We define four metrics between probability measures on a space equipped with a hierarchical partition tree, and prove their equivalence. Similar metrics have previously been defined in more restrictive settings; in particular, the well-known Earth Mover's Distance is widely used in machine learning. We adapt the definitions of these metrics to a much broader class of geometries, and use machinery from abstract harmonic analysis to prove they are all equivalent. We validate the theoretical results with numerical experiments.

Earth Mover's Distance and Equivalent Metrics for Spaces with Hierarchical Partition trees

R.R. Coifman[†] and W.E. Leeb[†]

Technical Report YALEU/DCS/TR-1482
July 22, 2013

[†] Dept. of Mathematics, Yale University, New Haven CT 06511

Approved for public release: distribution is unlimited.
Keywords: *earth mover's distance, partition trees, wavelets, haar basis, martingale*

1 Introduction

A basic problem in data analysis is to measure the similarity between two functions on some space of points. For example, two documents can be described by the relative frequencies of a collection of keywords; in this case, a document is a probability distribution over the space of words, and can be compared using one of the many of similarity measures in [1] defined for probability distributions. When the space on which the functions are defined has an underlying geometry, it is prudent to exploit this in the definition of the distance between the functions. In the document example, if we can measure the similarity in meaning of two words, two documents should be close if the words they contain are similar in meaning, even if the words themselves are distinct.

This basic premise informs the definition of the Earth Mover's Distance (EMD), a distance between probability measures widely used in machine learning. There are various precise ways of formulating EMD; see, for example [2], where it is defined as a distance between signatures. We will define it formally in Section 3 as a distance between two probability measures; intuitively, it measures the minimal cost of transforming one probability measure into another by moving mass, where the cost of moving a piece of mass between two locations is specified a priori by the geometry on the underlying measure space, say by a metric. If one probability measure is a small distortion of the other, the EMD between them will be small. EMD is therefore insensitive to perturbations, which is often a desirable property of a metric.

Another property we might desire for a metric between functions is that it give different weight to activity at different scales, for whatever notion of 'scale' one might have for a given problem. In signal and image processing, for example, we might be primarily interested in comparing the low frequencies of two signals, while filtering out high frequency behavior which is often noise. On the other hand, we might not want to completely discard the higher frequencies, if, say, we are comparing two textured images, where the high frequency variations are what characterize the texture. More generally, we seek distances that give higher weight to large-scale differences between functions, while still retaining some sensitivity to their small-scale variations.

We introduce three metrics that formalize this intuition, where in our setting the notion of 'scale' is captured by the size of folders in a hierarchical partition tree on the data set Ω. Two of the metrics we define, labeled D_2 and D_4 and defined in Sections 4 and 6, respectively, measure the changes between scales, where differences in the behavior across larger scales contribute more heavily to the distance. The metric D_3 defined in Section 5 on the other hand, simply measures the activity at each scale, again giving weight to the larger scales.

We remark that, although we formulate our definitions as distances between probability distributions p_1 and p_2, all of the metrics depend only the difference $p_1 - p_2$; furthermore, the proofs of equivalence between the metrics depend only the fact that $p_1 - p_2$ has mean zero. Consequently, the four distances we define are really norms on the space of mean zero L^1 functions (or mean zero finite measures). However, because the EMD 'norm' has such a natural interpretation when applied to the difference of two probability measures, we will stick with the language of probability and measure the distance between two probability distributions p_1 and p_2.

2 Definitions, Notation, and Basic Results

2.1 Definitions and Notation for Hierarchical Partition Trees

Let Ω be a set with a hierarchical partition tree $\mathcal{T}$, as in [3, 4]. We impose the critical assumption, found in that paper, that there are constants B_U, B_L such that

$$0 < B_L < \frac{|child|}{|parent|} < B_U < 1. \tag{1}$$

where *child* is a subfolder of *parent*.

