \int_X f(y|x) \log \frac{f(y|x)}{\overline{f}(y)} d\mu(x) d\eta(y) \tag{51}$$

$$= \int_X \int_Y f(y|x) \log \frac{f(y|x)}{\overline{f}(y)} d\eta(y) d\mu(x). \tag{52}$$

Comparing (52) with (4) we realize that the integrated variation breadth for Markov kernels is a natural extension of the general Jensen-Shannon divergence. We have thus achieved the announced embedding of Jensen-Shannon divergence into a natural two-parameter family of divergence measures. We shall write, somewhat incorrectly (but in line with the notation (4)), $\mathrm{VB}_{\beta|\beta_0}(\int P_x d\mu)$ for $\mathrm{VB}_{\beta|\beta_0}(f)$ in case f is defined via a Markov kernel as above. We write JSD for $\mathrm{VB}_{1|1}$. Let us state explicitly the version of (52) in the form which generalizes (4):

$$\mathrm{JSD}\left(\int P_x \, d\mu(x)\right) = \int \mathrm{D}(P_x \| \overline{P}) \, d\mu(x). \tag{53}$$

This follows from (52) as

$$\mathrm{D}(P_x \| \overline{P}) = \int_Y f(y|x) \log \frac{f(y|x)}{\overline{f}(y)} \, d\eta(y).$$

Basic properties follow from the results of Section 4 and we refrain from a more systematic study. Instead, we shall turn in Section 7 to the special Jensen-Shannon divergence measure. As a preparation, which is of interest in its own right, we first look into a case which at least in principle is exceptionally simple.

6 Discrimination in $\mathbb{R}_+$

In this section we consider variation breadth of the simple non-integrated type – i.e. Y is a one-element set – and we further simplify by taking as $(X, \mathcal{A}, \mu)$ the uniform distribution on a two-element set. Really, what variation breadth then amounts to is a measure of discrimination between two non-negative real numbers. Not to overload the notation we still stick to the VB_notation, but refer to it as *discrimination*. We denote by (x, y) the function (vector) assuming the two values x and y and by $\|(x, y)\|_\beta$ the β-norm of this function w.r.t. the uniform measure with weights $\frac{1}{2}, \frac{1}{2}$.

From Section 4 we inherit a number of notions and results. For instance, we have

$$\mathrm{VB}_{\beta|\beta_0}(x, y) = -\frac{\|(x, y)\|_\beta - \|(x, y)\|_{\beta_0}}{\frac{1}{\beta} - \frac{1}{\beta_0}} \tag{54}$$

$$= \frac{\beta\,\beta_0}{\beta - \beta_0} \left((\frac{1}{2}x^\beta + \frac{1}{2}y^\beta)^{\frac{1}{\beta}} - (\frac{1}{2}x^{\beta_0} + \frac{1}{2}y^{\beta_0})^{\frac{1}{\beta_0}} \right), \tag{55}$$

and $\mathrm{VB}_{\beta|\beta}(x,y)$ is given by the expression

$$\|(x,y)\|_\beta^{1-\beta}\left(\frac{x^\beta}{2}\log x^\beta+\frac{y^\beta}{2}\log y^\beta-\frac{x^\beta+y^\beta}{2}\log(\frac{1}{2}x^\beta+\frac{1}{2}y^\beta)\right) \tag{56}$$

or, say, by the formula

$$\mathrm{VB}_{\beta|\beta}(x,y)=(\frac{1}{2}x^\beta+\frac{1}{2}y^\beta)^{\frac{1}{\beta}}\,\mathrm{D}\left((\frac{x^\beta}{x^\beta+y^\beta},\frac{y^\beta}{x^\beta+y^\beta})\|(\frac{1}{2},\frac{1}{2})\right). \tag{57}$$

