CCIT Report #855 March 2014

On Improved Bounds for Probability Metrics and f-Divergences

Igal Sason

Abstract

Derivation of tight bounds for probability metrics and f-divergences is of interest in information theory and statistics. This paper provides elementary proofs that lead, in some cases, to significant improvements over existing bounds; they also lead to the derivation of some existing bounds in a simplified way. The inequalities derived in this paper relate between the Bhattacharyya parameter, capacitory discrimination, chi-squared divergence, Chernoff information, Hellinger distance, relative entropy, and the total variation distance. The presentation is aimed to be self-contained.

Index Terms – Bhattacharyya parameter, capacitory discrimination, Chernoff information, chi-squared divergence, f-divergence, Hellinger distance, relative entropy, total variation distance.

I. Introduction

Derivation of tight bounds for probability metrics and f-divergences is of interest in information theory and statistics, as is reflected from the bibliography of this paper and references therein. Following previous work in this area, elementary proofs are used in this paper for the derivation of bounds. In some cases, existing bounds are re-derived in a simplified way, and in some others, significant improvements over existing bounds are obtained.

The paper is structured as follows: the bounds and their proofs are introduced in Section II, followed by various discussions and remarks that link the new bounds to the literature. This section is separated into four parts: the first part introduces bounds on the Hellinger distance and Bhattacharyya parameter in terms of the total variation distance and the relative entropy (see Section II-A), the second part introduces a lower bound on the Chernoff information in terms of the total variation distance (see Section II-B), the third part provides bounds on the chi-squared divergence and some related inequalities on the relative entropy and total variation distance (see Section II-C), and the last part considers bounds on the capacitory discrimination (see Section II-D). A summary, which outlines the contributions made in this work, is provided in Section III.

Preliminaries

We introduce, in the following, some preliminary material that is essential to make the presentation self-contained.

Definition 1: Let f be a convex function defined on $(0,\infty)$ with $f(1)=0$, and let P and Q be two probability distributions defined on a common set $\mathcal{X}$. The *f-divergence* of P from Q is defined by

$$D_f(P||Q) \triangleq \sum_{x\in\mathcal{X}} Q(x)\, f\left(\frac{P(x)}{Q(x)}\right) \tag{1}$$

where sums may be replaced by integrals. Here we take

$$0f\left(\frac{0}{0}\right)=0, \quad f(0)=\lim_{t\to 0^+} f(t), \quad 0f\left(\frac{a}{0}\right)=\lim_{t\to 0^+} tf\left(\frac{a}{t}\right)=a\lim_{u\to\infty}\frac{f(u)}{u},\ \forall\, a>0.$$

Definition 2: An f-divergence is said to be *symmetric* if the equality $f(x)=xf\left(\frac{1}{x}\right)$ holds for every $x>0$. This requirement on f implies that $D_f(P||Q)=D_f(Q||P)$ for every pair of probability distributions P and Q.

From [13] and [15, Corollary 5.4], the following lower bound holds for a symmetric f-divergence:

$$D_f(P||Q) \geq \big(1-d_{\mathrm{TV}}(P,Q)\big)\, f\left(\frac{1+d_{\mathrm{TV}}(P,Q)}{1-d_{\mathrm{TV}}(P,Q)}\right). \tag{2}$$

Definition 3: Let P and Q be two probability distributions defined on a set $\mathcal{X}$. The *total variation distance* between P and Q is defined by

$$d_{\mathrm{TV}}(P,Q) \triangleq \sup_{\text{Borel } A\subseteq\mathcal{X}} |P(A)-Q(A)| \tag{3}$$

where the supremum is taken over all the Borel subsets A of $\mathcal{X}$.

I. Sason is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: sason@ee.technion.ac.il).