A new metric for probability distributions

Dominik M. Endres, Johannes E. Schindelin

Abstract— **We introduce a metric for probability distributions, which is bounded, information theoretically motivated and has a natural Bayesian interpretation. The square root of the well-known χ^2 distance is an asymptotic approximation to it. Moreover, it is a close relative of the capacitory discrimination and Jensen-Shannon divergence.**

I. Introduction

This paper is the result of the authors' search for a probability metric that is bounded and can be easily interpreted in terms of both information theoretical as well as probabilistic concepts. Metric properties are the prerequisites for several important convergence theorems for iterative algorithms, i.e. Banach's fixed point theorem [2], which is the basis of several pattern-matching algorithms. Boundedness is a valuable property, too, when numerical applications are considered.

We will limit the following discussion to discrete probability distributions, but the result can be generalized to probability density functions.

II. Motivation

The motivation we are presenting in this section is aimed at providing the reader with an idea of the meaning of the metric. As such it is not to be understood as a derivation in a strict mathematical sense. However, we will observe mathematical rigor in the following section, which contains the actual proof of the metric properties.

Let X be a discrete random variable which can take on N different values $\in \Omega_N = \{\omega_1, \ldots, \omega_N\}$. We now draw an *i.i.d.* sample $\tilde{X}$, where each observation is drawn from one of two known distributions, P and Q. Each of those is used with equal probability. However, we do not know, which one is used when. Now we wish to find the coding strategy that gives the shortest average codelength for the representation of the data. In other words, we are looking for the most *efficient* distribution R.

Let us call this code κ. The codelengths are $\kappa_i = -\log r_i$, where $i \in \{1, \ldots, N\}$ and r_i is the probability of $X = \omega_i$ under R. Denoting the expectation of κ w.r.t. P by $\mathcal{E}(\kappa, P)$, the average codelength $<\kappa>$ is then $\frac{1}{2}\mathcal{E}(\kappa, P) + \frac{1}{2}\mathcal{E}(\kappa, Q)$. By the very definition of the entropy, the *minimum* $<\kappa>$ is obtained by setting $R = \frac{1}{2}(P + Q)$, i.e. $<\kappa>= H(R)$.

An ideal observer, i.e. one who knows which distribution is used to generate the individual data, could reach an even shorter average codelength $\frac{1}{2}H(P) + \frac{1}{2}H(Q)$. Hence the redundancy of κ is $H(R) - \frac{1}{2}H(P) - \frac{1}{2}H(Q)$. The distance measure we studied is twice that redundancy

$$\begin{aligned} D^2_{PQ} &= 2H(R) - H(P) - H(Q) \\ &= D\left(P\| R\right) + D\left(Q\| R\right) \\ &= \sum_{i=1}^{N} \left(p_i \log \frac{2p_i}{p_i + q_i} + q_i \log \frac{2q_i}{p_i + q_i} \right) \end{aligned} \quad (1)$$

Since the Kullback divergence $D\left(P\| R\right)$ can be interpreted as the inefficiency of assuming that the true distribution is R when it really is P, D^2_{PQ} could be seen as a minimum inefficiency distance.

We are not the first ones to introduce this distance measure. Topsøe, in [9], called it *capacitory discrimination* and introduced it from an information transmission point of view. In that paper, its properties are studied in depth. We will relate his results to ours in the discussion. Now D^2_{PQ} is obviously symmetric and vanishes for $P = Q$, but it does not fulfill the triangle inequality. However, its square root, D_{PQ}, does. The proof of the metric properties of D_{PQ} is the subject of the next section.

III. Proof of metric properties of $\mathbf{D_{PQ}}$

In the following, $\mathbb{R}^+$ includes 0.

Definition 1: Let the function $L(p, q) : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ be defined by

$$L(p, q) := p \log \frac{2p}{p + q} + q \log \frac{2q}{p + q}. \quad (2)$$

This function can be taken to be any one of the summands of D^2_{PQ} (see eqn. (1)). By standard inequalities we realize that $L(p, q) \geq 0$ with equality only for $p = q$.