Special values are given by

$$\mathrm{VB}_{\beta|\beta_0}(0,1)=\frac{\beta\,\beta_0}{\beta-\beta_0}\left(2^{-\frac{1}{\beta}}-2^{-\frac{1}{\beta_0}}\right) \tag{58}$$

$$\mathrm{VB}_{\beta|\beta}(0,1)=2^{-\frac{1}{\beta}}\log 2. \tag{59}$$

For $\beta_0=1$ we point to the special cases

$$\mathrm{VB}_{\infty|1}(x,y)=\frac{1}{2}|x-y|\,, \tag{60}$$

$$\mathrm{VB}_{2|1}(x,y)=\left(\sqrt{2x^2+2y^2}-(x+y)\right)\,, \tag{61}$$

$$\mathrm{VB}_{1|1}(x,y)=\frac{1}{2}x\log\frac{2x}{x+y}+\frac{1}{2}y\log\frac{2y}{x+y}\,, \tag{62}$$

$$\mathrm{VB}_{\frac{1}{2}|1}(x,y)=\frac{1}{4}\left(\sqrt{x}-\sqrt{y}\right)^2\,. \tag{63}$$

which we refer to, respectively, as *discrimination in* $\mathbb{R}_+$ *by total variation*, *Österreicher discrimination in* $\mathbb{R}_+$, *Jensen-Shannon discrimination in* $\mathbb{R}_+$ and *Hellinger discrimination in* $\mathbb{R}_+$. Note that for fixed β_0 and fixed (x,y), $\mathrm{VB}_{\beta|\beta_0}$ is increasing in β by Lemma 1.

In the special setting we are dealing with here, we obtain a simple extra structural property, viz. that

$$\mathrm{VB}_{\beta|\beta_0}(\frac{1}{x},\frac{1}{y})=\frac{1}{xy}\,\mathrm{VB}_{\beta|\beta_0}(x,y) \tag{64}$$

as is easily seen. According to (46), the left hand side here defines the kernel $\mathrm{VB}_{-\beta|-\beta_0}$. These kernels we refer to as *inverse kernels* and denote them by adding a superscript:

$$\mathrm{VB}^{inv}_{\beta|\beta_0}(x,y)=\mathrm{VB}_{\beta|\beta_0}(\frac{1}{x},\frac{1}{y}). \tag{65}$$

These kernels are only defined on $]0,\infty[$, not on $\mathbb{R}_+ = [0,\infty[$. The inverse kernels are not bounded and behave differently from the usual "direct" kernels. However, in view of (65) it is evident that for a lot of properties such as negative definiteness or being a squared metric, if one of the kernels $\mathrm{VB}_{\beta|\beta_0}$ or $\mathrm{VB}^{inv}_{\beta|\beta_0}$ has the property, then so does the other.

Note that whereas the direct kernels are homogeneous, the inverse kernels are *inverse homogeneous*, i.e.

$$\mathrm{VB}^{inv}_{\beta|\beta_0}(cx,cy) = \frac{1}{c}\,\mathrm{VB}^{inv}_{\beta|\beta_0}(x,y).$$

Both inverse and direct kernels satisfy (11) - (13) from Section 2.

We now aim at characterizing those parameter values for which $\mathrm{VB}_{\beta|\beta_0}$ is a squared metric. The tools are those developed in Section 2. From the discussion there one easily deduces the following auxiliary result which may also be found in [3, Example 6.5.15].

Lemma 2. *The kernel $(x,y) \curvearrowright (x+y)^\alpha$ is negative definite on $\mathbb{R}_+$ for $0 \le \alpha \le 1$ and the kernel $(x,y) \curvearrowright -(x+y)^\alpha$ is negative definite on $\mathbb{R}_+$ for $1 \le \alpha \le 2$,*

The result may also be expressed by saying that the functions $x \curvearrowright x^\alpha$ for $0 \le \alpha \le 1$ and $x \curvearrowright -x^\alpha$ for $1 \le \alpha \le 2$ are negative definite on $\mathbb{R}_+$.

A metric d on $\mathbb{R}_+$ is *squareroot homogeneous* if the identity $d(cx,cy) = \sqrt{c}d(x,y)$ holds generally. We can now state a main result:

Theorem 10. *For $\frac{1}{2} \le \beta_0 \le 1 \le \beta \le \infty$, the discrimination measures $\mathrm{VB}_{\beta|\beta_0}$ on $\mathbb{R}_+$ are squared metrics. Metrics obtained in this way are bounded, squareroot homogeneous and topologizes the usual topology on $\mathbb{R}_+$. For fixed β_0, the metrics are increasing in β and, for the stated parameter values, the inequalities*

$$\mathrm{VB}_{\frac{1}{2}|1}(x,y) \le \mathrm{VB}_{\beta|\beta_0}(x,y) \le \mathrm{VB}_{\infty|1}(x,y) \tag{66}$$

hold. Furthermore, the metrics in question can be embedded isometrically in Hilbert space.