We assume we are given a measure on Ω, and that $|\Omega| = 1$. Let $\{\psi\}$ denote a Haar-like basis on the tree. These functions, together with the constant function, are an orthonormal basis for $L^2(\Omega)$. For any two points $x, y \in \Omega$ we define the tree metric

$$d(x, y) = \inf\{|X| : X \in \mathcal{T}, x, y \in X\}.$$

Given a function f on Ω, define the Hölder seminorm by

$$C_H(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)^\alpha}.$$

In all that follows, we fix a parameter α, $0 < \alpha \leq 1$. Whenever we say that a function is 'Hölder' we mean with exponent α.

For a function f on Ω and a subset $I \subset \Omega$, let

$$m(f, I) = \frac{1}{|I|} \int_I f(x) dx$$

denote the average value of f on I. Let $\mathcal{P}_l$ denote the set of folders at level l. For each $l \geq 0$, define the expectation operator $\mathbb{E}_l$ by

$$\mathbb{E}_l f(x) = \sum_{I \in \mathcal{P}_l} m(f, I) \chi_I(x) \tag{2}$$

that is, $\mathbb{E}_l f$ is constant on each of the folders in partition l, with value there equal to the average value of f over that folder.

We also define the operators Δ_l by

$$\Delta_l f(x) = \mathbb{E}_{l+1} f(x) - \mathbb{E}_l f(x) \tag{3}$$

for $l \geq 0$.

2.2 General Facts about Haar Functions

The results of this section are essentially found in the paper *Wavelets on Trees, Graphs, and High Dimensional Data*; however, we derive tighter estimates here.

Theorem 1. *Suppose $f : \Omega \to \mathbb{R}$ is a function with Hölder constant $C_H(f)$. Then $|\langle f, \psi\rangle| \leq C_H(f)|I(\psi)|^{\alpha+\frac{1}{2}}$ for all ψ.*

Proof. If f has Hölder constant C, then for all Haar functions ψ, if $I = I(\psi)$

$$\begin{aligned}
|\langle f, \psi\rangle| &= \left| \int_I f(x)\psi(x)dx \right| = \left| \int_I \left\{ f(x) - \frac{1}{|I|}\int_I f(y)dy \right\} \psi(x)dx \right| \\
&= \left| \int_I \left(\frac{1}{|I|}\int_I (f(x) - f(y))dy \right) \psi(x)dx \right| \\
&\leq \left\{ \int_I \left(\frac{1}{|I|}\int_I (f(x) - f(y))dy \right)^2 dx \right\}^{1/2} ||\psi||_2
\end{aligned}$$

(by Cauchy-Schwarz)

$$\leq \left\{ \int_I \left(\frac{1}{|I|}\int_I |f(x) - f(y)|dy \right)^2 dx \right\}^{1/2}$$

(since ψ has L^2 norm 1)

$$\leq \left\{ \int_I \left(\frac{1}{|I|}\int_I C|I|^\alpha dy \right)^2 dx \right\}^{1/2}$$

(since f has Hölder constant C and $d(x, y) \leq |I|$)

$$= C|I|^{\alpha+1/2}.$$

□

Theorem 2. *If f is a function on Ω with $|\langle f, \psi\rangle| \leq C|I(\psi)|^{\alpha+\frac{1}{2}}$ for some $C > 0$ and for all ψ, then f has Hölder constant $C_H(f) \leq C\frac{2(1-B_L)}{B_L(1-B_U^\alpha)}$.*

The proof of this theorem requires some preliminary lemmas. For a folder I at level l, let $sub(I)$ denote the set of its subfolders at level $l + 1$.

