

# Normalized Forms for Two Common Metrics

Peter N. Yianilos\*

12/5/91 Rev 2/27/92 Rev 7/7/2002

*Abstract* — We show that symmetric set difference and Euclidian distance have particular [0–1] normalized forms that remain metrics.

The first of these  $|A\Delta B|/|A\cup B|$  is easily established and generalizes to measure spaces.

The second applies to vectors in  $\mathbb{R}^n$  and is given by  $\|X-Y\|/(\|X\|+\|Y\|)$ . That this is a metric is more difficult to demonstrate and is true for Euclidian distance (the  $L_2$  norm) but for no other integral Minkowski metric. The short and elegant proof we give is due to David Robbins and Marshall Buck [1].

We also explore a number of variations.

*Keywords* — Metric Space, Distance Function, Similarity Function/Coefficient, Euclidian Distance, Association, Clustering, Vector Quantization, Pattern Recognition, Statistical Methods.

## I. INTRODUCTION

The notion of Metric Space [2], is a cornerstone of mathematical topology and analysis, and is often employed in pattern recognition and clustering systems.

**Definition 1** Let  $X$  be a set and  $d$  a non-negative real valued function on  $X \times X$  satisfying for  $a, b, c \in X$ :

1.  $d(a, b) = d(b, a)$
2.  $d(a, b) = 0$  iff  $a = b$
3.  $d(a, b) + d(b, c) \geq d(a, c)$

Then  $(X, d)$  is a metric space. Alternatively  $d$  is said to be a distance function and impose a metric on  $X$ .

The third item in this definition is usually referred to as the *triangle inequality*. It is worthwhile noting that not all approaches to pattern recognition and clustering impose this requirement. The term *similarity measure* or *dissimilarity measure*, or some variation thereof, is then used to describe the comparison function.

Two well known examples of metric spaces are Euclidian  $n$ -space with:

$$d(A, B) = \|A - B\|_2 = \sqrt{\sum_i^n (A_i - B_i)^2}$$

\*This is a revision of an earlier unpublished Technical Memorandum of NEC Research Institute, 4 Independence Way, Princeton, NJ 08540

and the symmetric difference metric on members of  $\mathcal{F}$ , the set consisting of all finite sets:

$$d_\Delta(A, B) = |(A \setminus B) \cup (B \setminus A)| = |A\Delta B|$$

$d(A, B)$  is of course geometrical *length*, and  $d_\Delta(A, B)$  simply counts the number of elements on which  $A$  and  $B$  disagree.

Both of these metrics are in general unbounded, and their value is independent of the *size* of  $A$  and  $B$ . I.e. two very large sets that have three points of disagreement, are just as distant under  $d_\Delta$  as two very small sets which also differ by three elements. In Euclidian 1-space, values 1,000 and 1,000.01 are just as distant as 0.01 and 0.02.

Hence we may in a sense view  $d$  and  $d_\Delta$  as measures of *absolute* distance or difference.  $d_\Delta$  is seen to be insensitive to  $|A \cap B|$  while  $d$  does not independently consider  $\|A\|$  and  $\|B\|$  when measuring distance.

It is therefore natural to consider forming *relative* distance measures for these underlying spaces, since such measures may be more effective in the solution of certain problems. Certainly the notion of relative error is an important one in numerical analysis. For sets, and Euclidian space, *size* might naturally be taken to mean cardinality, and norm respectively. Thus we are interested in set metrics which measure *relative* to the empty set, and in an alternative to the Euclidian distance metric which measures *relative* to the origin.

With no domain assumptions,  $d$  and  $d_\Delta$  are unbounded. This may create algorithmic difficulty (or at a minimum inconvenience). Thus converting these metrics to bounded forms may sometimes be useful.

The simplest way in general to effect a bound is to compose the metric with another function  $f$  which acts as a range *componder* such that the combination still satisfies definition 1. One well known example of such a function that bounds any metric to  $[0, 1]$  is:

$$\bar{d}(A, B) = \frac{d(A, B)}{1 + d(A, B)}$$

There are many such formulas for bounding metrics. It may for example be shown that the sum of metrics is a metric, and beyond this that a metric results from composition with any continuous, differentiable, strictly increasing function  $f$  such that  $f(0) = 0$ , and  $f'$  is non-increasing.<sup>1</sup>

Since however any such method depends only on the value  $d(A, B)$ , the bounded form will inherit the absolute or relative behavior of its unbounded parent.

<sup>1</sup>The key is that  $f(a + b) \leq f(a) + f(b)$ .

These metric *companding* methods are in contrast to *normalization* to which we now turn.

In the sections that follow we will introduce the metric  $d_{\Delta_n}$  defined by  $d_{\Delta}(A, B)/|A \cup B|$  and also metric  $d_n$  defined <sup>2</sup> by  $d(A, B)/(\|A\| + \|B\|)$ .