Proof. Assume that $\frac{1}{2} \le \beta_0 \le 1 \le \beta$ with $\beta_0 < \beta$. The key fact needed is the negative definiteness of $\mathrm{VB}_{\beta|\beta_0}$. To establish this fact, let $(c_i)_{i\le n}$ and $(x_i)_{i\le n}$ be real numbers with $\sum c_i = 0$ and all the x_i non-negative. Put $\xi_i = \frac{1}{2}x_i^\beta$, $\eta_i = \frac{1}{2}x_i^{\beta_0}$. By Lemma 2,

$$\sum_{i,j} c_i c_j (\xi_i + \xi_j)^{\frac{1}{\beta}}$$

and

$$-\sum_{i,j} c_i c_j (\eta_i + \eta_j)^{\frac{1}{\beta_0}}$$

are both non-positive. Therefore, the sum which, apart from a positive factor, equals

$$\sum_{i,j} c_i c_j \, \mathrm{VB}_{\beta|\beta_0}(x_i, x_j)$$

which is non-positive as well. We conclude, as desired, that $\mathrm{VB}_{\beta|\beta_0}$ is negative-definite.

The remaining parts of the proof follow by Theorem 1 and by Corollary 7 in connection with geometric considerations regarding the slope defining $\mathrm{VB}_{\beta|\beta_0}$. It should also be remarked that the limit cases $\beta = \beta_0$ (and $\beta = \infty$) are obtained by a trivial passage to the limit. □

In particular, the kernels (60) - (63) are all squared metrics and so are the associated inverse kernels.

The central property of negative definiteness of course holds for many other kernels on $\mathbb{R}_+$ than those considered here, cf. [3]. Based on the kernels appearing in the theorem, more negative definite kernels can be generated by simple transformations. Indeed, the proof shows that also kernels of the form $(x, y) \curvearrowright \mathrm{VB}_{\beta|\beta_0}((\phi(x), \phi(y)))$ are negative definite. This could lead to the inverse kernels and also implies that one may change the uniform $(\frac{1}{2}, \frac{1}{2})$-distribution to any probability distribution (s, t). By symmetrization one can then obtain new types of kernels which are squared distances. Kernels of this type include mappings of the form $(x, y) \curvearrowright \|(x, y)\|_\beta + \|(y, x)\|_\beta - \|(x, y)\|_{\beta_0} - \|(y, x)\|_{\beta_0}$ with the norms defined w.r.t. the weights (s, t) in place of $(\frac{1}{2}, \frac{1}{2})$. Such kernels will also be homogeneous and the associated metrics squareroot homogeneous.

There are, therefore, many squareroot homogeneous metrics on $\mathbb{R}_+$. One may ask if the metrics found in Theorem 10 (and their extensions to other parameter values) are uniquely characterized in some way.

We shall now investigate cases where $\mathrm{VB}_{\beta|\beta_0}$ is not a squared metric. By Proposition 4 we know one way of expressing this. Let us analyze the problem and consider fixed values of β and β_0 with $0 < \beta_0 < \beta < \infty$. Furthermore, let x, y, z in $\mathbb{R}_+$ be given (to be chosen appropriastely below). Put

$$
\begin{aligned}
A &= \frac{\beta - \beta_0}{\beta\beta_0}\,\mathrm{VB}_{\beta|\beta_0}(x,y) \\
B &= \frac{\beta - \beta_0}{\beta\beta_0}\,\mathrm{VB}_{\beta|\beta_0}(y,z) \\
C &= \frac{\beta - \beta_0}{\beta\beta_0}\,\mathrm{VB}_{\beta|\beta_0}(z,x)\,,
\end{aligned}
$$

