A Lower Bound on Arbitrary f–Divergences in Terms of the Total Variation

Jochen Bröcker*
Max–Planck–Institut für Physik komplexer Systeme
Nöthnitzer Strasse 34
01187 Dresden
Germany

March 10, 2009

Abstract

An important tool to quantify the likeness of two probability measures are f–divergences, which have seen widespread application in statistics and information theory. An example is the total variation, which plays an exceptional role among the f–divergences. It is shown that every f–divergence is bounded from below by a monotonous function of the total variation. Under appropriate regularity conditions, this function is shown to be monotonous.

Remark: The proof of the main proposition is relatively easy, whence it is highly likely that the result is known. The author would be very grateful for any information regarding references or related work.

1 The total variation

Let (Ω, σ) be a probability space. A *signed measure* ν is a σ–additive set function with values in $\mathbb{R} \cup \{-\infty, \infty\}$, and so that either $\nu > -\infty$ or $\nu < \infty$. I will use the standard term *measure* if ν is nonnegative. To any signed measure ν, there corresponds a *Hahn–Jordan decomposition* of Ω into two measurable sets P, N so that $P \cup N = \Omega$, $P \cap N = \emptyset$ and

$$\nu^+(.) = \nu(. \cap P), \qquad \nu^-(.) = -\nu(. \cap N) \tag{1}$$

are both (nonnegative) measures. Obviously, $\nu = \nu^+ - \nu^-$. Furthermore, the representation

$$\nu^+(A) = \sup_{B \subset A} \nu(B), \qquad \nu^-(A) = -\inf_{B \subset A} \nu(B) \tag{2}$$

holds for every measurable set A. For a proof of these facts see [2]. The measure $\langle \nu \rangle = \nu^+ + \nu^-$ is called the *variation measure* of ν, which in turn defines the

*email: `broecker@pks.mpg.de`

total variation $\|\nu\| = \langle\nu\rangle(\Omega)$. If $\nu(\Omega) = 0$, it follows easily from the previous statements that

$$\langle\nu\rangle(\Omega) = 2\sup_{B\in\sigma}|\nu(B)|. \tag{3}$$

A *probability measure* is a measure μ so that $\mu(\Omega) = 1$. For any two probability measures, μ, ν, the difference $\mu - \nu$ is a signed measure, and Equation (3) applies. Hence,

$$\|\mu - \nu\| = \langle\mu - \nu\rangle(\Omega) = 2\sup_{B\in\sigma}|\mu(B) - \nu(B)|. \tag{4}$$

Obviously, $\|\mu - \nu\|$ is a metric for probability measures, namely the *total variation metric*, with Equation (4) providing two possible representations. If μ is absolutely continuous with respect to μ, then there is a third representation, namely

$$\|\mu - \nu\| = \int |\frac{\mathrm{d}\mu}{\mathrm{d}\nu} - 1|\mathrm{d}\nu. \tag{5}$$

`Proof of this fact`

2 The f-divergences

Equation (5) can be read as follows:

$$\|\mu - \nu\| = \int f(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\mathrm{d}\nu, \tag{6}$$

with $f(x) = |x - 1|$. There is a way to generalise this approach by using other forms of f. Let f be a convex function on $\mathbb{R}_{\geq 0}$ that vanishes at $x = 1$. Let μ, ν two probability measures with μ being absolutely continuous with respect to ν (which will be written as $\mu \ll \nu$). The *f–divergence* between μ and ν is given by

$$D_f(\mu, \nu) = \int f(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\mathrm{d}\nu. \tag{7}$$

For, if $\mu = \nu$ we have $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = 1$, we see that $f(\mu, \nu)$ vanishes in this case. Furthermore, $D_f(\mu, \nu)$ is non-negative. Indeed, by Jensen's inequality,

$$0 = f(1) = f(\int \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\mathrm{d}\nu) \leq \int f(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\mathrm{d}\nu = f(\mu, \nu).$$

Note though that $f(\mu, \nu)$ may be infinite. Furthermore $f(\mu, \nu)$ may vanish even if $\mu \neq \nu$. To exclude this, further conditions on f have to be imposed, for example as in the following

2.1. *Lemma*. Suppose there is an $a \in \mathbb{R}$ so that the function

$$g(x) := f(x) - a(x - 1)$$

is non-negative and vanishes only if $x = 1$, then $f(\mu, \nu)$ vanishes only if $\mu = \nu$.

