Chapter 1

Metrics for more than two points at once

David H. Wolpert
NASA Ames Research Center
dhw@email.arc.nasa.gov

The conventional definition of a topological metric over a space specifies properties of any measure of "how separated" two points in that space are. Here it is shown how to extend that definition, and in particular the triangle inequality, to concern arbitrary numbers of points. Such a measure of how separated the points within a collection are can be bootstrapped, to measure "how separated" from each other are two (or more) collections. The measure presented here also allows fractional membership of an element in a collection, and therefore measures "how spread out" a probability distribution over a space is. When such a measure is bootstrapped to compare two collections, it measures how separated two probability distributions are, or more generally, how separated a distribution of distributions is.

1.1 Introduction

The conventional definition of a topological metric formalizes the concept of distance. It specifies properties required of any function that purports to measure "how separated" two elements of a space are. However often one wants to measure "how separated" the members of a collection of more than two elements is. The conventional, ad hoc way to do this is to combine the pair-wise metric values for all pairs of elements in the collection, into an aggregate measure.

As an alternative, here the formal definition of a topological metric is extended to collections of more than two elements. In particular, the triangle inequality is extended to such collections. The measure presented here applies to collections with duplicate elements. It also applies to collections with "frac-

tional" numbers of elements, i.e., to probability distributions. Furthermore, this measure can be bootstrapped to measure "how separated" from each other two collections are. In other words, given a measure ρ of how separated from each other the elements in an arbitrary collection ξ are, one can define a measure of how separated from each other two collections ξ_1 and ξ_2 are. More generally, one can measure how separated a collection of such collections is. Indeed, with fractional memberships such bootstrapping allows us to measure how separated a distribution of distributions is.

In the next section the definition of a multi-argument metric (**multimetric**, for short) is presented, together with extensive set of examples. For instance, it is shown that the standard deviation of a probability distribution across $\mathbb{R}^N$ is a multimetric, whereas the variance of that distribution is not. The following section shows how to bootstrap from a multimetric for elements within a collection to a multimetric over collections. A fuller version of this paper, discussing applications and vector-valued multimetrics, and including all proofs, can be found at http://....

1.2 Multimetrics

Collections of elements from a space X are represented as vectors of counts, i.e., functions from $x \in X \to \{0, 1, 2, \ldots\}$. So for example, if $X = \{A, B, C\}$, and we have the collection of three A's, no B's, and one C, we represent that as the vector $(3, 0, 1)$. It is natural to extend this to functions from $x \in X$ to $\mathbb{R}$. In particular, doing this will allow us to represent probability distributions (or density functions, depending on the cardinality of X) over X. Accordingly, our formalization of multimetrics will provide a measure for how how spread out a distribution over distributions is.[1] Given X, the associated space of all functions from X to $\mathbb{R}$ is written as $\mathbb{R}^X$. The subspace of functions that are nowhere-negative is written as $(\mathbb{R}^+)^X$.

As a notational comment, integrals are written with the measure implicitly set by the associated space. In particular, for a finite space, the point-mass measure is implied, and the integral symbol indicates a sum. In addition δ_x is used to indicate the appropriate type of delta function (Dirac, Kronecker, etc.) about x. Other shorthand is $\mathcal{R}^X \equiv (\mathbb{R}^+)^X - \{0\}$ and $||v||$ to mean $\int dx\, v(x)$.

In this representation of collections of elements from X, any conventional metric taking two arguments in X can be written as a function ρ over a subset of the vectors in $\mathcal{R}^X$. That subset consists of all vectors that either have two of their components equal to 1 and all others 0, or one component equal to 2 and all others 0. For example, for $X = \{A, B, C\}$, the metric distance between A and B is $\rho(1, 0, 1)$, and from A to itself is $\rho(2, 0, 0)$.

Generalizing this, a multimetric for $T(X) \subseteq \mathcal{R}^X$ is defined as a real-valued function ρ over $\mathcal{R}^X$ such that $\forall u, v, w \in \mathcal{R}^X$,

[1] See [4, 2, 3, 5, 1, 6] and references therein for work on how spread out a pair of distributions is.