i.e. $A = \|(x,y)\|_\beta - \|(x,y)\|_{\beta_0}$ and similarly for B and C. Clearly, $\mathrm{VB}_{\beta|\beta_0}(x,y)$, $\mathrm{VB}_{\beta|\beta_0}(y,z)$ and $\mathrm{VB}_{\beta|\beta_0}(z,x)$ are squared distances if and only if A, B and C are so and this is equivalent to the inequality $F \geq 0$ where we have put

$$
F = (2BC - A^2) + (2AC - B^2) + (2AB - C^2)\,,
$$

cf. the proof of Proposition 4. Now choose $(x,y,z) = (0,1,1+\varepsilon)$ with ε a parameter. Then A, B and C are functions of ε with A a constant function. For $\epsilon = 0$, straight forward differentiation shows that, denoting differentiation by primes, $B^{'} = B^{''} = C^{''} = 0$ and $C = C^{'} = A$. Plugging these values into $F = F(\varepsilon)$ gives

$$
\begin{aligned}
F(0) &= F^{'}(0) = 0 \\
F^{''}(0) &= 4A(0)B^{''}(0) - 2A(0)^2.
\end{aligned}
$$

Therefore, $F^{''}$ has the same sign as $2B^{''}(0) - A(0)$. Clearly, if $F^{''}(0)$ is negative, $F(\varepsilon)$ will be negative for sufficiently small numerical values of ε and $\mathrm{VB}_{\beta|\beta_0}$ cannot be a squared metric. As $B^{''}(0) = \frac{1}{4}(\beta - \beta_0)$ and $A(0) = 2^{-\frac{1}{\beta}} - 2^{-\frac{1}{\beta_0}}$, we see that $F^{''}(0)$ is negative if and only if

$$
\frac{1}{2}(\beta - \beta_0) < 2^{-\frac{1}{\beta}} - 2^{-\frac{1}{\beta_0}}. \tag{67}
$$

At this stage we find it convenient to introduce a simple special function which we believe stears the issue of having or not having the metric property in question. We denote the function by π and call it the *pilot function* (associated with the problem considered). It is defined on $[0,\infty[$ by

$$
\pi(x) = \frac{1}{2}x - 2^{-\frac{1}{x}}. \tag{68}
$$

We can now state an important auxiliary result:

Lemma 3 (pilot function lemma). *Let β and β_0 be distinct positive numbers. If the slope of the chord on the graph of the pilot function* (68) *which connects $(\beta_0, \pi(\beta_0)$ with $(\beta, \pi(\beta)$ is negative, then $\mathrm{VB}_{\beta|\beta_0}$ is not a squared metric.*

Proof. We assume as in the foregoing analysis that $\beta > \beta_0$. The slope in question is

$$\frac{\pi(\beta) - \pi(\beta_0)}{\beta - \beta_0} = \frac{1}{2} - \frac{2^{-\frac{1}{\beta}} - 2^{-\frac{1}{\beta_0}}}{\beta - \beta_0}$$

and we realize, as $\beta > \beta_0$ that the slope is negative if and only if (67) holds. The result then follows by the above analysis. The case when $\beta < \beta_0$ is of course easily discussed based on the case $\beta > \beta_0$. □

We can now formulaste a conjecture, the truth of which will justify the key role played by the pilot function.

Conjecture 6.1. *For arbitrary parameters β and β_0, $\mathrm{VB}_{\beta|\beta_0}$ is a squared metric in $\mathbb{R}_+$ if and only if the slope of the chord connecting $(\beta_0, \pi(\beta_0))$ with $(\beta, \pi(\beta))$ is non-negative, it being understood that the chord is to be replaced by the tangent of the pilot function in the point $(\beta, \pi(\beta))$ in case $\beta_0 = \beta$.*

As evidence for the conjecture we prove the following main result:

Theorem 11. *The conjecture is true if either $\beta_0 = 1$ or $\beta_0 = \frac{1}{2}$. Accordingly, $\mathrm{VB}_{\beta|1}$ is a squared metric if and only if $\beta \geq \frac{1}{2}$ and $\mathrm{VB}_{\beta|\frac{1}{2}}$ is a squared metric if and only if $\beta \geq 1$.*