In contrast to the *absolute* behavior of  $d$  and  $d_{\Delta}$  these functions judge distance with consideration to the relative location of the origin and empty set respectively.

That they are in fact metrics is not obvious and a considerable portion of this paper is devoted to the required proofs.

We will also present a number of alternative forms including mixed metrics that combine absolute and relative behavior.

## II. NORMALIZED SYMMETRIC DIFFERENCE

To normalize the symmetric difference metric we choose to divide its value by the size of the union of its arguments. More formally we have:

**Definition 2** Let  $\mathcal{F}$  be the set consisting of all finite sets We define function  $d_{\Delta_n}$  on  $\mathcal{F} \times \mathcal{F}$  by:

$$d_{\Delta_n}(A, B) = \begin{cases} \frac{|A \Delta B|}{|A \cup B|} & \text{if } A \cup B \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

We would point out here that not every reasonable attempt at normalization results in a metric. Dividing instead by  $|A| + |B|$ , an altogether sensible thing to do, fails to satisfy the triangle inequality. This may be seen by letting  $A = \{x\}$ ,  $B = \{x, y\}$ , and  $C = \{y\}$ . The  $AB$  and  $BC$  distances are then both  $1/3$  while the distance  $AC$  is 1. <sup>3</sup>

While it is clear that this function is bounded by  $[0, 1]$ , and that its behavior is more *relative* in that  $|A \cap B|$  will affect its value, it must be demonstrated that  $d_{\Delta_n}$  is in fact a metric.

**Theorem 1**  $(\mathcal{F}, d_{\Delta_n})$  is a metric space.

*Proof:* We must show that for all finite sets  $A$ ,  $B$ , and  $C$ , the following are true:

- i)  $d_{\Delta_n}(A, B) = 0 \Leftrightarrow A = B$
- ii)  $d_{\Delta_n}(A, B) = d_{\Delta_n}(B, A)$
- iii)  $d_{\Delta_n}(A, C) \leq d_{\Delta_n}(A, B) + d_{\Delta_n}(B, C)$

We have *i*) because  $A \cap B = A \cup B$  only when  $A = B$ . The second requirement *ii*) is clear from the commutativity of basic set operations.

Item *iii*), the *triangle inequality*, requires more work. Using EQ 1, we must show:

$$1 - \frac{|A \cap B|}{|A \cup B|} + 1 - \frac{|B \cap C|}{|B \cup C|} \geq 1 - \frac{|A \cap C|}{|A \cup C|} \quad (2)$$

<sup>2</sup>Both of these metrics are defined to have zero value if  $A = B$ .

<sup>3</sup>This example noticed by S.R. Buss who also pointed out that this definition satisfies a weakened triangle inequality.

It is easy to verify that this is true if  $A \cup B = \emptyset$ ,  $B \cup C = \emptyset$ , or  $A \cup C = \emptyset$ . So we restrict our attention to the case in which none of the denominators in EQ 2 is zero.

Now the union of sets  $A, B$ , and  $C$ , may be partitioned (see figure 1) into seven disjoint subsets whose orders we denote:

$$\begin{aligned} a &= |A \setminus (B \cup C)| \\ b &= |B \setminus (A \cup C)| \\ c &= |C \setminus (A \cup B)| \\ ab &= |(A \cap B) \setminus (A \cap B \cap C)| \\ bc &= |(B \cap C) \setminus (A \cap B \cap C)| \\ ac &= |(A \cap C) \setminus (A \cap B \cap C)| \\ abc &= |A \cap B \cap C| \end{aligned}$$

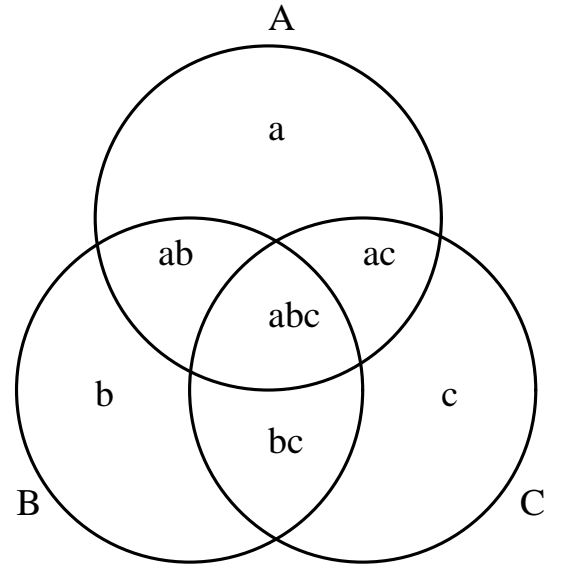


Figure 1: Partitioning  $A \cup B \cup C$

For later convenience we also define  $\gamma = ab + ac + bc + abc$ .