1) $u, v, w \in T(X) \;\; \Rightarrow \;\; \rho(u+v) \;\leq\; \rho(u+w) + \rho(v+w)$.
2) $\rho(u) \geq 0, \;\; \rho(k\delta_x) = 0 \; \forall x, k > 0$.
3) $\rho(u) = 0 \;\; \Rightarrow \;\; u = k\delta_x$ for some k, x.

In this representation of collections, if only one $x \in X$ is in a collection (perhaps occurring more than once), then only one component of u is non-zero. Accordingly, conditions (2) and (3) are extensions of the usual condition defining a metric that it be non-negative and equal 0 iff its arguments are the same. Condition (1) is an extension of the triangle inequality, to both allow repeats of elements from X and/or more than two elements from X to be in the collection. Note though that condition (1) involves sums in its argument rather than (as in a conventional norm-based metric for a Euclidean space) differences. Intuitively, $T(X)$ is that subset of $\mathcal{R}^X$ over which the generalized version of the triangle inequality holds.

Condition (1) implies that multimetrics obey a second triangle inequality, just as conventional metrics do:

$$\rho(u+v) \geq |\rho(u+w) - \rho(v+w)|.$$

(This follows by rewriting condition (1) as $\rho(u+w) \geq \rho(u+v) - \rho(v+w)$, and then relabeling twice.)

Example 1: Set $X = \mathbb{R}^N$. Take $T(X)$ to be those elements of $\mathcal{R}^X$ whose norm equals 1, i.e., the probability density functions over $\mathbb{R}^N$. Then have $\rho(s)$ for any $s \in \mathcal{R}^X$ (whether in $T(X)$ or not) be the standard deviation of the distribution $\frac{s}{||s||}$, i.e., $\rho(s) = \sqrt{\frac{1}{2} \int dx dx' \; \frac{s(x)s(x')}{||s||^2} (x - x')^2}$.

Conditions (2) and (3) are immediate. To understand condition (1), first, as an example, say that all three of u, v and w are separate single delta functions over X. Then condition (1) reduces to the conventional triangle inequality over $\mathbb{R}^N$, here relating the points (in the supports of) u, v and w. This example also demonstrates that the variance (i.e., the square of our ρ) is not a multimetric.

For a vector s that involves multiple delta functions, $\rho(s)$ measures the square root of the sum of the squares of the Euclidean distances between the points (in the support of) s. In this sense it tells us how "spread out" those points are. Condition (1) even holds for vectors that are not sums of delta functions however.

Example 2: As a variant of Ex. 1, have X be the unit simplex in $\mathbb{R}^N$, and use the same ρ as in Ex. 1. In this case any element of X is a probability distribution over a variable with N possible values. So any element of $T(X)$ is a probability density function over such probability distributions. In particular, say s is a sum of some delta functions for such an X. Then $\rho(s)$ measures how spread out the probability distributions in (the support of) s are. If those probability distributions are themselves sums of delta functions, they just constitute subsets of our N values, and $\rho(s)$ measures how spread out from one another those subsets are.

Example 3: As another variant of Ex. 1, for any X, take $T(X) = \mathcal{R}^X$. Define the tensor contraction $\langle s \mid t \rangle \equiv \int dx dx' s(x)t(x)F(x,x')$ where F is symmetric and nowhere-negative, and where $F(x,x') = 0 \Leftrightarrow x = x'$. Then $\rho(s) \equiv \sqrt{\langle s \mid s \rangle}$ obeys conditions (2) and (3) by inspection. It also obeys condition (1).

Note that the $\langle ., . \rangle$ operator is not an inner product over $\mathbb{R}^X$, the extension of $T(X)$ to a full vector space. When components of s can be negative, $\langle s, s \rangle$ may be as well. Note also that there is a natural differential geometric interpretation of this ρ when X consists of N values. Say we have a curve on an N-dimensional manifold with metric tensor F at a particular point on the curve, and that at that point the tangent vector to the curve is s. Then $\rho(s)$ is the derivative of arc length along that curve, evaluated at that point.