Proof. This depends on the pilot function lemma and on Theorem 10 in connection with a simple analysis of the pilot function. Instead of carrying out this analysis, we choose to show the result in the form of the graph of the pilot function, depicted in Figure 3. Clearly, a glimpse on the figure reveals the fact that the slope of the chord connecting $(\beta, \pi(\beta))$ with $(1, \pi(1)) = (1, 0)$, respectively with $(\frac{1}{2}, \pi(\frac{1}{2})) = (\frac{1}{2}, 0)$, is non-negative if and only if $\beta \geq \frac{1}{2}$, respectively $\beta \geq 1$. The result follows. □

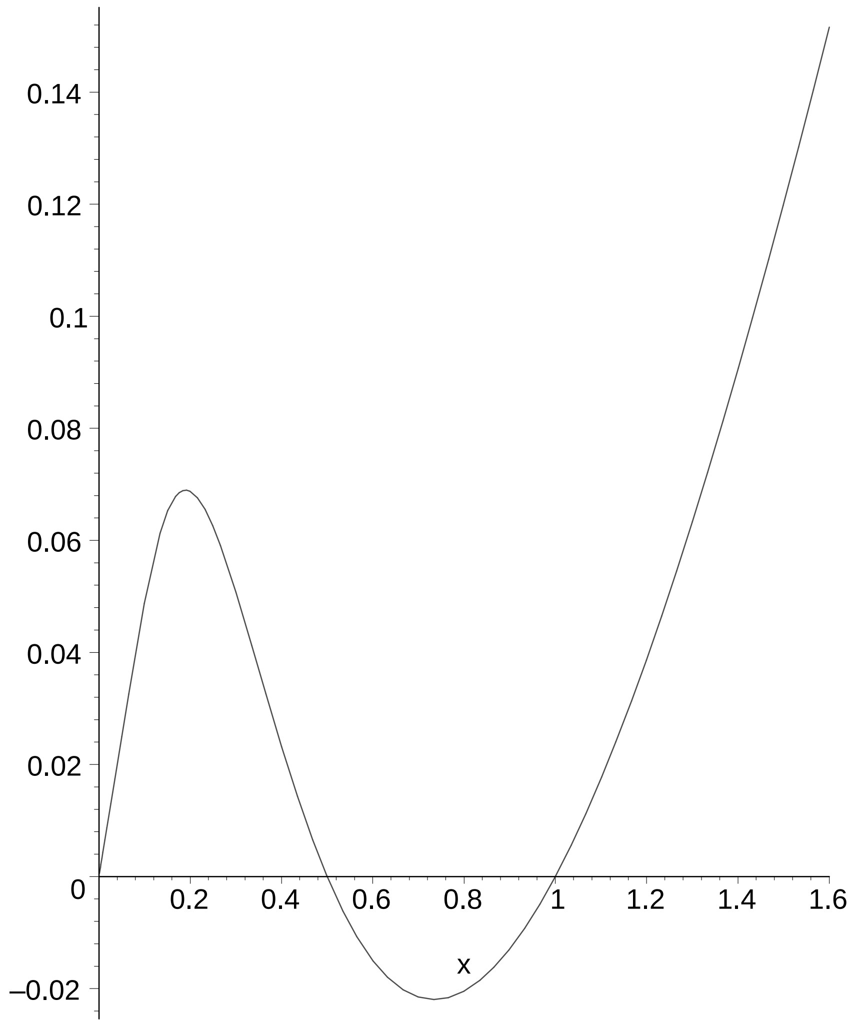

Fig. 3. The pilot function.

7 Norm-based measures of discrimination for distributions

We return to a study of measures which discriminate between just two probability distributions, P and Q. These measures of discrimination are special cases of integrated variation breadth introduced in Section 5.

So, what we consider is the case when the feature space is the two-element set $\{0,1\}$ provided with the uniform distribution and when the space of loci, $(Y, \mathcal{B}, \eta)$, is any σ-finite measure space. We shall study discrimination among probability measures which are absolutely continuous with respect to the measure η. Let P_0 and P_1 be two such distributions and denote by f_0 and f_1

their density functions with respect to η.

