Jensen-Shannon Divergence and Hilbert space embedding

Bent Fuglede and Flemming Topsøe[1]
Department of Mathematics, University of Copenhagen
2100 Copenhagen, Denmark. e-mail: {fuglede, topsoe}@math.ku.dk

Consider a mixture $\sum_\nu \alpha_\nu P_\nu$ of probability distributions and put $\overline{P} = \sum_\nu \alpha_\nu P_\nu$. Then, with H for entropy and $\mathrm{D}(\cdot\|\cdot)$ for Kullback-Leibler divergence,

$$H(\sum_\nu \alpha_\nu P_\nu) - \sum_\nu \alpha_\nu H(P_\nu) = \sum_\nu \alpha_\nu D(P_\nu\|\overline{P}) \qquad (1)$$

provided $\sum_\nu \alpha_\nu H(P_\nu) < \infty$. We call this quantity the *general Jensen-Shannon divergence* pertaining to the mixture. Using the right hand side of (1) as definition, it is defined for distributions over arbitrary Borel spaces. Note the interpretation related to concavity of H as well as the similar interpretation related to convexity of $\mathrm{D}(\cdot\|Q)$ for any distribution Q:

$$\sum_\nu \alpha_\nu \mathrm{D}(P_\nu\|Q) - \mathrm{D}(\sum_\nu \alpha_\nu P_\nu\|Q) = \sum_\nu \alpha_\nu D(P_\nu\|\overline{P})\,.$$

Another interpretation relates to the *switching model* where a source generates a string $x_1x_2\cdots$ of letters, selected independently and each 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 and α_ν's but does not know which distribution is used at any particular time instant. Compare with an *ideal observer* who also knows which distribution is used at each time instant. The observer wants to design a code such that the expected *redundancy* is minimized. With natural definitions making these considerations precise, one finds that the general Jensen-Shannon divergence related to the mixture is the minimum redundancy which can be achieved by the observer.

Now turn to the *specific Jensen-Shannon divergence* which is the symmetrized and smoothed version of $\mathrm{D}(\cdot\|\cdot)$ given by $\mathrm{JSD}(P,Q) = \frac{1}{2}\mathrm{D}(P\|M) + \frac{1}{2}\mathrm{D}(Q\|M)$ with $M = \frac{1}{2}(P+Q)$. It thus corresponds to the uniform mixture $\frac{1}{2}P + \frac{1}{2}Q$. Previous research includes: [1] (implicit definition), [2] (simple properties), [3] (repetition of these), [4] (implicitly contains the result that triggered the authors' research, viz. the fact that $\sqrt{\mathrm{JSD}}$ is a metric), [5] (some identities and inequalities), [6] (explicit proof of the metric property) and [7] (another independent explicit proof). As is easily seen, $\sqrt{\mathrm{JSD}}$ metrizes convergence in total variation.

Theorem. The set of distributions with the metric $\sqrt{\mathrm{JSD}}$ can even be embedded isometrically into Hilbert space and the embedding can be identified.

The proof depends on a study of the *kernel* on $\mathbb{R}_+$: $K(x,y) = \frac{x}{2}\ln\frac{2x}{x+y} + \frac{y}{2}\ln\frac{2y}{x+y}$. It suffices to characterize the embedding of $(\mathbb{R}_+, \sqrt{K})$ in Hilbert space as JSD is obtained by integration of this kernel.

A kernel K on X is *negative definite* if, for real numbers $(c_i)_{i\le n}$ and points $(x_i)_{i\le n}$ in X, $\sum_{i,j} c_ic_jK(x_i,x_j) \le 0$, whenever $\sum_i c_i = 0$. A kernel on $\mathbb{R}_+$ is 2α-*homogeneous* if $K(tx,ty) = t^{2\alpha}K(x,y)$ for x, y, $t \in \mathbb{R}_+$.

By a *logarithmic spiral of order* α in (real) Hilbert space, we understand a curve $t \curvearrowright x(t)$; $t \in \mathbb{R}$ for which $\|x(t_1+t) - x(t_2+t)\| = e^{\alpha t}\|x(t_1) - x(t_2)\|$. For $\alpha = 0$, these are helixes.

Generalizing spectral properties developed in [8] for helixes, one can prove:

Theorem. The 2α-homogeneous negative definite kernels on $\mathbb{R}_+$ can be identified by the representation

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

with μ a bounded measure on $\mathbb{R}_+$. If (2) holds with $\mu(\{0\}) = 0$, then $(\mathbb{R}_+, \sqrt{K})$ can be embedded isometrically into $L^2(\mu) \oplus L^2(\mu)$ by $x \curvearrowright (Re(f_x), Im(f_x))$ where $f_x(\lambda) = (x^{\alpha+i\lambda} - 1)\frac{-\alpha+i\lambda}{\alpha+i\lambda}$.

For the concrete kernel above,

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

Other applications concern generalizations of divergence measures considered by Arimoto [9], cf. also [6].

References

[1] 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.

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

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

[4] F. Österreicher and I. Vajda. Statistical information and discrimination. *IEEE Trans. Inform. Theory*, 39:1036–1039, 1993.

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

[6] 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.

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

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

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

[1] Supported by INTAS, Project 00-738 and by the Danish Natural Science Research Council.

0-7803-8280-3/04/$20.00 ©2004 IEEE