Lemma 1. *Assuming the balance condition (1), $|sub(I)| \leq \frac{1}{B_L}$.*

Proof. If $sub(I) = \{I_1, \ldots, I_n\}$, then by the balance condition (1), $|I_i| \geq B_L|I|$. Summing over all i gives $|I| = \sum_{i=1}^n |I_i| \geq B_L n|I|$, from which it follows $|sub(I)| = n \leq 1/B_L$, as claimed. □

Lemma 2. *Let I be any folder of the tree. For a Haar function ψ, denote by $I(\psi)$ the folder supporting ψ. Then for all $x \in I$*

$$\sum_{\psi: I(\psi)=I} |\psi(x)| \leq \left(\frac{1}{B_L} - 1 \right) \frac{1}{|I|^{1/2}}.$$

Proof. Suppose I has n subfolders, $I_1, \ldots, I_n$. Let $\psi_1, \ldots, \psi_{n-1}$ denote the $n - 1$ Haar functions supported on I. Each ψ_j is constant on the folders I_i; let $\psi_j(I_i)$ denote its values. The orthogonality gives us:

$$\sum_{i=1}^n \psi_j(I_i)\psi_k(I_i)|I_i| = \delta_{jk}.$$

Furthermore, the functions ψ_j are orthogonal to the constant function on I. It follows that if we define the matrix W by

$$W = \begin{pmatrix} |I|^{-1/2} & |I|^{-1/2} & \cdots & |I|^{-1/2} \\ \psi_1(I_1) & \psi_1(I_2) & \cdots & \psi_1(I_n) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n-1}(I_1) & \psi_{n-1}(I_2) & \cdots & \psi_{n-1}(I_n) \end{pmatrix}$$

and the diagonal matrix D by

$$D = \begin{pmatrix} |I_1| & & & \\ & |I_2| & & \\ & & \ddots & \\ & & & |I_n| \end{pmatrix}$$

then $WDW^T = id$. This implies $DW^TW = id$, and so $W^TW = D^{-1}$. This last equality implies that for each folder I_i,

$$\frac{1}{|I|} + \sum_{j=1}^{n-1} \psi_j(I_i)^2 = \frac{1}{|I_i|}$$

from which it follows that

$$\left(\sum_{j=1}^{n-1} |\psi_j(I_i)| \right)^2 \leq (n-1) \sum_{j=1}^{n-1} \psi_j(I_i)^2 \leq (n-1) \left(\frac{1}{|I_i|} - \frac{1}{|I|} \right).$$

By Lemma 1, $n \leq 1/|B_L|$; and the balance condition (1) implies $1/|I_i| \leq 1/(B_L|I|)$. Consequently,

$$\sum_{j=1}^{n-1} |\psi_j(I_i)| \leq \sqrt{n-1} \sqrt{\frac{1}{B_L} \frac{1}{|I|} - \frac{1}{|I|}} \leq \left(\frac{1}{B_L} - 1 \right) \frac{1}{|I|^{1/2}},$$

as claimed. □

Proof of Theorem 2. For each $l \geq 0$ and each $x \in \Omega$, let $J_{l,x}$ denote the unique folder at level l containing x. Suppose without loss of generality that $C = 1$, i.e. $|\langle f, \psi \rangle| \leq |I(\psi)|^{\alpha+1/2}$ for all ψ. Take any $x, y \in \Omega$, and suppose I is the smallest folder containing both points, and I is at level L. So $d(x, y) = |I|$, and all Haar functions supported on

folders at levels $l < L$ have the same value on x and y. Consequently,

$$\begin{aligned}
|f(x) - f(y)| &= \left| \sum_{l \geq 0} \sum_{\psi : I(\psi) = J_{l,x}} \langle f, \psi \rangle \psi(x) - \sum_{l \geq 0} \sum_{\psi : I(\psi) = J_{l,y}} \langle f, \psi \rangle \psi(y) \right| \\
&= \left| \sum_{l \geq L} \sum_{\psi : I(\psi) = J_{l,x}} \langle f, \psi \rangle \psi(x) - \sum_{l \geq L} \sum_{\psi : I(\psi) = J_{l,x}} \langle f, \psi \rangle \psi(y) \right| \\
&\leq \sum_{l \geq L} |J_{l,x}|^{\alpha + 1/2} \sum_{\psi : I(\psi) = J} |\psi(x)| + \sum_{l \geq L} |J_{l,Y}|^{\alpha + 1/2} \sum_{\psi : I(\psi) = J} |\psi(x)| \\
&\leq \left(\frac{1}{B_L} - 1 \right) \left(\sum_{l \geq L} |J_{l,x}|^{\alpha + 1/2} \frac{1}{|J_{l,x}|^{1/2}} + \sum_{l \geq L} |J_{l,Y}|^{\alpha + 1/2} \frac{1}{|J_{l,y}|^{1/2}} \right) \\
&\leq \left(\frac{1}{B_L} - 1 \right) 2 \sum_{l \geq L} B_U^{(l-L)\alpha} |I|^{\alpha} \\
&= \frac{2(1 - B_L)}{B_L (1 - B_U^{\alpha})} d(x, y)^{\alpha}
\end{aligned}$$