Referring to Section 5, we define a two-parameter family of discrete measures, denoted $(P_0, P_1) \curvearrowright \mathrm{VB}_{\beta|\beta_0}(P_0, P_1)$ by considering integrated variation breadth based on the density functions involved, i.e. we put

$$\mathrm{VB}_{\beta|\beta_0}(P_0, P_1) = \mathrm{VB}_{\beta|\beta_0}(f_0, f_1) = \int_Y \mathrm{VB}_{\beta|\beta_0}(f_0(y), f_1(y))d\eta(y) \tag{69}$$

with the integrand given by (54) or (55) or, in case $\beta_0 = \beta$, by (56) or (57). In case $\beta_0 = \beta = 1$, we srite $\mathrm{JSD}(P_0, P_1)$ for $\mathrm{VB}_{1|1}(P_0, P_1)$ and refer to this as the (specific) *Jensen-Shannon divergence measure*. In case $(Y, \mathcal{B}, \eta)$ is a discrete space with η as counting measure, this is in line with the terminology of Section 1.

In the general case, one may refer to the discrimination measure $\mathrm{VB}_{\beta|\beta_0}$ as *discrimination by variation breadth*.

All these measures are potential *f-divergences* in the sense of Csiszár [5] or Ali and Silvey [1]. However, the property of convexity which is essential for many results may not hold. In order to clarify the relation, we define the family $\mathrm{f}_{\beta|\beta_0}$ of functions on $\mathbb{R}_+$ by

$$\mathrm{f}_{\beta|\beta_0}(u) = \frac{\beta\beta_0}{\beta - \beta_0}\left((\frac{1}{2}u^\beta + \frac{1}{2})^{\frac{1}{\beta}} - (\frac{1}{2}u^{\beta_0} + \frac{1}{2})^{\frac{1}{\beta_0}}\right) \tag{70}$$

for $\beta \neq \beta_0$, and by

$$\mathrm{f}_{\beta|\beta}(u) = (\frac{1}{2}u^\beta + \frac{1}{2})^{\frac{1}{\beta}-1}\left(\frac{1}{2}u^\beta \log u^\beta - (\frac{1}{2}u^\beta + \frac{1}{2})\log(\frac{1}{2}u^\beta + \frac{1}{2})\right) \tag{71}$$

when $\beta_0 = \beta$. In all cases then, $\mathrm{f}_{\beta|\beta_0}(u) = \mathrm{VB}_{\beta|\beta_0}(u, 1)$.

With these definitions,

$$\mathrm{VB}_{\beta|\beta_0}(P_0, P_1) = \int \mathrm{f}_{\beta|\beta_0}(\frac{dP_0}{dP_1})dP_1, \tag{72}$$

which is the standard formula for f-divergences. In order to decide which of these divergences are genuine f-divergences, it remains to investigate when the functions $\mathrm{f}_{\beta|\beta_0}$ are convex functions. The following result holds.

Theorem 12. *If both β and β_0 are greater than 1, then $\mathrm{f}_{\beta|\beta_0}$ is not convex. If both β and β_0 are less than $\frac{1}{2}$, there are examples both of convex and of non-convex functions $\mathrm{f}_{\beta|\beta_0}$. In all remaining cases, $\mathrm{f}_{\beta|\beta_0}$ is convex. In cases of convexity the functions are strictly convex.*

Proof. □

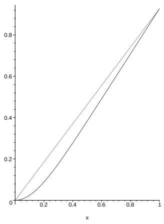

Fig. 4. V / $\mathrm{VB}_{\beta|\beta_0}$-diagram for $\beta = 5\,,\beta_0 = 1$.

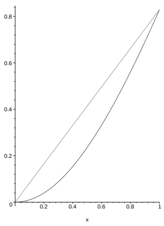

Fig. 5. V / $\mathrm{VB}_{\beta|\beta_0}$-diagram for $\beta = 2\,,\beta_0 = 1$.

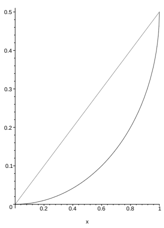

Fig. 6. V / $\mathrm{VB}_{\beta|\beta_0}$-diagram for $\beta = \frac{1}{2}\,,\beta_0 = 1$.

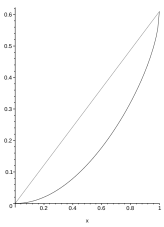

Fig. 7. V / $\mathrm{VB}_{\beta|\beta_0}$-diagram for $\beta = 2\,,\beta_0 = \frac{1}{2}$.