This suggest an extension of this multimetric, in which rather than a tensor contraction between two vectors, we form the tensor contraction of n vectors: $\langle s^1, \ldots, s^n \rangle \equiv \int dx^1 \ldots dx^n s^1(x^1) \ldots s^n(x^n) F(x^1, \ldots, x^n)$, where F is invariant under permutation of its arguments, nowhere-negative, and equals 0 if and only if all its arguments have the same value. Any $\rho(s)$ that is a monotonically increasing function of $\langle s, s, \ldots, s \rangle^{1/n}$ automatically obeys conditions (2) and (3).

It is worth collecting a few elementary results concerning multimetrics:

Lemma 1:

1. Let $\{\rho_i\}$ be a set of functions that obey conditions (2) and (3), and $\{a_i\}$ a set of non-negative real numbers at least one of which is non-zero. Then $\sum_i a_i \rho_i$ also obeys conditions (2) and (3).
2. Let $\{\rho_i\}$ be a set of functions that obey condition (1), and $\{a_i\}$ a set of non-negative real numbers at least one of which is non-zero. Then $\sum_i a_i \rho_i$ also obeys conditions (1).
3. Let $f : \mathbb{R} \to \mathbb{R}^+$ be a monotonically increasing concave function that equals 0 when its argument does. Then if ρ is a multimetric for some $T(X)$, $f(\rho)$ is also a multimetric for $T(X)$.
4. Let $f : X \to Y$ be invertible, and let ρ_Y be a multimetric over Y. Define the operator $B_f : \mathcal{R}^X \to \mathcal{R}^Y$ by $[B_f(s)](y) \equiv s(f^{-1}(y))$ if $f^{-1}(y)$ exists, 0 otherwise. B_f is a linear operator. This means $\rho_X(s) \equiv \rho_Y(A_f(s))$ is a multimetric.

Example 4: Take $X = \mathbb{R}^N$ again, and let $T(X)$ be all of $\mathcal{R}^X$ with bounded support. Then by Lemma 1, the width along x_1 of (the support of) $s \in T(X)$ is a multimetric function of s.

This means that the average of the width in x_1 over all possible rotations of X is also a multimetric. Similarly, consider the smallest axis-parallel box enclosing the (support of the) Euclidean points in s. Then the sum of the lengths of the edges of that box is a multimetric function of s.

On the other hand, while the volume of that box obeys conditions (2) and (3), in general it can violate condition (1). Similarly, the volume of the convex hull of the (support of) the points in s obeys conditions (2) and (3) but can violate (1). (In general, multimetrics have the dimension of a length, so volumes have to be raised to the appropriate power to make them be multimetrics.)

It is worth comparing the sum-of-edge-lengths multimetric to the standard deviation multimetric of Ex. 1 for the case where all arguments s are finite sums of delta functions (i.e., "consist of a finite number of points"). For such an s we can write the sum-of-edge-lengths multimetric as a sum over all N dimensions i of $\max_j s_i^j - \min_j s_i^j$, where s^j is the j'th point in s. In contrast, the (square of the) standard deviation multimetric is also a sum over all i, but of the (square of the) standard deviation of the i'th components of the points in s. Another difference is that the standard deviation multimetric is a continuous function of its argument, unlike the sum-of-edge-lengths multimetric.

Example 5: Let X be countable and have $T(X) = \mathcal{R}^X$. Then $\rho(s) = \int dx \Theta(s(x)) - 1$ where Θ is the Heaviside function is a multimetric. This is the volume of the support of s, minus 1.

Example 6: Let X be countable and have $T(X) = \mathcal{R}^X$. Then $\rho(s) = ||s|| - \max_x s(x)$ obeys conditions (2) and (3), by inspection. Canceling terms, for this ρ condition (1) holds iff $\max_x(u(x) + v(x)) \geq \max_x(u(x) + w(x)) + \max_x(v(x) + w(x)) - 2||w||$. This is not true in general, for example when $||w|| = 0$ and the supports of u and v are disjoint. However if we take $T(X)$ to be the unit simplex in $\mathcal{R}^X$, then condition (1) is obeyed, and ρ is a multimetric.

