Jensen-Shannon Divergence and Hilbert space embedding

Bent Fuglede and Flemming Topsøe
University of Copenhagen, Department of Mathematics

Consider the set $M_+^1(A)$ of probability distributions where A is a set provided with some σ-algebra.

Jensen-Shannon divergence JSD : $M_+^1(A) \times M_+^1(A) \to [0, \infty]$ is a symmetrized and smoothed version of the all important divergence measure of informaton theory, Kullback-Leibler divergence $\mathrm{D}(P\|Q)$. It is defined by

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

with $M = \frac{1}{2}(P+Q)$.

Apparently, it is gaining in popularity, especially among statisticians. Somewhat implicitly the quantity was introduced in Wong and You [15, Definition 13]. Lin and Wong derive some simple properties in [9] and this information is essentially repeated in Lin [8]. Further identities and inequalities appeared in Topsøe [13] (where the name "capacitory discrimination" was used). The result which triggered the research behind the present submission and which is a main reason why it is safe to predict a rise of interest in JSD-divergence is the fact that JSD is the square of a metric. In view of previous research, cf. references in [11], the result is not that surprising. However, it was first published very recently, in Endres and Schindelin [4] and, independently, in Österreicher and Vajda, [11]. The metric $\sqrt{\mathrm{JSD}}$ metrizes convergence in total variation as is clear in view of well known inequalities for certain classes of divergence measures (see [7] for a systematic treatment). The interest is in the metric itself and the information theoretic interpretations and results it gives rise to. The further results which are the basis for the present submission will be published in Fuglede [5] and Topsøe [3].

On pages 4 and 5 we give some simple arguments in favour of JSD-divergence. But first we turn to some technically more involved considerations which depend on partly classical results and methods from harmonic

analysis. The monograph [2] may be helpful regarding terminology used and basic results from that field.

The starting point is the observation that $\left(M_+^1(A), \sqrt{\mathrm{JSD}}\right)$ is isometrically isomorphic to a subset – a certain curve – in Hilbert space. This may be proved as an application of *Schoenberg's theorem*, cf. [12] or the relatively short proof in [2]. The result opens for an approach to certain parts of information theory based on differential geometry.

A drawback with a proof of the metric property of JSD-divergence via Schoenberg's theorem is that it is non-constructive and leaves the question open as to the actual embedding in Hilbert space. The answer to this question requires a good bit of further work which we shall now briefly indicate.

First observe that JSD-divergence is defined by integration (or summation) of terms generated by the *kernel* on $\mathbb{R}_+ = [0, \infty]$:

$$K(x,y) = \frac{x}{2}\ln\frac{2x}{x+y} + \frac{y}{2}\ln\frac{2y}{x+y}\,. \tag{2}$$

It therefore suffices to characterize the embedding of $(\mathbb{R}_+, \sqrt{K})$ in Hilbert space. The image turns out to be what we shall call a $\frac{1}{2}$-spiral. By an α-*spiral* in real Hilbert space, more precisely, a *logarithmic spiral of order* α, we understand a curve $t \curvearrowright x(t)\,;\, t \in \mathbb{R}$ for which

$$\|x(t_1+t) - x(t_2+t)\| = e^{\alpha t}\|x(t_1) - x(t_2)\|\,. \tag{3}$$

For $\alpha = 0$, the α-spirals become helixes as studied by Kolmogorov [6] and von Neumann and Schoenberg, [14]. An important structural result concerns the spectral representation of helixes. It turns out that the elegant proof by Masani in [10] of this result can be adapted to spirals. This leads to a characterization of spirals in terms of isometries to L^2-spaces.

Another approach to α-spirals is through negative definite kernels. The kernel $K : X \times X \to \mathbb{R}$ is *negative definite* if, 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

$$\sum_{i=1}^{n} c_i = 0 \implies \sum_{i,j} c_i c_j K(x_i, x_j) \le 0 \tag{4}$$

holds. A kernel on $\mathbb{R}_+$ is 2α-*homogeneous* if $K(tx, ty) = t^{2\alpha}K(x,y)$ for $x\,,\, y\,,\, t \in \mathbb{R}_+$.

Theorem 1. *The 2α-homogeneous negative definite kernels on $\mathbb{R}_+$ can be identified as kernels which have a representation*

$$K(x,y) = \int_0^\infty |x^{\alpha+i\lambda} - y^{\alpha+i\lambda}|^2 d\mu(\lambda) \tag{5}$$

for a bounded measure μ on $\mathbb{R}_+$.