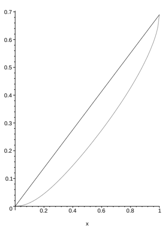

Fig. 8. V / $\mathrm{VB}_{\beta|\beta_0}$-diagram for $\beta = 5\,,\beta_0 = \frac{1}{2}$.

For two probability distributions, (P_0, P_1) on $(Y, \mathcal{B}$ we may then, for each locus y, consider variation from P_0 to P_1 at that locus. Integrating this over $(Y, \mathcal{B}, \eta)$ we obtain a measure of discrimination between P_0 and P_1. To be precise, for fixed non-negative parameters β and β_0 we define the discrimination measure $\mathrm{D}_{\beta|\beta_0}$ by

$$\mathrm{VB}_{\beta|\beta_0}(P_0 \| P_1) = \mathrm{VB}_{\beta|\beta_0}(f_0, f_1) \tag{73}$$

with (f_0, f_1) considerd as the map $(x, y) \curvearrowright f_x(y)$

8 General $\mathrm{V} \,/\, \mathrm{VB}_{\beta|\beta_0}$-diagrams

The special nature of total variation makes it possible to generalize the result from Section

References

[1] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser B*, 28:131–142, 1966.

[2] S. Arimoto. Information-theoretical considerations on estimation problems. *Information and Control*, 19:181–194, 1971.

[3] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer-Verlag, New York, 1984.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[5] I. Csiszár. On topological properties of f-divergences. *Studia Sci. Math. Hungar.*, 2:329–339, 1967.

[6] F. Topsøe. Information theory at the service of science. to appear, 200x.

[7] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, pages xx–xx, 2003. to appear.

[8] Alexei Fedotov, Peter Harremoës, and Flemming Topsøe. Refinements of Pinsker's inequality. *IEEE Trans. Inf. Theory*, pages xx–xx, 2003. to appear.

[9] G. H. Hardy, J. E. Littlewood, and G. Pólya. *Inequalities*. Cambridge University Press, second edition edition, 1959.

[10] Peter Harremoës and Flemming Topsøe. Inequalities between entropy and index of coincidence derived from information diagrams. *IEEE Trans. Inform. Theory*, 47:2944–2960, 2001.

[11] L. Hörmander. *Notions of Convexity.* Birkhäuser, Boston, 1994.

[12] S. Kullback. *Informaton Theory and Statistics.* Wiley, New York, 1959.

[13] S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22:79–86, 1951.

[14] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inform. Theory*, 37:145–151, 1991.

[15] J. Lin and S. K. M. Wong. A new directed divergence measure and its characterization. *Int. J. General Systems*, 17:73–81, 1990.

[16] A. Lyapunov. Nouvelle forme du théoreme sur la limite de probabilité. *Memoires de l'Acad. de St-Pétersbourg*, 12, 1901.

[17] T. P. Minka. Bayesian inference, entropy, and the multinomial distribution. Internet manuscript, 2001. available at http://www.stat.cmu.edu/ minka/papers/multinomial.html.

[18] D. S. Mitrinović. *Analytic Inequalities.* Springer, Berlin, 1970.

[19] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55:xxx–xxx, 2003. to appear.

[20] S. Fine R. El-Yaniv and N. Tishby. Agnostic classification of markovian sequences. *NIPS, MIT-Press*, pages 465–471, 1997.

[21] I. J. Schoenberg. Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, 44:522–536, 1938.

[22] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, 1948.

[23] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inform. Theory*, 46:1602–1609, 2000.

[24] F. Topsøe. Basic concepts, identities and inequalities – the toolkit of information theory. *Entropy*, 3:162–190, 2001. http://www.unibas.ch/mdpi/entropy/ [ONLINE].

[25] F. Topsøe. Maximum entropy versus minimum risk and applications to some classical discrete distributions. *IEEE Trans. Inform. Theory*, 48:2368–2376, 2002.

[26] A. K. C. Wong and M. You. Entropy and distance of random graphs with application to structural pattern recognition. *IEEE Trans. Pattern Anal. Machine Intell.*, 7:599–609, 1985.