Example 7: Let X have a finite number of elements and set $T(X) = \mathcal{R}^X$. Say that $\rho(s) = 0$ along all of the axes, and that everywhere else, $k \leq \rho(s) \leq 2k$ for some fixed $k > 0$. Then ρ is a multimetric.

1.3 Concavity gaps and dispersions

In Ex. 1, ρ can be used to tell us how spread out a distribution over $\mathbb{R}^N$ is. One would like to be able to use that ρ to construct a measure of how spread out a collection of multiple distributions over $\mathbb{R}^N$ is. Intuitively, we want a way to construct a metric for a space of sets (generalized to be able to work sets with duplicates, fractional memberships, etc.) from a metric for the members within a single set. This would allow us to directly incorporate the distance relation governing X into the distance relation for $\mathcal{R}^X$.

To do this, first let $\{Y, S(Y)\}$ be any pair of a subset of a vector space together with a subset of $\mathbb{R}^Y$ such that $\forall g \in S(Y), \frac{\int dy g(y) y}{||g||} \in Y$. (As an example, we could take Y to be any convex subspace of a vector space, with $S(Y)$ any subset of $\mathcal{R}^Y$.) Then the associated **concavity gap** operator $\mathcal{C} : S(Y) \to \mathbb{R}^{S(Y)}$ is

$$(\mathcal{C}\sigma)(g) = \sigma(\frac{\int dy\ g(y) y}{||g||}) - \frac{\int dy\ g(y)\sigma(y)}{||g||}$$

where $y \in Y$, and both σ and g are arbitrary elements in $S(Y)$. So the concavity

gap operator takes any single member of the space $S(Y)$ (namely σ) and uses it to generate a function (namely, $\mathcal{C}\sigma$) over all of $S(Y)$.[2]

In particular, say $Y = T(X)$ for some space X. Say we are given a multimetric σ measuring the (X-space) spread specified by any element of Y. Say we are also given a g which is a normalized distribution over Y. Then $\mathcal{C}\sigma(g)$ is a measure of how spread out the distribution g is. Note that in this example space $S(Y)$ is both the space of multimetrics over Y and the space of distributions for Y, exemplified by σ and g, respectively.

We can rewrite the definition of the concavity gap in several ways:

$$\begin{aligned} \mathcal{C}\sigma(g) &= \sigma(E_g(y)) - E_g(\sigma) \\ &= \sigma(\frac{\mathbf{y} \cdot g}{||g||}) - \frac{\sigma \cdot g}{||g||} \end{aligned}$$

where E_g means expected value evaluated according to the probability distribution $\frac{g}{||g||}$, and in the last expression $\mathbf{y}$ is the (infinite-dimensional) matrix whose y'th column is just the vector y, and the inner products are over the vector space $S(Y)$. Taken together, these equations say that the concavity gap of σ, applied to the distribution g, is given by evaluating σ at the center of mass of the distribution g, and then subtracting the inner product between σ and g.

Example 9: Let $Y = \mathbb{R}^N$, and choose $S(Y)$ to be the set of nowhere-negative functions of Y with non-zero magnitude. Choose $\sigma(y) = 1 - \sum_{i=1}^{N} y_i^2$. Then $\mathcal{C}\sigma(g) = Var(\frac{g}{||g||})$.

Example 10: Say X has N values, with $T(X) = \mathcal{R}^X$. Consider a $u \in T(X)$ whose components are all either 0 or some particular constant a such that $\int dx\ u(x) = 1$. So u is a point on the unit hypercube in $T(X)$, projected down to the unit simplex. Let $\mathcal{T}$ be the set of all such points u. In the usual way, the support of each element of $\mathcal{T}$ specifies a set of elements of X.