Proof. The function $g(x)$ is convex as well. Furthermore $D_f(\mu,\nu) = D_g(\mu,\nu)$. But since g is non-negative,

$$D_g(\mu,\nu) = \int g(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\mathrm{d}\nu$$

can only vanish if $g(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})$ is identical to zero, which implies that $\frac{\mathrm{d}\mu}{\mathrm{d}\nu} = 1$ ν-a.s. But this means $\mu = \nu$. □

The concept of f-divergences was introduced by Csiszár [1], who also noted the result in Lemma 2.1. Common choices for f are

$(\sqrt{x}-1)^2$	Hellinger divergence HE
$\lvert x-1\rvert$	total–variation divergence TV
$x\log(x)$	Kullback–Leibler divergence KL
$(x-1)^2$	Pearson divergence PE

The transformation $f^*(x) = xf(1/x)$ yields a divergence D_{f^*} which is equal to D_f but with interchanged arguments. Applying this transformation to the Kullback–Leibler divergence for example, we get a divergence which is also sometimes referred to as the Kullback–Leibler divergence, or alternatively as the Shannon divergence SH. The total variation divergence plays a central role, since all f–divergences allow for an estimate against TV, as will be shown in the following proposition, which forms the main result of this short note.

2.2. *Proposition*. For two probability measures μ, ν, it holds in general that

$$f(1+\frac{1}{2}\mathsf{TV}(\mu,\nu)) + f(1-\frac{1}{2}\mathsf{TV}(\mu,\nu)) \leq D_f(\mu,\nu).$$

Proof. The proof of this fact is a generalisation of the method used in [3] to prove the special case of the KL divergence. Since $f(1) = 0$, we have the general property that

$$f(x) = f(\max\{x,1\}) + f(\min\{x,1\}).$$

Using this fact and the convexity of f we get the general estimate

$$\begin{aligned} D_f(\mu,\nu) &= \int f(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\mathrm{d}\nu \\ &= \int f(\max\{\frac{\mathrm{d}\mu}{\mathrm{d}\nu},1\})\mathrm{d}\nu + \int f(\min\{\frac{\mathrm{d}\mu}{\mathrm{d}\nu},1\})\mathrm{d}\nu \\ &\geq f(\int \max\{\frac{\mathrm{d}\mu}{\mathrm{d}\nu},1\}\mathrm{d}\nu) + f(\int \max\{\frac{\mathrm{d}\mu}{\mathrm{d}\nu},1\}\mathrm{d}\nu). \end{aligned}$$

Now use that

$$\begin{aligned} \max\{x,1\} &= \frac{1+x+\lvert 1-x\rvert}{2} \\ \min\{x,1\} &= \frac{1+x-\lvert 1-x\rvert}{2} \end{aligned}$$

to complete the theorem. □

Recalling that always $\mathsf{TV} \leq 2$, the proposition rises the question as to when the function $f(1+x) + f(1-x)$ is monotonous on $x \in [0,1]$. The following lemma partially answers this.

2.3. *Lemma*. Under the conditions of Lemma 2.1, the function $\phi(x) = f(1+x) + f(1-x)$ is strictly monotonous on $x \in [0,1]$.

Proof. The conditions imply that $\phi(0) = 0$, $\phi(x) > 0$ for $x > 0$, and that ϕ is convex. Let $0 \leq x_1 < x_2 \leq 1$. For any $\tau \in]0,1[$,

$$(1-\tau)\phi(0) + \tau\phi(x_2) > \phi((1-\tau)0 + \tau x_2)$$

which obviously implies $\phi(x_2) > \tau\phi(x_2) > \phi(\tau x_2)$ (since $\tau \in]0,1[$). Now take $\tau = x_1/x_2$ to get the result. □

As a corollary of Proposition 2.2, we get the following well known estimates between TV and KL

2.4. *Corollary (Bretagnole–Huber and Furstemberg inequality).*

$$\mathsf{TV}(\mu,\nu) \leq 2\sqrt{1-\exp\left(-\mathsf{SH}(\mu,\nu)\right)} \leq 2\sqrt{\mathsf{SH}(\mu,\nu)}$$

Recall that $\mathsf{SH}(\mu,\nu) = \mathsf{KL}(\nu,\mu)$. A further useful estimate concerns the Hellinger divergence

2.5. *Corollary*. For the Hellinger divergence HE, the estimate

$$\mathsf{TV} \leq \begin{cases} 2 - 2\left(1-\sqrt{\mathsf{HE}}\right)^2 & \text{if } \mathsf{HE} < 1 \\ 2 & \text{otherwise} \end{cases} \tag{8}$$

holds.

Proof. Theorem 2.2 gives the inequality

$$\mathsf{HE} \geq \left(\sqrt{1+\frac{1}{2}\mathsf{TV}} - 1\right)^2 + \left(\sqrt{1-\frac{1}{2}\mathsf{TV}} - 1\right)^2. \tag{9}$$

The right hand side of Equation (9) is larger than $\left(\sqrt{1-\frac{1}{2}\mathsf{TV}} - 1\right)^2$, whence

$$\mathsf{HE} \geq \left(\sqrt{1-\frac{1}{2}\mathsf{TV}} - 1\right)^2,$$

which, after solving for TV, yields the result. □

References

[1] Imre Csiszar. Information-type measues of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[2] Joseph L. Doob. *Measure Theory.* Springer, 1994.

[3] Vladimir N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.