With these definitions, EQ 2 becomes with some simplification:

$$\frac{ab + abc}{a + b + \gamma} + \frac{bc + abc}{b + c + \gamma} \leq \frac{ac + abc}{a + c + \gamma} + 1 \quad (3)$$

Now observe that  $b$  may be removed without loss of generality from the left denominator, since its removal can only increase that side. It will then suffice to show that:

$$\frac{ab + abc}{a + \gamma} + \frac{bc + abc}{c + \gamma} \leq \frac{ac + abc}{a + c + \gamma} + 1$$

Now replacing 1 with  $\gamma/\gamma$  and adding fractions, we arrive at:

$$\frac{(c + \gamma)(ab + abc) + (a + \gamma)(bc + abc)}{(a + \gamma)(c + \gamma)} \leq \quad (4)$$

$$\frac{(ac + abc)\gamma + (a + c + \gamma)\gamma}{(a + c + \gamma)\gamma} \quad (5)$$

The denominator of EQ 4 is equal to the denominator in EQ 5 plus  $ac$ , and is therefore no smaller.

It therefore suffices to show that the inequality is true for the numerators alone, i.e. that:

$$(c + \gamma)(ab + abc) + (a + \gamma)(bc + abc) \leq (ac + abc)\gamma + (a + c + \gamma)\gamma$$

Starting from the left side of this inequality we have:

$$\begin{aligned} (c + \gamma)(ab + abc) + (a + \gamma)(bc + abc) &\leq \\ c\gamma + \gamma(ab + abc) + a\gamma + \gamma(bc + abc) &= \\ c\gamma + \gamma(ab + bc + abc + abc) + a\gamma &\leq \\ c\gamma + \gamma^2 + \gamma abc + a\gamma &\leq \\ (ac + abc)\gamma + (a + c + \gamma)\gamma & \end{aligned}$$

and we are done.  $\square$

Our arguments for sets in  $\mathcal{F}$  generalize easily to measure spaces[3, 4]. The case of finite non-zero measures is straightforward since these correspond well with finite sets and our earlier arguments.

To see this, remember that for any  $A, B, C \in \mathcal{F}$ , we earlier expressed the triangle inequality in terms of the orders of their natural decomposition into seven disjoint parts. Our proof of theorem 1 may then be viewed as establishing the truth of an inequality in these seven independent variables – an inequality that holds for any assignment of finite non-negative values; subject only to the restriction that the denominators may not vanish. The same decomposition applies if  $A, B, C \in \mathcal{M}$ , with the inequality relating measure instead of set order, thus motivating the connection between  $d_{\Delta_n}$  and measure spaces.

As a practical matter the finite setting is the most important case. However the generalization extends fully to measures which assume value  $+\infty$ , and we will take the time to show this.

For elements of finite non-zero measure, definition of our metric will correspond to simple generalization of the notion of set order, to that of measure. For elements in general, several cases are necessary to patch together a definition. While these special cases manage to define the metric for all members of the space, only the simplest discrete distance notion applies to elements with zero or infinite measure.

**Definition 3** Let  $\mathcal{M} = (X, \mathcal{B}, \mu)$ , be a measure space, and  $A, B \in \mathcal{B}$ . We define function  $d_\mu$  on  $\mathcal{M} \times \mathcal{M}$  by:

$$d_\mu(A, B) = \begin{cases} \frac{\mu(A \Delta B)}{\mu(A \cup B)} & : \mu(A \cup B) \neq 0, \mu(A \cup B) \neq \infty \\ 0 & : \mu(A \cup B) = 0, A = B \\ 1 & : \mu(A \cup B) = 0, A \neq B \\ 0 & : \mu(A \cup B) = \infty, A = B \\ 1 & : \mu(A \cup B) = \infty, A \neq B \end{cases} \quad (6)$$

We now state the corollary:

**Corollary 1**  $(\mathcal{M}, d_\mu)$  is a metric space.

*Proof:* Everything but the triangle inequality follows immediately from the definition.

Note first that if  $d_\mu(A, C) = 0$  or either one of  $d_\mu(A, B)$ ,  $d_\mu(B, C)$  is one, then the triangle inequality is trivially established. So in particular the inequality is satisfied if  $A = B = C$ .

Further observe that if  $\mu(A)$  or  $\mu(B)$  is infinite or zero, then  $d_\mu(A, B) = 0$  iff  $A = B$ , and one otherwise. I.e. the definition reduces to a simple equality test for these cases.

With these points in mind we distinguish three cases:

1.  $\mu(A)$ ,  $\mu(B)$ ,  $\mu(C)$  are finite with no two zero: Interpreting set order  $|\cdot|$  instead as measure  $\mu(\cdot)$ , the denominators of EQ 2 are all non-zero and our proof of theorem 1 applies.