as desired. □

We derive an upper bound on the L^{∞} norm of the Haar functions that will be useful later.

Lemma 3. *If ψ is a Haar function supported on a folder F of the tree, then*

$$||\psi||_{\infty} \leq \frac{\sqrt{1 - B_L}}{\sqrt{B_L |F|}}.$$

Proof. Suppose $F = I \sqcup J$, where I is the subfolder of F on which ψ attains its maximum value; call this maximum value c, so that $|c| = ||\psi||_{\infty}$ (note that J is not necessarily a subfolder of F; it is the union of all the subfolders of F, excluding I). Because ψ has L^2 norm 1

$$1 = c^2 |I| + \int_J \psi(x)^2 dx.$$

Since $\int_F \psi(x) dx = 0$, we get

$$\int_J \psi(x) dx = -c|I|.$$

From this follows easily the algebraic identity

$$\int_J \left(\psi(x) + c \frac{|I|}{|J|} \right)^2 dx = \int_J \psi(x)^2 dx - c^2 \frac{|I|^2}{|J|}$$

and combining this with the above gives:

$$\begin{aligned}
1 &= c^2|I| + \int_J \left(\psi(x) + c\frac{|I|}{|J|}\right)^2 dx + c^2\frac{|I|^2}{|J|} \\
&= c^2|I|\left(1 + \frac{|I|}{|J|}\right) + \int_J \left(\psi(x) + c\frac{|I|}{|J|}\right)^2 dx \\
&\geq c^2|I|\left(1 + \frac{|I|}{|J|}\right)
\end{aligned}$$

from which it follows

$$|c| \leq \sqrt{\frac{1-A}{A}}\frac{1}{\sqrt{|F|}}$$

where $A = |I|/|F|$. Since $\sqrt{\frac{1-A}{A}}$ is decreasing as a function of $A \in (0,1)$, and the smallest value A can assume is B_L, and $|c| = ||\psi||_\infty$, we get

$$||\psi||_\infty \leq \sqrt{\frac{1-B_L}{B_L}}\frac{1}{\sqrt{|F|}}$$

which is the desired result. □

Note that the result is tight, since we could take $\psi(x) = \sqrt{\frac{1-B_L}{B_L}}\frac{1}{\sqrt{|F|}} \equiv c$ on a subfolder I of size $B_L|F|$, and equal to $-c\frac{|I|}{|F\setminus I|}$ on $F \setminus I$; this will have L^2 norm 1 and be mean zero.

2.3 General Facts about $\mathbb{E}_l$ and Δ_l

Lemma 4. *If f has Hölder constant 1 (with exponent α), then for each folder $J \in \mathcal{P}_l$,*

$$\sup_{x\in J} |\Delta_l f(x)| \leq |J|^\alpha.$$

Proof. Let $J \in \mathcal{P}_l$, and let $I \in sub(J)$ be any subfolder of J. On I, $\mathbb{E}_l f \equiv \frac{1}{|J|}\int_J f(x)dx$ and $\mathbb{E}_{l+1} f \equiv \frac{1}{|I|}\int_I f(x)dx$. Therefore