Theorem 1 uses some properties of the partial derivative of $L(p, q)$ and to show these we introduce the function $g : \mathbb{R}^+\backslash\{1\} \to \mathbb{R}$ defined by

$$g(x) := \frac{\log \frac{2}{x+1}}{\sqrt{L(x, 1)}}.$$

Lemma 1: Let g be defined as above. Then
1. $\lim_{x\to 1^\mp} g(x) = \pm 1$, i.e. g jumps from $+1$ to -1 at $x = 1$.
2. The derivative $\frac{\mathrm{d}}{\mathrm{d}x} g$ is positive for $x \in \mathbb{R}^+\backslash\{1\}$.

A consequence of this lemma is that $|g(x)| \leq 1$ with equality only at $x = 1$. Also, it is easy to see that $|g|$ is continuous, but not g.

Proof: First note that g changes sign at $x = 1$.

A straightforward application of l'Hôspital's rule (differentiate twice) yields $\lim_{x\to 1} g^2(x) = 1$.

By differentiation one finds that $\frac{\mathrm{d}}{\mathrm{d}x} g$ is positive if and only if $f < 0$ where f is given by

$$f(x) = \log \frac{2}{1 + x} + \log \frac{2x}{1 + x}.$$

D. Endres, School of Psychology, University of St Andrews, St Andrews KY16 9JU, U.K. E-mail: dme2@st-andrews.ac.uk

J. Schindelin, Institut für Genetik, Biozentrum, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany. E-mail:gene099@mail.uni-wuerzburg.de

Straightforward differentiation shows that $f(1) = f'(1) = 0$ and that

$$f''(x) = \frac{-1}{x^2(1+x)}\left(\log\frac{2}{1+x} + x^2\log\frac{2x}{1+x}\right).$$

Using the standard inequality $\log a \geq 1 - \frac{1}{a}$, we find that $f'' < 0$, hence f is concave. Combined with the first found facts, $f < 0$ for $x \neq 1$◇

We will now prove

Theorem 1: Let $\mathcal{F}_N$ be the set of all discrete probability distributions over Ω_N, $N \in \mathbb{N}$. The function $D_{PQ} : \mathcal{F}_N \times \mathcal{F}_N \to \mathbb{R}^+$ is a metric.

Proof: To show this, we recall that $D\,(P\|\,Q)$ is 0 for $P = Q$ and strictly positive otherwise (see e.g. [3]). In addition, D^2_{PQ} is symmetric in P, Q and so is D_{PQ}. Therefore, we only have to show that the triangle inequality holds.

Lemma 2: Let $p, q, r \in \mathbb{R}^+$. Then

$$\sqrt{L(p,q)} \leq \sqrt{L(p,r)} + \sqrt{L(r,q)}.$$

Proof: It is easy to see that this holds if any of p, q, r are zero. Now we assume $p \leq q$, denote by **rhs** the right hand side as a function of r, and show that

1. **rhs** has 2 minima, namely one at $r = p$ and one at $r = q$, and
2. only 1 maximum somewhere between p and q.

We show this by way of the derivative

$$\frac{\partial \mathbf{rhs}}{\partial r} = \frac{\log\frac{2r}{p+r}}{2\cdot\sqrt{L(p,r)}} + \frac{\log\frac{2r}{q+r}}{2\cdot\sqrt{L(q,r)}}. \quad (3)$$

With g as in Lemma 1 and $x := \frac{p}{r}$ and $\beta\cdot x := \frac{q}{r}$ ($\beta > 1$), we find that

$$2\cdot\sqrt{r}\cdot\frac{\partial \mathbf{rhs}}{\partial r} = g(x) + g(\beta x).$$

With $|g(x)| \leq 1$ with equality only at $x = 1$, and the fact that g jumps from $+1$ to -1 at $x = 1$ (see lemma 1), the derivative $\frac{\partial \mathbf{rhs}}{\partial r}$ indeed changes sign at $r = p$, because then $x = 1$ and $|g(x)| > |g(\beta x)|$, and likewise at $r = q$. Those extrema are minima because r is reciprocal to x.