2.  $\mu(A)$ ,  $\mu(B)$ ,  $\mu(C)$  are finite and at least two are zero: If  $\mu(A) = \mu(C) = 0$  and  $A = C$ , then  $d_\mu(A, C) = 0$ . On the other hand  $A \neq C$  implies  $A \neq B$  or  $B \neq C$  so that either  $d_\mu(A, B) = 1$  or  $d_\mu(B, C) = 1$ .

Otherwise we may assume without loss of generality that  $\mu(A) = \mu(B) = 0$ . Here  $A \neq B$  implies  $d_\mu(A, B) = 1$  – and if  $A = B$ , then  $d_\mu(B, C) = 1$  unless  $A = B = C$ .

3. At least one of  $\mu(A)$ ,  $\mu(B)$ ,  $\mu(C)$  is infinite: If  $\mu(B) = \infty$  then  $d_\mu(A, B) = 1$  or  $d_\mu(B, C) = 1$ , unless  $A = B = C$ .

Otherwise we may assume without loss of generality that  $\mu(A) = \infty$ . But then  $d_\mu(A, B) = 1$  unless  $A = B$ , in which case  $\mu(B) = \infty$ ; a situation covered by the preceding argument.  $\square$

Proceeding further we can develop a slightly more general form for  $d_\mu$ . Consider some  $A, B, C \in \mathcal{M}$  and their associated decomposition. Now for some  $\alpha \geq 0$ , adding  $\alpha/2$  to each of  $a$ ,  $b$ , and  $c$ , leaves them non-negative and the inequality therefore valid.

This may be thought of as extending each of  $A$ ,  $B$ , and  $C$ , by a new region which does not intersect the others.

Equation 3 then becomes:

$$\frac{ab + abc}{a + b + \gamma + \alpha} + \frac{bc + abc}{b + c + \gamma + \alpha} \leq \frac{ac + abc}{a + c + \gamma + \alpha} + 1 \quad (7)$$

This all motivates:

**Definition 4** Let  $\mathcal{M} = (X, \mathcal{B}, \mu)$  be a measure space, and  $A, B \in \mathcal{B}$ . We define function  $d_{\mu_\alpha}$  on  $\mathcal{M} \times \mathcal{M}$  by:

$$d_{\mu_\alpha}(A, B) = \begin{cases} \frac{\mu(A \Delta B)}{\mu(A \cup B) + \alpha} & : \mu(A \cup B) \neq 0, \mu(A \cup B) \neq \infty \\ 0 & : \mu(A \cup B) = 0, A = B \\ 1 & : \mu(A \cup B) = 0, A \neq B \\ 0 & : \mu(A \cup B) = \infty, A = B \\ 1 & : \mu(A \cup B) = \infty, A \neq B \end{cases} \quad (8)$$

for  $\alpha \geq 0$ .

And from our discussion above it follows that:

**Corollary 2**  $(\mathcal{M}, d_{\mu_\alpha})$  is a metric space.

Thus we see that the  $d_\mu$  discontinuity when  $A = B$  may be eliminated <sup>4</sup> by *biasing* the denominator term  $|A \cup B|$ . Function  $d_{\mu_0}$  is just  $d_\mu$ , and as  $\alpha \rightarrow \infty$ , the behavior of  $d_{\mu_\alpha}$  approaches that of simple symmetric difference.

In a later section we will present several examples of metrics that arise from the results above. Before doing so however, we turn our attention back to Euclidian space and our goal of establishing a normalized metric there.

### III. THE NORMALIZED EUCLIDIAN METRIC

In this section we will demonstrate that Euclidian distance normalized by combined norm, defines an alternative and  $[0, 1]$  bounded metric on  $\mathbb{R}^n$ .

**Definition 5** Let  $(\mathbb{R}^n, d)$  denote Euclidian  $N$ -space (The  $L_2$  norm and corresponding standard distance function). Then the following  $[0,1]$  bounded function  $d_n$  of any two vectors  $X, Y$  is defined to be the Normalized Euclidian Distance between  $X$  and  $Y$ :

$$d_n(X, Y) = \begin{cases} \frac{\|X-Y\|}{\|X\|+\|Y\|} & X \neq 0 \text{ or } Y \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

That this definition is in fact a metric will later be stated as a theorem. Its proof will require a series of Lemmas which build insight into the function.

It is worthwhile noting that in  $\mathbb{R}^1$ , an easy proof exists and that a very similar  $\mathbb{R}^1$  metric (a special case of our definition 6), is employed in [5].

