Norm-based measures of Discrimination and Variation

F. Topsøe [1]
Department of Mathematics
University of Copenhagen
topsoe@math.ku.dk

Abstract — **A function $f : X \to \mathbb{R}$ varies between its minimum and maximum. Using L^p-norms, a two-parameter family of measures of variation is defined, based on a given "Bayesian" distribution on X. Introducing a second measure space over a set Y, we define an integrated notion of variation for functions $f : X \times Y \to \mathbb{R}$.**

Applications include hypothesis testing in statistics, feature selection problems in pattern recognition and others.

When X is a two-element set provided with the uniform distribution we obtain a two-parameter family of measures of discrimination between distributions P and Q over Y. The contribution is, therefore, an addition to the extensive literature on measures of discrimination or divergence between probability distributions. In particular, we build on Österreicher and Vajda [5] who in turn builds on Arimoto [1]. In fact, the two-parameter family considered here extends in a natural way one-parameter families appearing in the research just cited.

I. NORM-BASED MEASURES OF VARIATION

Let $(X, \mathcal{A}, \mu)$ be a probability space and consider a measurable, non-negative function on X. Consider the *p-norm* of f: $\|f\|_p = \left(\int_X f(x)^p d\mu(x)\right)^{\frac{1}{p}}$. We need the following rather well known result:

Lemma 1. *The function ϕ defined on $\mathbb{R}$ (or even on the extended real line) by $\phi(p) = \|f\|_{1/p}$ 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 $\phi(0^+) = f_{\max}$. In particular, ϕ is convex and decreasing in $\mathbb{R}_+$.*

Figure 1 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 by

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

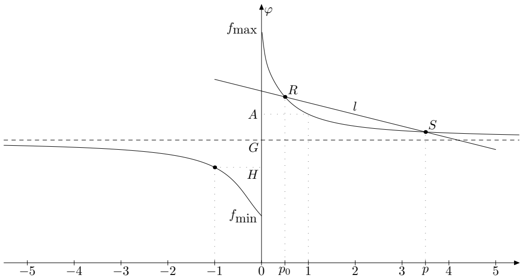

Fig. 1. The function $\phi : p \curvearrowright \|f\|_{1/p}$.

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 $(p_0, \phi(p_0))$. A direct calculation shows that

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

in particular,

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

Some general properties are *symmetry*: $\mathrm{VB}_{\beta|\beta_0} = \mathrm{VB}_{\beta_0|\beta}$, *homogeneity*: $\mathrm{VB}_{\beta|\beta_0}(cf) = c\,\mathrm{VB}_{\beta|\beta_0}(f)$, *monotonicity*: $\beta \curvearrowright \mathrm{VB}_{\beta|\beta_0}(f)$ is increasing in $\mathbb{R}_+$ and varies from 0 to $\beta_0\,(f_{\max} - \|f\|_{\beta_0})$, and *consistency*: $\mathrm{VB}_{\beta|\beta_0}(f) = 0$ if and only if f is constant a.e. Also a property of datareduction can be established for certain parameter values.

The notion may be extended to situations where we have also given a σ-finite measure space $(Y, \mathcal{B}, \eta)$. If $f : (x, y) \curvearrowright f(y|x)$ is non-negative on $X \times Y$, we define:

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

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. η. Denote by $y \curvearrowright f(y|x)$ the corresponding density. We denote by $\overline{P}$ the natural mixture of the P_x's: $\overline{P} = \int P_x d\mu(x)$. Clearly, $\overline{f}$ given by $\overline{f}(y) = \int f(y|x) d\mu(x)$ is the density of $\overline{P}$ w.r.t. η.

---

[0] Supported by INTAS (project 00-738) and by the Danish Natural Science Research Council.

The special case $\beta = \beta_0 = 1$ gives the quantity $\int \mathrm{D}(P_x\|\overline{P})\,d\mu(x)$ which is the *Jensen-Shannon variation measure* associated with the family of distributions.

II. Discrimination in $\mathbb{R}_+$

We consider variation breadth of the simple non-integrated type – i.e. Y is a one-element set – and we take as $(X, \mathcal{A}, \mu)$ the uniform distribution on a two-element set. Then variation breadth amounts to a measure of discrimination between two non-negative real numbers. We stick to the VB$_-$notation, but refer to it as *discrimination.* We denote by (x, y) the function assuming the two values x and y and by $\|(x, y)\|_\beta$ the β-norm w.r.t. the uniform measure. Then, for $\beta \neq \beta_0$,

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

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

$$\begin{aligned}
\mathrm{VB}_{\infty|1}(x,y) &= \frac{1}{2}|x-y|\,,\\
\mathrm{VB}_{2|1}(x,y) &= \left(\sqrt{2x^2+2y^2} - (x+y)\right)\,,\\
\mathrm{VB}_{1|1}(x,y) &= \frac{1}{2}x\log\frac{2x}{x+y} + \frac{1}{2}y\log\frac{2y}{x+y}\,,\\
\mathrm{VB}_{\frac{1}{2}|1}(x,y) &= \frac{1}{4}\left(\sqrt{x}-\sqrt{y}\right)^2\,.
\end{aligned}$$

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.

Theorem 1. *For $\frac{1}{2} \leq \beta_0 \leq 1 \leq \beta \leq \infty$, 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*

$$\mathrm{VB}_{\frac{1}{2}|1}(x,y) \leq \mathrm{VB}_{\beta|\beta_0}(x,y) \leq \mathrm{VB}_{\infty|1}(x,y) \qquad (1)$$

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

The proof depends on rather deep facts from harmonic analysis which can be found in [2].

III. Norm-based measures of discrimination for distributions

Now, X is again a two-element set with uniform distribution and $(Y, \mathcal{B}, \eta)$ is any σ-finite measure space. We study discrimination among probability measures which are absolutely continuous with respect to η. Let P_0 and P_1 be two such distributions and denote by f_0 and f_1 their density functions. We define:

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

If $\beta_0 = \beta = 1$, we write $\mathrm{JSD}(P_0, P_1)$ for $\mathrm{VB}_{1|1}(P_0, P_1)$ and refer to this as the (specific) *Jensen-Shannon divergence measure.* In 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 or Ali and Silvey (see standard references). However, the property of convexity which is essential for many results may not hold. It is easy to identify a family $\mathrm{f}_{\beta|\beta_0}$ of functions on $\mathbb{R}_+$ such that

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

which is the standard formula for f-divergences.

Theorem 2. *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.*

Other properties concern the issue of metrics. The natural analogue of Theorem 1 holds. The proof depends on Theorem 1 and hence on the indicated results from harmonic analysis. The result obtained extends previous results of Endres and Schindelin [4] and Österreicher and Vajda [5]. A precise conjecture is formulated regarding cases when $\mathrm{VB}_{\beta|\beta_0}(P, Q)$ defines a squared metric (details can be found on the net, cf. [3]). Finally we mention that the issue of information diagrams, comparing $\mathrm{VB}_{\beta|\beta_0}(P, Q)$ with total variation has been settled in the case of Csiszár divergences, following a new general result, likewise submitted to ITW2003.

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*, pages xx–xx, 2003. To appear.

[5] 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.