Also, $\frac{\mathrm{d}}{\mathrm{d}x}g(x) \geq 0$, therefore between $x = \frac{1}{\beta}$ and $x = 1$, $g(x) + g(\beta x)$ is monotonic increasing and as a consequence has at most one sign change. ◇

Applying Minkowski's inequality to the square root of the sum which defines D_{PQ}, we see that the triangle inequality is fulfilled.

Whence D_{PQ} is a metric.◇

The generalization of this result to continuous random variables is straightforward. Let P and Q be probability measures defined on a measurable space (Ω, A) and let $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$ be their Radon-Nikodym derivatives w.r.t. a dominating σ-finite measure μ. Then

$$D_{PQ} = \sqrt{\int_\Omega \left(p\log\frac{2p}{p+q} + q\log\frac{2q}{p+q}\right)d\mu} \quad (4)$$

is a metric, too.

An alternative proof could be constructed using results presented in [4]. Since D^2_{PQ} is an instance of a class of distances known as f-divergences (cf. [1]) (let $f(t) = t\log\frac{2t}{1+t} + \log\frac{2}{1+t}$, then $D^2_{PQ} = \sum_{i=1}^N q_i f(\frac{p_i}{q_i})$), the theorems proven in [4] apply.

Now we will look at the maxima and minima of D_{PQ}. Its minimum is, of course, located at $P = Q$, where $D_{PQ} = 0$. To find its maximum, rewrite (2) in the form

$$L(p,q) = \underbrace{(p+q)\log 2}_{\geq 0} + \underbrace{p\log\left(\frac{p}{p+q}\right)}_{\leq 0} + \underbrace{q\log\left(\frac{q}{p+q}\right)}_{\leq 0} \quad (5)$$

It follows that when P and Q are two distinct deterministic distributions, D_{PQ} assumes its maximum value $\sqrt{2\log 2}$.

IV. Asymptotic approximation

Next, we shall investigate the limit

$$\lim_{P\to Q} D^2_{PQ} \quad (6)$$

A term-by-term expansion of D_{PQ} to second order in p_j yields:

$$D^2_{PQ} \approx \sum_{j=1}^N \frac{1}{4q_j}(p_j - q_j)^2 = \frac{1}{4}\chi^2(P,Q) \quad (7)$$

where $\chi^2(P,Q)$ is the well-known χ^2-distance (see e.g [5]).

V. Discussion

The D_{PQ} metric can also be interpreted as the square root of an entropy approximation to the logarithm of an evidence ratio when testing if two (equally long) samples have been drawn from the same underlying distribution [6]. In that paper, it is also argued that $\frac{1}{2}D^2_{PQ}$ should be named Jensen-Shannon divergence, or rather, a special instance of that divergence, which is defined as

$$\begin{aligned} D_\lambda(P,Q) &= \lambda D\,(P\|\,R) + (1-\lambda)D\,(Q\|\,R) \\ R &= \lambda P + (1-\lambda)Q \end{aligned}$$

and therefore $\frac{1}{2}D^2_{PQ} = D_{\frac{1}{2}}(P,Q)$.

Topsøe [9] has interpreted capacitory discrimination as twice an information transmission rate and related it to a variety of other distance measures, such as the Kullback divergence, triangular discrimination, variational distance and Hellinger distance. Many of the inequalities found by him can now be rewritten to become relationships between metrics.

Österreicher, in [7], proved the triangle inequality for square roots of f_β divergences defined by the functions

$$f_\beta(t) = \frac{(1+t^\beta)^{\frac{1}{\beta}} - 2^{\frac{1-\beta}{\beta}}(1+t)}{1-\frac{1}{\beta}} \quad (8)$$

for $\beta > 1$. Since the f_β divergence one obtains by taking the limit $\beta \to 1$ is D^2_{PQ} (a fact pointed out to us by one of the reviewers), our result extends the theorem proven in [7] to include the case $\beta = 1$.