We would also observe that this definition fails to generate a metric if the  $L_1$  norm is substituted. <sup>5</sup> For if  $A = (1, 0)$ ,  $B = (1, 1)$ , and  $C = (0, 1)$ , the triangle inequality does not hold. The  $L_\infty$  norm also fails; consider  $A = (1, -1)$ ,  $B = (2, 0)$ , and  $C = (1, 1)$ . For this example, and the  $L_p$  norm, the triangle inequality becomes:

$$\frac{2 \cdot 2^{1/p}}{2^{1/p} + 2} \geq \frac{1}{2^{1/p}}$$

This fails for  $p > 2$ . Therefore, we have that our definition does not generate a metric for any of the integral Minkowski metrics  $L_p$  where  $p \neq 2$ .

Our to-be-proven positive result for  $L_2$  is thus that much more interesting.

**Theorem 2**  $(\mathbb{R}^n, d_n)$  is a metric space.

*Proof:*

Let  $O, A, B, C$  be four distinct points in Euclidean space. Then

$$\frac{AB}{OA+OB} + \frac{BC}{OB+OC} \geq \frac{AC}{OA+OC}.$$

<sup>4</sup>The same applies for  $d_{\Delta_n}$ .

<sup>5</sup>This may also be seen as a result of our earlier comments regarding other methods for normalizing  $d_\Delta$ .

proof: Let  $OA = a$ ,  $OB = b$ ,  $OC = c$ ,  $BC = \alpha$ ,  $AC = \beta$ ,  $BC = \gamma$ . Then we need to prove that

$$\alpha a^2 - \beta b^2 + \gamma c^2 + (\alpha - \beta + \gamma)(ab + bc + ac) \geq 0.$$

Note that we have  $\alpha - \beta + \gamma \geq 0$  by the triangle inequality and  $\alpha a - \beta b + \gamma c \geq 0$  by Ptolemy's inequality. We may suppose that  $a \leq c$ .

Case 1:  $b \leq a$ . Then we have

$$\alpha a^2 - \beta b^2 + \gamma c^2 \geq (\alpha - \beta + \gamma)a^2 \geq 0$$

and the theorem follows immediately.

Case 2:  $a \leq b \leq c$ . Then we have

$$\begin{aligned} & \alpha a^2 - \beta b^2 + \gamma c^2 + (\alpha - \beta + \gamma)(ac) & (10) \\ = & (a+c)(\alpha a - \beta b + \gamma c) - \beta(b-a)(b-c) \geq 0 & (11) \end{aligned}$$

from which the result follows immediately as well.

Case 3:  $c \leq b$ . Observe that the theorem is true for four points  $O, A, B, C$  if and only if it is true for  $O$  and the points  $A', B', C'$  obtained by inverting  $A, B$  and  $C$  in the sphere of radius 1 centered at  $O$ . (Recall that  $A$  and  $A'$  are on the same ray from  $O$  and satisfy  $OA \cdot OA' = 1$ . The key fact that is needed is  $A'B' = AB/(OA \cdot OB)$ .) This reduces case 3 to case 1.  $\square$

If we now label the origin as  $D$ , we may re-state the theorem as a corollary in purely geometrical terms:

**Corollary 3** Let  $A, B, C, D$  be four points in Euclidian Space and  $ab, ac, ad, bc, bd, cd$ , denote their interpoint segment lengths. Then:

$$(bd + cd)(ad + cd)ab + (ad + bd)(ad + cd)bc \geq (ad + bd)(bd + cd)ac$$

So despite a singularity at the origin,  $d_n$  imposes a metric on  $\mathbb{R}^n$ . This amounts to a *fully relative* distance measure which clearly cannot well cope with zero. By contrast, standard Euclidian distance may be viewed as a *fully absolute* distance measure. It is easy to imagine a continuum of intermediate metrics which are *partially relative*.

Furthermore, some applications, while requiring a modicum of *relative* behavior, may also require better behavior at the origin than  $d_n$  provides.

All of this motivates the following definition:

**Definition 6** Let  $(\mathbb{R}^n, d)$  denote Euclidian  $N$ -space. Then for values  $\alpha, \beta \geq 0$ , with  $\alpha \neq 0$  or  $\beta \neq 0$ , the following  $[0, 1/\beta]$  bounded function  $d_{n_{\alpha\beta}}$  of any two vectors  $X, Y$  is defined to be the  $\alpha\beta$ -Normalized Euclidian Distance between  $X$  and  $Y$ :

$$d_{n_{\alpha\beta}}(X, Y) = \begin{cases} \frac{\|X-Y\|}{\alpha+\beta(\|X\|+\|Y\|)} & X \neq 0 \text{ or } Y \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

The Normalized Euclidian Distance of definition 9 corresponds to  $\alpha = 0, \beta = 1$ . Note that the definition 6's special condition at  $X = Y = 0$  is only required when  $\alpha = 0$ , for when  $\alpha > 0$ , the denominator cannot vanish.