$$\begin{aligned}
|\Delta_l f(I)| &= \left|\frac{1}{|I|}\int_I f(x)dx - \frac{1}{|J|}\int_J f(y)dy\right| \\
&= \left|\frac{1}{|J|}\int_J \frac{1}{|I|}\int_I f(x)dxdy - \frac{1}{|I|}\int_I \frac{1}{|J|}\int_J f(y)dydx\right| \\
&= \left|\frac{1}{|I||J|}\int_I\int_J (f(x) - f(y))dydx\right| \\
&\leq \frac{1}{|I||J|}\int_I\int_J |f(x) - f(y)|dydx \\
&\leq \frac{1}{|I||J|}\int_I\int_J |J|^\alpha dydx
\end{aligned}$$

(since f is Hölder)

$$= |J|^\alpha$$

as claimed. □

In fact the converse to this result is also true, that is, we can fully characterize Hölder functions by the size of $\Delta_l f$ on level l folders. The proof is nearly identical to the characterization of Hölder functions by the decay of their wavelet coefficients given earlier, and as we do not need it, we will not present it here.

It is easy to check the identity $\mathbb{E}_l \mathbb{E}_k = \mathbb{E}_{\min(k,l)}$, from which it follows

$$\Delta_l^2 = \Delta_l. \tag{4}$$

Furthermore, $\mathbb{E}_l$, and hence Δ_l, is self-adjoint. In fact, the operator $\mathbb{E}_l$ is given by the symmetric kernel

$$a_l(x,y) = \begin{cases} |I|^{-1}, \text{ if } x,y \in I, I \in \mathcal{P}_l \\ 0, \text{ otherwise.} \end{cases}$$

Lemma 5. *For any $J \in \mathcal{P}_l$,*

$$\int_J |\Delta_l f(x)| dx = \sum_{I \in sub(J)} |I||m(f,I) - m(f,J)| \tag{5}$$

Proof. $\mathbb{E}_{l+1} f$ and $\mathbb{E}_l f$ are constant on each $I \in sub(J)$, with values $m(f,I)$, $m(f,J)$ there, respectively; so

$$\begin{aligned} \int_J |\Delta_l f(x)| dx &= \sum_{I \in sub(J)} \int_I |\Delta_l f(x)| dx = \sum_{I \in sub(J)} \int_I |\mathbb{E}_{l+1} f(x) - \mathbb{E}_l f(x)| dx \\ &= \sum_{I \in sub(J)} \int_I |m(f,I) - m(f,J)| dx \\ &= \sum_{I \in sub(J)} |I||m(f,I) - m(f,J)| dx \end{aligned}$$

as desired. □

3 First Metric: Earth Mover's Distance

Given two probability densities p_1 and p_2 on Ω, we define the Earth Mover's Distance

$$D_1(p_1,p_2) = \inf_\pi \int_\Omega \int_\Omega d(x,y)^\alpha d\pi(x,y)$$

where the infimum is over all non-negative measures π on $\Omega \times \Omega$ satisfying

$$\pi(E \times \Omega) - \pi(\Omega \times E) = \int_E p_1(x) dx - \int_E p_2(x) dx \text{ for measurable subsets } E \subseteq \Omega. \tag{6}$$

The interpretation of this metric as as follows. Suppose p_1 and p_2 describe two mounds of equal mass sitting on Ω, and our goal is to start with the mound given by p_1 and move mass around to reshape it into p_2. We are allowed to dig holes (i.e. take more mass out of a location than is there to begin with), just so long as whatever deficit we create sending mass out is made up by the mass coming in. Then the measure π describes the mass transfer, that is, if $A, B \subset \Omega$, then $\pi(A \times B)$ is the amount of mass sent out from A and arriving in B. The difference of marginals constraint (6) expresses the fact that for any set $E \subset \Omega$, the net change in mass must be the amount of mass we start with, namely $\int_E p_1(x)dx$, minus the amount of mass we end up with, $\int_E p_2(x)dx$.

