On entropies, divergences, and mean values

Michèle Basseville and Jean-François Cardoso[1]

IRISA/CNRS, Campus de Beaulieu, 35042 Rennes Cedex, France - E-mail: basseville@irisa.fr,
and Télécom Paris/CNRS, 46 rue Barrault, 75634 Paris Cedex 13, France - E-mail: cardoso@sig.enst.fr.

Abstract — **Two entropy-based divergence classes are compared using the associated quadratic differential metrics, mean values and projections.**

I. Two classes of divergences

The design concepts of divergences are of interest because of the key role they play in statistical inference and signal processing. Most of the existing divergences $\mathbf{D}$ between two probability distributions may be associated with an integral or non integral entropy functional $\mathbf{H}_\nu(\mu)$ with respect to some reference measure ν. We distinguish two different classes of divergences built on entropies. The first one is the well known class of f-divergences $\mathbf{I}_f$ [4] which are based on the likelihood ratio and are formally identical to the above entropies $\mathbf{I}(\mu,\nu) \stackrel{\Delta}{=} -\mathbf{H}_\nu(\mu)$. In the integral case, this yields the relative entropy class, which includes Kullback information as its most prominent member [4]. The most important instance of non integral f-divergences is Rényi information [13].

The second class of divergences builds upon the concavity of an entropy functional, which entails that, for $0 < \alpha < 1$, $\mathbf{J}_H^{(\alpha)}(\mu,\nu) \stackrel{\Delta}{=} \mathbf{H}((1-\alpha)\mu+\alpha\nu) - (1-\alpha)\mathbf{H}(\mu) - \alpha\mathbf{H}(\nu)$ is positive. One can then construct $\mathbf{C}_H(\mu,\nu) = \max_\alpha \mathbf{J}_H^{(\alpha)}(\mu,\nu)$, a Jensen difference $\mathbf{J}_H(\mu,\nu) = \mathbf{J}_H^{(1/2)}(\mu,\nu)$, and a Bregman distance $\mathbf{D}_H(\mu,\nu) = \lim_{\alpha\to 0} \alpha^{-1}\mathbf{J}_H^{(\alpha)}(\mu,\nu)$. Bregman distances enjoy an Euclidian-like property, similar to the Pythagorean theorem [9, 7], when involved in projections onto exponential or mixture families. This may be further generalized to projections onto 'α-families' as shown in [2] where families of distributions are dealt with as differential manifolds. Still in this geometrical vein, the interplay between $\mathbf{C}_H, \mathbf{J}_H$ and $\mathbf{D}_H$ can be understood via Thales theorem.

A local quadratic differential metric is associated with any divergence measure [12, 2]. Based on the fact that f-divergences are locally equivalent to the Riemannian metric defined by the Fisher information matrix, we characterize the intersection of the two above divergence classes. In particular, it is easily found that the only Bregman distance $\mathbf{D}_H$ which is a f-divergence is Kullback information [9, 7]. Similarly, it is found that the only f-divergences which can be written as a Jensen difference $\mathbf{J}_H^{(\alpha)}$ are those introduced in [11, 10].

II. Associated mean values and projections

Mean values can be associated with entropy-based divergences in two different ways. The first way [13, 1] consists in writing explicitly the *generalized mean values* $\phi^{-1}\left(\sum_{i=1}^n \beta_i\, \phi(p_i)\right)$ underlying integral and non integral f-divergences. Here the β's are normalized positive weights. For Rényi information, $\phi(u) = u^\alpha$, and this results in *α-mean values* $\left(\sum_{i=1}^n \beta_i\, p_i^\alpha\right)^{1/\alpha}$.

The second way [3] consists in defining mean values by $\arg\min_v \sum_{i=1}^n \beta_i\, d(v,u_i)$, namely as *projections*, in the sense of distance d, onto the half-line $u_1 = \ldots = u_n > 0$ [7]. When d is an integral f-divergence $d(v,u_i) = u_i f\left(\frac{v}{u_i}\right)$, this gives the *entropic means* [3], which are characterized implicitly by $\sum_{i=1}^n \beta_i\, f'\left(\frac{v}{u_i}\right) = 0$, and necessarily homogeneous (scale invariant). The class of entropic means includes all available integral means and, when applied to a random variable, contains most of centrality measures (moments, quantiles). When d is a Bregman distance $d_h(u,v) = h(u) - h(v) - (u-v)h'(v)$, the corresponding mean values are exactly the above generalized mean values (for $\phi = h'$), which are generally not homogeneous.

The only generalized mean value which is also an entropic mean, and thus both an f-divergence-projection and a Bregman-projection, is the above α-mean value, corresponding to Rényi information [3]. This agrees with invariant properties of means [8] and the axiomatic of inference in [5].

Finally, we mention that mutual information (viewed both as relative entropy and Jensen difference) and the related concepts of channel capacity [6] and information radius [14], can be seen as another manner of investigating the intersection of the above two divergence classes.

[1] The authors are also with GDR CNRS no 134 'Traitement du Signal et Images'.

References

[1] J. Aczél and Z. Daròczy, "*On Measures of Information and Their Characterizations*," Academic Press, 1975.

[2] S-I. Amari, "*Differential-Geometrical Methods in Statistics*," Lecture Notes in Statistics, vol.28, Springer-Verlag, 1985.

[3] A. Ben-Tal, A. Charnes and M. Teboulle, "Entropic means," *Jal Math. Anal. Appl.*, vol.139, pp.537-551, 1989.

[4] I. Csiszár, "Information measures: a critical survey," *7th Prague Conf. Inf.Th., Stat.Dec.Funct. and Rand.Proc.*, pp.73-86, 1974.

[5] I. Csiszár, "Why least-squares and maximum entropy ? An axiomatic approach to inference for linear inverse problems," *Annals Statistics*, vol.19, pp.2032-2066, Jul. 1991.

[6] I. Csiszár, "Generalized cutoff rates and Rinyi's information measures," *IEEE Trans. Information Theory*, vol.IT-40, pp.26-34, Jan. 1995.

[7] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol.68, pp.161-185, Jan. 1995.

[8] G.H. Hardy, J.E. Littlewood and G. Polya, "*Inequalities*," Cambridge Univ. Press, 1952.

[9] L.K. Jones and C.L. Byrne, "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Trans. Inform. Theory*, vol.IT-36, pp.23-30, Jan. 1990.

[10] L. Knockaert, "A class of statistical and spectral distance measures based on Bose-Einstein statistics," *IEEE Trans. Signal Proc.*, vol.SP-41, pp.3171-3174, Nov. 1993.

[11] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inform. Theory*, vol.IT-37, pp.145-151, Jan. 1991.

[12] C.R. Rao, "Differential metrics in probability spaces," in *IMS Lecture Notes*, vol.10, S. Gupta (ed.), pp.217-240, 1987.

[13] A. Rényi, "On measures of entropy and information," *4th Berkeley Symp. Math. Stat. Proba.*, vol.I, pp.547-561, 1961.

[14] R. Sibson, "Information radius," *Z. Wahrscheinlichkeitsth. Werw. Gebiete*, vol.14, pp.149-160, 1969.