The case  $\beta = 0$ , merely corresponds to standard Euclidian distance scaled by  $\alpha$ .

In between these two extremes lies a family of hybrid functions that combine the relative behavior of  $d_n$  with the absolute behavior of  $d$ .

To establish that  $(\mathbb{R}^n, d_{n\alpha\beta})$  is in general a metric space, we will require one lemma concerning the behavior of quasi-quadrilaterals in  $\mathbb{R}^n$ .

**Lemma 1** *Let  $A, B, C, D$  be any four points in  $\mathbb{R}^n$ , then with respect to the six interpoint segment lengths:  $ab, ac, ad, bc, bd, cd$ , the following is true:*

$$ab \cdot cd + bc \cdot ad \geq bd \cdot ac$$

*I.e. the sum of the product of the lengths of opposite sides of an arbitrary quadrilateral, is no smaller than the product of its diagonals.*

*Proof:* We have only to argue that we may reduce setting to  $\mathbb{R}^2$  since there, the lemma is recognized as a standard theorem of inversive geometry ([6] Theorem 5.12) – a direct generalization of Ptolemy's theorem.

Since any four points may be embedded in  $\mathbb{R}^3$ , we may assume our setting is at most this.

Within  $\mathbb{R}^3$ , choose coordinates so that  $A, C, D$  lie on the  $XY$  plane, and  $B$  is located above (or on) it. I.e.  $A_z = C_z = D_z = 0$  and  $B_z >= 0$ .

Now translate so that  $D$  is the origin.

Further choose rotation and orientation so that line  $AC$  is parallel to and above the  $X$ -axis with  $A$  left of  $C$ . I.e.  $A_y = C_y$  and  $A_x \leq C_x$ .

Now consider spheres  $S(A, ab)$  and  $S(C, bc)$ . The intersection of these spheres is a circle perpendicular to the  $XY$  plane and parallel to the  $Y$ -axis, having center along  $AC$ .

Any choice for  $B$  along on this circle it will leave everything unchanged in our inequality but for  $bd$ .

Notice that the point on this circle farthest from the origin, is just its intersection with the  $XY$  plane.

Denoting this point  $B'$  and relocating  $B$  there, increases  $bd$  and and the inequality's right side, while leaving the left unchanged.

Thus it is clear that the inequality is true for points  $A, B, C, D$  in  $\mathbb{R}^3$  iff it is true for  $A, B', C, D$  in  $\mathbb{R}^2$ .  $\square$

We are now ready to establish that  $d_{n\alpha\beta}$  is a metric:

**Corollary 4**  $(\mathbb{R}^n, d_{n\alpha\beta})$  is a metric space.

*Proof:* The first and second metric conditions are straightforward. We therefore turn to the triangle inequality.

If  $\beta = 0$ , then by definition  $\alpha \neq 0$  and  $d_{n\alpha\beta}$  is nothing more than scaled Euclidian distance. If  $\alpha = 0$ , then by definition  $\beta \neq 0$  and  $d_{n\alpha\beta}$  is just scaled normalized Euclidian distance.

So the only interesting case is that in which  $\alpha, \beta \neq 0$ . Here, we may assume without loss of generality that  $\beta = 1$  since other cases may be reduced to this one by multiplying each term in the triangle inequality by  $\beta$  and then reducing.

Now defining for convenience:  $a = \|A\|, b = \|B\|, c = \|C\|$ , it remains for us to show that:

$$\frac{d(A, B)}{\alpha + (a + b)} + \frac{d(B, C)}{\alpha + (b + c)} \geq \frac{d(A, C)}{\alpha + (a + c)}$$

Now forming a common denominator, discarding it, and then rearranging, we have:

$$\left. \begin{aligned} &\alpha^2 d(A, B) + \alpha^2 d(B, C) + \\ &(a+c)(b+c)d(A, B) + (a+b)(a+c)d(B, C) + \\ &\alpha[(a+c) + (b+c)d(A, B)] + \alpha[(a+b) + (a+c)d(B, C)] \end{aligned} \right\} \geq \left\{ \begin{aligned} &\alpha^2 d(A, C) + \\ &(a+b)(b+c)d(A, C) + \\ &\alpha[(a+b) + (b+c)d(A, C)] \end{aligned} \right.$$

We now separately consider matched lines from either side of this expression. It turns out that the inequality holds for each such pairing, thus establishing the overall inequality. The inequality for the first line is just scaled ordinary Euclidian distance. Then, letting  $D$  denote the origin, the inequality holds for the second line due to corollary 3, and for the third <sup>6</sup> due to lemma 1.  $\square$

#### IV. ALTERNATIVE FORMS AND INTERPRETATIONS

##### A. Association, and more about Finite Sets

To illustrate the wide variety of metrics that can be constructed with our earlier results, we start by returning to metrics on  $\mathcal{F}$ .