If K has such a representation with $\mu(\{0\}) = 0$, then $(\mathbb{R}_+, \sqrt{K})$ is a metric space which can be embedded isometrically into the real Hilbert space $L^2(\mu) \oplus L^2(\mu)$ by the map $x \curvearrowright (Re(f_x), Im(f_x))$ where

$$f_x(\lambda) = (x^{\alpha+i\lambda} - 1)\frac{-\alpha + i\lambda}{\alpha + i\lambda}\,. \tag{6}$$

For the kernels (2), the *representing measure* is given by

$$d\mu(\lambda) = \frac{1}{\cosh(\pi\lambda)}\frac{1}{1+\lambda^2}d\lambda\,. \tag{7}$$

The generality of Theorem 1 points to the possibility to consider more general divergence measures. This is indeed possible. In fact you may generalize a one-parameter family of divergence measures considered by Arimoto [1] (see also [11]) to a two-parameter family. The kernels concerned on $\mathbb{R}_+$ are, for $\beta_0 \neq \beta$ given by

$$K_{\beta|\beta_0}(x,y) = \frac{\beta\,\beta_0}{\beta - \beta_0}\left((\frac{1}{2}x^\beta + \frac{1}{2}y^\beta)^{\frac{1}{\beta}} - (\frac{1}{2}x^{\beta_0} + \frac{1}{2}y^{\beta_0})^{\frac{1}{\beta_0}}\right)\,, \tag{8}$$

and for $\beta_0 = \beta$ by the expression

$$\|(x,y)\|_\beta^{1-\beta}\left(\frac{x^\beta}{2}\log x^\beta + \frac{y^\beta}{2}\log y^\beta - \frac{x^\beta + y^\beta}{2}\log(\frac{1}{2}x^\beta + \frac{1}{2}y^\beta)\right) \tag{9}$$

where the norm is with respect to the uniform measure $(\frac{1}{2}, \frac{1}{2})$.

Theorem 2. *Assume that $0 < \beta_0 \leq \beta$. The above kernels are negative definite if and only if $\beta_0 \geq \frac{1}{2}$ and $\beta \geq 1$. Clearly, the kernels are 1-homogeneous.*

The representing measures can be determined in terms of the complex gamma function.

We end by some simple considerations aiming at illuminating the significance of JSD-divergence.

Consider the discrete case and introduce entropy as usual, i.e. $H(P) = -\sum_n p_n \log p_n$ (with p_n's for the point probabilities of P).

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

$$H\left(\sum_\nu \alpha_\nu P_\nu\right) = \sum_\nu \alpha_\nu H(P_\nu) + \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) \tag{10}$$

and Jensen's inequality confirming the concavity of the entropy function,

$$H(\sum \alpha_\nu P_\nu) \geq \sum \alpha_\nu H(P_\nu)\,, \tag{11}$$

follows from the fundamental inequality $D(P\|Q) \geq 0$. In case $\sum_\nu \alpha_\nu H(P_\nu) < \infty$, we may write (10) in the form

$$H\left(\sum_\nu \alpha_\nu P_\nu\right) - \sum_\nu \alpha_\nu H(P_\nu) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P})\,. \tag{12}$$

The left hand side of (12) 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.

The right hand side of (12) 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

$$\mathrm{JSD}(\sum_\nu \alpha_\nu P_\nu) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) \tag{13}$$

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

Another advantage of the expression (13) is the interpretations it gives rise to. In fact, the quantity is the 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.

A second interpretation relates to the model which we shall refer to as the *switching model* where a source generates a string $x_1 x_2 \cdots$ 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 is α_ν. Consider an observer who knows the P_ν's as well as the α_ν's but does not know 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 choose, at each

instant, a code adapted to the distribution selected. 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 when P_ν is selected.

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. Associate with κ the distribution Q with $q_n = exp(-\kappa_n)$. The average amount of nits used by the observer is $\sum_n p_{\nu,n} \log \frac{1}{q_n}$ in case $P_\nu = (p_{\nu,n})$ is the actual distribution selected. 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 $\sum_\nu \alpha_\nu D(P_\nu \| Q)$. Therefore the observer should choose that distribution Q as the basis for coding which minimizes average redundancy given by $R(Q) = \sum_\nu \alpha_\nu D(P_\nu \| Q)$. In order to identify the associated argmin-distribution, we refer to the socalled *compensation identity* which states that

$$\sum_\nu \alpha_\nu D(P_\nu \| Q) = \sum_\nu \alpha_\nu D(P_\nu \| \overline{P}) + D(\overline{P} \| Q) \tag{14}$$

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

By (14), for any fixed Q, divergence $D(\cdot \| Q)$ is a convex function:

$$D\left(\sum_\nu \alpha_\nu P_\nu \| Q\right) \leq \sum_\nu \alpha_\nu D(P_\nu \| Q)\,. \tag{15}$$

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 (15), we are back to the quantity (13), we started with and this is independent of the distribution Q we choose to take as our reference.

References

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

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

[3] F. Topsøe. Jensen-Shannon divergence and norm-based measures of discrimination and variation. In preparation, draft available at http://www.math.ku.dk/ topsoe/.

[4] D. M. Endres and J. E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49:1858–60, 2003.

[5] B. Fuglede. Spirals in Hilbert space. With an application in information theory. Submitted for publication, November 2003.

[6] A. N. Kolmogorov. Curves in Hilbert space which are invariant with respect to a one-parameter group of motions. *Doklady Akad. Nauk*, 26:6–9, 1940. (russian).

[7] F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner, Leipzig, 1987.

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

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

[10] P. Masani. On helixes in Hilbert space. *Theory of Prob. and Appl.*, 17:1–19, 1972.

[11] F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55:639–653, 2003.

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

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

[14] J. von Neumann and I. J. Schoenberg. Fourier integrals and metric geometry. *Trasns. Amer. Math. Soc.*, 50:226–251, 1941.

[15] 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.