Available online at www.sciencedirect.com
SCIENCE DIRECT®
Nonlinear Analysis 61 (2005) 309–339
Nonlinear Analysis
www.elsevier.com/locate/na

Higher order optimal approximation of Csiszar's f-divergence

George A. Anastassiou

Department of Mathematical Sciences, University of Memphis, Memphis, TN 38152 USA

Received 30 June 2004; accepted 20 September 2004

Abstract

Here are established various sharp and nearly optimal probabilistic inequalities giving the high order approximation of Csiszar's *f*-divergence between two probability measures, which is the most essential and general tool for their comparison. The above are done through Taylor's formula, generalized Taylor–Widder's formula, an alternative recent expansion formula. Based on these we give many representation formulae of Csiszar's distance, then we estimate in all directions their remainders by using either the norms approach or the modulus of continuity way. Most of the last probabilistic estimates are sharp or nearly sharp, attained by basic simple functions.

MSC: 26B40; 26D15; 28A25; 41A17; 41A25; 60B10; 60E15

Keywords: Csiszar's discrimination; Csiszar's distance; Representation; Rate of approximation; Probabilistic inequality

1. Basics

Throughout this article we use the following. Let f be a convex function from $(0, +\infty)$ into $\mathbb{R}$ which is strictly convex at 1 with $f(1) = 0$.

Let $(X, \mathscr{A}, \lambda)$ be a measure space, where λ is a finite or a σ-finite measure on $(X, \mathscr{A})$. And let μ_1, μ_2 be two probability measures on $(X, \mathscr{A})$ such that $\mu_1 \ll \lambda$, $\mu_2 \ll \lambda$ (absolutely continuous), e.g. $\lambda = \mu_1 + \mu_2$.

E-mail address: ganastss@memphis.edu (G.A. Anastassiou).

0362-546X/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.na.2004.09.049

Denote by $p=\frac{d\mu_1}{d\lambda}$, $q=\frac{d\mu_2}{d\lambda}$ the (densities) Radon–Nikodym derivatives of μ_1, μ_2 with respect to λ. Here we assume that

$$0<a\leqslant \frac{p}{q}\leqslant b,\ \text{ a.e. on } X \text{ and } a\leqslant 1\leqslant b.$$

The quantity

$$\Gamma_f(\mu_1,\mu_2)=\int_X q(x)f\left(\frac{p(x)}{q(x)}\right)\mathrm{d}\lambda(x), \tag{1}$$

was introduced by I. Csiszar in 1967, see [8], and is called *f-divergence* of the probability measures μ_1 and μ_2.

By Lemma 1.1 of [8], the integral (1) is well-defined and $\Gamma_f(\mu_1,\mu_2)\geqslant 0$ with equality only when $\mu_1=\mu_2$. Furthermore $\Gamma_f(\mu_1,\mu_2)$ does not depend on the choice of λ. The concept of *f*-divergence was introduced first in [7] as a generalization of Kullback's "*information for discrimination*" or *I-divergence* (*generalized entropy*) [11,10] and of Rényi's "*information gain*" (*I-divergence of order* α) [12]. In fact the *I-divergence of order* 1 equals

$$\Gamma_{u\log_2 u}(\mu_1,\mu_2).$$

The choice $f(u)=(u-1)^2$ produces again a known measure of difference of distributions that is called χ^2*-divergence*. Of course the *total variation* distance

$$|\mu_1-\mu_2|=\int_X |p(x)-q(x)|\,\mathrm{d}\lambda(x)$$

is equal to $\Gamma_{|u-1|}(\mu_1,\mu_2)$. Here by assuming $f(1)=0$ we can consider $\Gamma_f(\mu_1,\mu_2)$, the *f*-divergence as a measure of the difference between the probability measures μ_1, μ_2.

The *f*-divergence is in general asymmetric in μ_1 and μ_2. But since f is convex and strictly convex at 1 so is

$$f^*(u)=uf\left(\frac{1}{u}\right) \tag{2}$$

and as in [8] we get

$$\Gamma_f(\mu_2,\mu_1)=\Gamma_{f^*}(\mu_1,\mu_2). \tag{3}$$

In Information Theory and Statistics many other divergences are used which are special cases of the above general Csiszar *f*-divergence, e.g. Hellinger distance D_{H}, α-divergence D_α, Bhattacharyya distance D_{B}, Harmonic distance D_{Ha}, Jeffrey's distance D_{J}, triangular discrimination D_Δ, for all these see, e.g. [5,9].

The problem of finding and estimating the *proper distance* (*or difference or discrimination*) of two probability distributions is one of the major ones in Probability Theory. The above *f*-divergence measures in their various forms have been also applied to Anthropology, Genetics, Finance, Economics, Political Science, Biology, Approximation of Probability distributions, Signal Processing and Pattern Recognition.

A great inspiration for this paper has been the very important monograph on the topic by Dragomir [9].