We first observe that finite sets may be regarded as binary vectors in Euclidian space. With this observation theorem 2 gives us that the following is a metric:

$$d_{\Delta_2}(A, B) = \begin{cases} \frac{|A \Delta B|^{1/2}}{|A|^{1/2} + |B|^{1/2}} & \text{if } A \cup B \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

It is interesting to note that without the square root operations, this definition corresponds to normalization of  $d_{\Delta}$  by  $|A| + |B|$ , which was earlier shown not to be a metric.

Again thinking of sets as binary vectors or variables, metric  $d_{\Delta_n}$  is recognized as a well known measure of association referred to as the *Jaccard coefficient* [7], or *S-coefficient*:

$$\frac{b + c}{a + b + c}$$

Where  $a, b, c, d$  refer to the standard entries in a 2-by-2 contingency (or association) table for binary variables.

<sup>6</sup>Factor  $\alpha$  may be disregarded as it applies to all terms, and regrouping the paired sums, yields a common term  $(a + b + c)$  which may be similarly disregarded.

This very basic comparison function may be expressed in many forms. Expressing essentially the same thing as a similarity measure corresponds to the Tanimoto Coefficient, [8], operating on binary vectors:

$$\frac{A^t B}{A^t A + B^t B - A^t B}$$

In the language of computer programming this is just the count of 1's in the exclusive *or* of bit vectors  $A$  and  $B$ , divided by the count of 1's in their logical *or*.

In [9], the measure is referred to as *association*. While the authors do not state that these various forms fail to satisfy the triangle inequality, they describe them along with other non-metric measures.

Other sources such as [10] mention explicitly the set function  $d_{\Delta_n}$ , but again fail to note that it satisfies the triangle inequality.

### B. Other Statistical Measures

The *Bray-Curtis* measure[11], is written:

$$\left( \sum_j |x_{1j} - x_{2j}| \right) / \left( \sum_j x_{1j} + \sum_j x_{2j} \right)$$

and is recognized as our  $d_n$  but under the  $L_1$  norm. Now we have seen that this does not form a metric, but since  $d_n$  is a metric we may write:

$$\left( \sum_j |x_{1j} - x_{2j}|^2 \right)^{1/2} / \left( \left[ \sum_j x_{1j}^2 \right]^{1/2} + \left[ \sum_j x_{2j}^2 \right]^{1/2} \right)$$

with the knowledge that this does form a metric.

Also in [11], the reader will recognize the *Canberra Metric* as  $d_n$  in  $\mathbb{R}^1$ , making  $d_n$  a considerable generalization of this apparently well known metric.

### C. Positive $n$ -tuples

In familiar settings such as  $\mathbb{R}^1$  and  $\mathbb{R}^N$ , with measure  $\mu$  defined as length/ area/ volume, the metric  $d_\mu$  corresponds to the amount of non-common length/ area/ volume, normalized by the total length/ area/ volume.

Here are some examples of metrics derived from corollary 1, that are defined for  $n$ -tuples of non-negative values. For brevity's sake, we will not show the  $A = B$  case for which any metric must evaluate to zero.

1. Thinking of our  $n$ -tuple as a histogram in the plane, we may write:

$$1 - \frac{\sum_i \min(A_i, B_i)}{\sum_i \max(A_i, B_i)} = \frac{\sum_i |A_i - B_i|}{\sum_i \max(A_i, B_i)}$$

This may be viewed as a *repaired* Bray-Curtis measure in that  $\max(A_i, B_i) = |A_i| + |B_i| - \min(A_i, B_i)$ . I.e. the denominator has an extra term which suffices to form a metric.

2. We may take the direct product of metric spaces, and combine metrics by any linear combination with the result still a metric. Thus we have in particular that:

$$1 - \sum_i \frac{\min(A_i, B_i)}{\max(A_i, B_i)} = \sum_i \frac{|A_i - B_i|}{\max(A_i, B_i)}$$

which resembles somewhat our first example but behaves quite differently.

3. Imagining now that our vectors represent the dimensions of an  $N$ -dimensional solid object rooted at the origin, and defining measure as volume, we have:

$$1 - \frac{\prod_i \min(A_i, B_i)}{\prod_i \max(A_i, B_i)} = \frac{\prod_i |A_i - B_i|}{\prod_i \max(A_i, B_i)}$$

An example in which a single component greatly influences the resulting value.