Let $Y = T(X)$, and have $S(Y) = \mathcal{R}^Y$. Have g be a uniform average of a countable set of delta functions, each of which is centered on a member of $\mathcal{T}$. So each of the delta functions making up g specifies a set of elements of X; g is a specification of a collection of such X-sets.

In this scenario $\sigma(E_g(y))$ is σ applied to the union (over X) of all the X-sets specified in g. In contrast, $E_g(\sigma)$ is the average value you get when you apply σ to one of the X-sets specified in g. $\mathcal{C}\sigma(g)$ is the difference between these two values. Intuitively, it reflects how much overlap there is among the X-sets specified in g.

Example 11: Say X has N values, with $T(X) = \mathcal{R}^X$. Have $Y = T(X)$, and $S(Y) = \mathcal{R}^Y$, i.e., the set of all nowhere-negative non-zero functions over those points in $\mathbb{R}^N$ with no negative components. Choose $\sigma(y) = H(y)\ \forall y \in Y$, where $H(.) = -\int dy\ y(x)\ln[y(x)]$, the Shannon entropy function extended to non-normalized y. This σ is a natural choice to measure how "spread out" any point in Y with magnitude 1 is.

[2]Equivalently, it can be viewed as a non-symmetric function from $S(Y) \times S(Y) \rightarrow \mathbb{R}$, although we will not exploit that perspective here.

Have g be a sum of a set of delta functions, about the distributions over $\mathbb{B}$, $\{v^1, v^2, \ldots\}$. Then $\mathcal{C}\sigma(g)$ is a measure of how "spread out" those distributions are. In the special case where $g = \delta_{v^1} + \delta_{v^2}$, $\mathcal{C}\sigma(g)$ is the Jensen-Shannon divergence between v^1 and v^2 [1, 4]. More generally, if g is a probability density function across the space of all distributions over $\mathbb{B}$, $\mathcal{C}\sigma(g)$ is a measure of how "spread out" that density function is.

Lemma 2:

1. $\mathcal{C}$ is linear.
2. $\mathcal{C}\sigma$ is linear $\Leftrightarrow$ it equals 0 everywhere $\Leftrightarrow$ σ is linear.
3. $\mathcal{C}\sigma$ is continuous $\Leftrightarrow$ σ is continuous.
4. $\mathcal{C}\sigma(g) = 0$ if $g \propto \delta_{y'}$ for some $y' \in Y$.
5. Giving $\mathcal{C}\sigma$ and the values of σ at $1+|Y|$ distinct points in Y fixes the value of σ across all Y. ($|Y|$ is the dimension of Y.)
6. The equivalence class of all σ' having a particular concavity gap $\mathcal{C}\sigma$ is the set of functions of $y \in Y$ having the form $\{\sigma(y) + b \cdot y + a \ : \ a \in \mathbb{R}, b \in Y, \sigma(y) + b \cdot y + a \in S(Y)\}$.

By (1.4), $\mathcal{C}\sigma$ necessarily obeys the second part of condition (2) if $S(Y) = \mathcal{R}^Y$.

Next define a (strict) **dispersion** over a space X as a (strictly) concave real-valued function over $\mathcal{R}^X$ that obeys conditions (2) and (3) of a multimetric $\forall u, v, w \in \mathcal{R}^X$.

Example 12: Take $X = \{1, 2\}$, with $T(X) = \mathcal{R}^X = \mathbb{R}^2 - \{0\}$. Define $\sigma(u \in \mathbb{R}^2)$ to equal 0 if $u_1 = 0$ or $u_2 = 0$, and equal $\ln(1+u_1) + \ln(1+u_2)$ otherwise. Then σ is a (not everywhere continuous) strict dispersion.
Example 13: The $X, \mathcal{R}^X$, and ρ of Ex. 3 form a strict dispersion.
Example 14: The $X, \mathcal{R}^X$, and σ of Ex. 5 form a dispersion.
Example 15: The $X, \mathcal{R}^X$, and σ of Ex. 11 form a strict dispersion.