Another way of looking at D^2_{PQ} is from the viewpoint of Bayesian inference. Consider the following scenario: We draw a sample $\tilde{X}_1 = \{x_1\}$ of length 1 from an unknown distribution R. What we do know about the distribution is that it is either P or Q, hence assigning each distribution the prior probability $\frac{1}{2}$. We now use Bayesian inference to calculate the posterior probabilities $P(R = P|\tilde{X}_1)$,$P(R = Q|\tilde{X}_1)$ of each distribution given the observation $\tilde{X}_1$:

$$\begin{aligned} P(R = P|\tilde{X}_1) &= \frac{\frac{1}{2}P(x_1)}{\frac{1}{2}P(x_1) + \frac{1}{2}Q(x_1)} \\ P(R = Q|\tilde{X}_1) &= \frac{\frac{1}{2}Q(x_1)}{\frac{1}{2}P(x_1) + \frac{1}{2}Q(x_1)} \end{aligned} \tag{9}$$

The information gain $\Delta I(x_1)$ resulting from the observation of $\tilde{X}_1$ is given by the Kullback divergence between the posterior and the prior

$$\Delta I(x_1) = \frac{P(x_1)\log\frac{2P(x_1)}{P(x_1)+Q(x_1)} + Q(x_1)\log\frac{2Q(x_1)}{P(x_1)+Q(x_1)}}{P(x_1) + Q(x_1)} \tag{10}$$

To find the expected value of this gain, we now average $\Delta I(x_1)$ over the prior distribution of x_1, which is given by $\frac{1}{2}P + \frac{1}{2}Q$. This yields, noting that $P(x_1 = \omega_i) = p_i$ and likewise for Q:

$$\begin{aligned} \mathcal{E}(\Delta I(x_1)) &= \frac{1}{2}\sum_{i=1}^{N} p_i \log\frac{2p_i}{p_i + q_i} \\ &\quad + \frac{1}{2}\sum_{i=1}^{N} q_i \log\frac{2q_i}{p_i + q_i} \\ &= \frac{1}{2}D^2_{PQ} \end{aligned} \tag{11}$$

Therefore, another interpretation of D_{PQ} is that it is twice the expected information gain when deciding (by means of a sample of length 1) between two distributions given a uniform prior over the distributions. Consider now the case that P and Q are such that D_{PQ} is maximized. Then, as stated above, $\frac{1}{2}D^2_{PQ} = 1$ (when using $\log_2$), i.e. the information gain is one bit. Thus, a sample of length 1 is sufficient to make the (binary) decision as to which distribution is the correct one. More general formulas than (11) can be found in [8], where relations between arbitrary f-divergences and information gains in decision problems are studied.

VI. Acknowledgment

We would like to thank Dr. Peter Földiák for his valuable comments on our manuscript. Moreover, we are grateful for the references and suggestions provided to us by the unknown reviewers.

References

[1] Ali, S.M. and Silvey, S.D. (1966). A general class of coefficients of divergence of one distribution from another. *J. Roy. Stat. Soc. Ser. B*, vol 28, 131-42.

[2] Brown, R.F.(1993). A topological introduction to nonlinear analysis. Birkhäuser, Kassel.

[3] Cover, T.M. and Thomas J.A. (1991). Elements of Information Theory. John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

[4] Kafka, P., Österreicher, F., and Vince, I. (1991). On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar*, vol 26, pp. 415-22.

[5] Liese, F. and Vajda, I. (1987). Convex Statistical Distances. B.G. Teubner Verlagsgesellschaft, Leipzig.

[6] Minka, T. P. (2001). Bayesian inference, entropy, and the multinomial distribution. http:// www.stat.cmu.edu/˜minka/papers/multinomial.html.

[7] Österreicher, F. (1996). On a class of perimeter-type distances of probability distributions. *Kybernetika*, vol 32, pp. 389-93.

[8] Österreicher, F. and Vajda, I. (1993). Statistical Information and Discriminiation. *IEEE Trans. Info. Theory*, vol 36, pp. 1036-39.

[9] Topsøe, F. (2000): Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Info. Theory*, vol IT-46,pp. 1602-09.