### D. Function Spaces

Now corollary 1 may also be applied to function spaces. Let  $C[0, 1]$  denote the space of non-negative continuous functions on  $[0, 1]$ . Given  $f, g \in C[0, 1]$ , the metric defined by:

$$1 - \frac{\int_0^1 \min(f(x), g(x)) dx}{\int_0^1 \max(f(x), g(x)) dx} = \frac{\int_0^1 |f(x) - g(x)| dx}{\int_0^1 \max(f(x), g(x)) dx}$$

may be thought of as an extension of our first positive  $n$ -tuple metric above, as  $N \rightarrow \infty$ , or in a Lebesgue/Measure-Space context.

Extending our second  $n$ -tuple metric to Riemann integrable functions allows us to write:

$$1 - \int_0^1 \frac{\min(f(x), g(x))}{\max(f(x), g(x))} dx = \int_0^1 \frac{|f(x) - g(x)|}{\max(f(x), g(x))} dx$$

Note that in some cases, and with the proper assumptions, the continuity and range of integration assumptions made above may be relaxed.

Now for the sake of brevity, we have focused mainly on examples using  $d_{\Delta_n} / d_\mu$ . Similar constructions apply to  $d_n$  for both  $n$ -tuples and function spaces.

### E. The Complex Plane

It is worthwhile noting (however obvious) that  $d_n$  may also be viewed as an alternate metric for the complex plane with  $L_2$  norm corresponding to complex absolute value.

### F. Numerical Analysis

We close this section by commenting that  $d_n$  used as a measure of relative error in numerical analysis, may prove interesting. Given a sequence  $\bar{X}_1, \bar{X}_2, \dots$  of approximate solutions, we might define  $R(i, j) = d_n(\bar{X}_i, \bar{X}_j)$  and the triangle inequality would then give us that:  $R(1, n) \leq \sum_{i=1}^{n-1} R(i, i+1)$ .

## V. CONCLUDING REMARKS

We have developed normalized forms for two important metrics and shown each to be an endpoint of a continuum of metric functions ranging to the original unnormalized forms.

In a combinatorial setting,  $d_{\Delta_n}$  may be viewed as a prototype for constructing a normalized bounded metric for many applications – for every injection of a pattern space into  $\mathcal{F}$  induces a metric on the original pattern space.<sup>7</sup>

Furthermore, evaluating this metric is possible given only  $|A \cap B|$  and the individual orders  $|A|$ , and  $|B|$  since  $|A \cup B| = |A| + |B| - |A \cap B|$ .

In practice, these values may sometimes be directly computed without explicitly forming injective images in set space. In [12] a VLSI chip is described which computes these three orders for a particular mapping of finite symbol strings into  $\mathcal{F}$ , and combines them to define a measure of string similarity.

In more Euclidian settings, the family  $d_{n_{\alpha\beta}}$  may be used where normalized behavior is required and the metric triangle inequality is of value, e.g. in finding nearest neighbors, or when bounding sequential sums of distances.

Beyond the existence and basic properties established in this paper, one may examine a number of topics including the notion of colinearity in these normalized spaces, the nature of the geodesics corresponding to  $d_n$ , and questions of statistical behavior.

## VI. ACKNOWLEDGEMENTS

I thank David Robbins and Marshall Buck for their proof that replaces the tortuous argument I first discovered and gave in an earlier version of this paper. I also thank C.W. Gear, N. Littlestone, I. Rivin, and W.D. Smith, for helpful discussions and references, E. Baum, for corrections and general assistance, – and S.R. Buss for his interest in and help with my work over many years.

## REFERENCES

- [1] D. Robbins and M. Buck. Private Communication, May 1993.
- [2] J. L. Kelly, *General Topology*. New York: D. Van Nostrand, 1955.
- [3] H. L. Royden, *Real Analysis*. New York: Macmillan Publishing, second ed., 1968.
- [4] R. G. Bartle, *The Elements of Integration*. John Wiley & Sons, Inc., 1966.
- [5] D. Haussler, “Decision theoretic generalizations of the pac model for neural net and other learning applications,” Tech. Rep. UCSC-CRL-91-02, University of California, Santa Cruz, December 1990.
- [6] H. Coxeter and S. Greitzer, *Geometry Revisited*. Washington, D.C.: Mathematical Association of America, 1967.
- [7] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, Inc., 1990.
- [8] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., 1973.
- [9] A. Kandel, *Fuzzy techniques in pattern recognition*. John Wiley & Sons, Inc., 1982.
- [10] R. J. Schalkoff, *Pattern recognition: statistical, structural, and neural approaches*. John Wiley & Sons, Inc., 1992.
- [11] A. Ralston, *Statistical Methods for Digital Computers*. John Wiley & Sons, Inc., 1977.
- [12] P. N. Yianilos, “A dedicated comparator matches symbol strings fast and intelligently,” *Electronics Magazine*, December 1983.

---

<sup>7</sup>For measure spaces we must additionally require that if  $A \neq B$ , then their images differ by more than a set of measure zero.