Lemma 3: Let $T(X) = \mathcal{R}^X$.

1. σ is a dispersion over $T(X) \Rightarrow$ σ is nowhere-decreasing over $T(X)$.
2. σ is a dispersion over $T(X)$ and $\sigma(s)$ is independent of $||s|| \ \forall \ s \neq 0 \in T(X)$ $\Rightarrow$ σ is constant over the interior of $T(X)$.
3. σ is (strictly) concave over $V(X) \Leftrightarrow \mathcal{C}\sigma$ obeys condition (2) in full (and condition (3)) over $T(X)$.
4. Say that σ is continuous over $T(X)$. Then $\mathcal{C}\sigma$ is separately (strictly) concave over each simplex in $T(X) \Leftrightarrow$ σ is (strictly) concave over $T(X)$.

Let $f : \mathbb{R} \to \mathbb{R}$ be monotonically increasing and strictly concave. Then by Lemma 3.3, if σ is strictly concave, $f(\mathcal{C}\sigma)$ obeys conditions (2) and (3). For example, this is the case for $\sqrt{\mathcal{C}\sigma}$. In other words, so long as σ is a strict dispersion, $\sqrt{\mathcal{C}\sigma}$ obeys those conditions. On the other hand, Lemma 3.2 means that any nontrivial σ that normalizes its argument (so that it is a probability distribution) and then evaluates a function of that normalized argument cannot be a dispersion. So for example, if a concavity gap is a dispersion, it must be constant. Fortunately it is not the case that if $f(\mathcal{C}\sigma)$ is a multimetric it must be constant. In particular, often for a strictly concave σ, $\sqrt{\mathcal{C}\sigma}$ for space $\{Y, S(Y)\}$ is a multimetric for an appropriate $T(Y) \subseteq S(Y)$.

Example 16: Choose $\{\sigma, Y, S(Y)\}$ as in Ex. 11, and take $T(Y)$ to be all elements of $S(Y)$ which are sums of two delta functions. This σ is strictly concave, so we know conditions (2) and (3) are obeyed by $\sqrt{\mathcal{C}\sigma}$. Furthermore, for this choice of $T(Y)$, obeying condition (1) reduces to obeying the conventional triangle inequality of two-argument metrics, and it is known that the square root of the Jensen Shannon divergence obeys that inequality [4, 5, 1]. Therefore all three conditions are met.

Example 17: Choose $\{\sigma, Y, S(Y)\}$ as in Ex. 9. This σ is strictly concave, and therefore $\sqrt{\mathcal{C}\sigma}$ automatically obeys conditions (2) and (3). Now take $T(Y) = S(Y)$. Write $\mathcal{C}\sigma(g)$ as $\langle g, g\rangle$ for the tensor contraction of Ex. 3, where $F(y, y') = \frac{(y-y')\cdot(y-y')}{2}$. So by that example, we know that $\sqrt{\mathcal{C}\sigma}$ is a multimetric.

Acknowledgements: I would like to thank Bill Macready and Creon Levit.

Bibliography

[1] Fuglede, Bent, and Flemming Topsoe, "Jensen-shannon divergence and hilbert space embedding", Submitted to ISIT2004 (2004).

[2] Hall, M., "Universal geometric approach to uncertainty, entropy, and information", *Physical Review A* **59** (1999), 2602.

[3] Judge, George, "Semiparametric moment based estimation for binary response models", RMBE-DE1-4-2-04.doc (2004).

[4] Lin, Jianhua, "Divergence measures based on the shannon entropy", *IEEE Trans. Info. Theory* **37**, 1 (1991), 145–151.

[5] Osterreicher, Ferdinand, and Igor Vajda, "A new class of metric divergiences on probability spaces and its applicability in statistics", *Ann. Inst. Statist. Math.* **55**, 3 (2003), 639–653.

[6] Wolpert, David H., and William Macready, "Self-dissimilarity as a high dimensional complexity measure", *Proceedings of the International Conference on Complex Systems, 2004*, (2004), in press.