If $d(x,y)^\alpha$ is the cost of moving a unit of mass from x to y, then $\int\int d(x,y)^\alpha d\pi(x,y)$ is the total cost of moving all the mass using the transport described by the measure π. So the metric $D_1(p_1, p_2)$ is the *minimal* cost over all rearrangements of p_1 to p_2.

There is another expression for the earth mover's distance which we will find convenient. We have the following well-known theorem:

Theorem 3 (Kantorovich-Rubinstein)**.** *The earth mover's distance is equal to the following:*

$$D_1(p_1, p_2) = \sup\Big\{ \int_\Omega f(x)(p_1(x) - p_2(x))dx : f \text{ s.t. } C_H(f) < 1 \Big\}$$

The proof of a very general version of this result can be found in, among other places, [8]. In the case where Ω is finite, the proof follows easily from the duality theorem of linear programming.

4 Second Metric: Weighted L^1 Norm of Haar Coefficients

We define another metric D_2 between p_1 and p_2, which we will prove is equivalent to the earth mover's distance D_1. The definition and the proof of equivalence are inspired by the paper *Approximate earth mover's distance in linear time* by Sameer Shirdhonkar and David W. Jacobs. Consider the expansion of $p_1 - p_2$ in the Haar-like basis:

$$p(x) = \sum_\psi \langle p_1 - p_2, \psi\rangle \psi(x).$$

(Note that we don't have to include the constant function in this expansion, since $p_1 - p_2$ has mean zero.) Then we define

$$D_2(p_1, p_2) = \sum_\psi |I(\psi)|^{\alpha + \frac{1}{2}} |\langle p_1 - p_2, \psi\rangle|$$

where $I(\psi)$ is the folder supporting ψ.

We state some trivial facts, without proof:

Lemma 6. *For $c > 0$,*

$$cD_1(p_1, p_2) = \sup\Big\{ \int_\Omega f(x)(p_1(x) - p_2(x))dx : C_H(f) < c \Big\}.$$

Lemma 7. *The metric D_2 can be expressed as*

$$D_2(p_1,p_2) = \sup\Big\{ \sum_{\psi} \langle p_1 - p_2, \psi\rangle\langle f,\psi\rangle : f \text{ s.t. } |\langle f,\psi\rangle| \le |I(\psi)|^{\alpha+\frac{1}{2}} \Big\}$$
$$= \sup\Big\{ \int_\Omega (p_1(x)-p_2(x))f(x)dx : f \text{ s.t. } |\langle f,\psi\rangle| \le |I(\psi)|^{\alpha+\frac{1}{2}} \Big\}$$

where $f(x) = \sum_\psi \langle f,\psi\rangle\psi(x)$ is the expansion of f in the basis $\{\psi\}$.

(The second equality above holds because the Haar-like basis is orthonormal, hence preserves inner products).

Lemma 8. *For $c > 0$,*

$$cD_2(p_1,p_2) = \sup\Big\{ \int_\Omega (p_1(x)-p_2(x))f(x)dx : f \text{ s.t. } |\langle f,\psi\rangle| \le c|I(\psi)|^{\alpha+\frac{1}{2}} \Big\}$$

By Lemma 7 and Theorem 3, both $D_1(p_1,p_2)$ and $D_2(p_1,p_2)$ are obtained by maximizing the inner product $\langle p_1 - p_2, f\rangle$ over some collection of f: to get $D_1(p_1,p_2)$ we restrict f to have Hölder norm not exceeding 1; and to get $D_2(p_1,p_2)$ we restrict f to have wavelet coefficients with a certain decay rate, namely $|\langle f,\psi\rangle| \le |I(\psi)|^{\alpha+\frac{1}{2}}$. However, Theorems 1 and 2 tell us that these constraints are nearly the same. Using the equivalence given by these theorems between the regularity of a function and the decay of its wavelet coefficients, we show the equivalence of the metrics D_1 and D_2.

We have:

$$\begin{aligned} D_1(p_1,p_2) &= \sup\Big\{ \langle p_1 - p_2, f\rangle : C_H(f) < 1 \Big\} \\ &\le \sup\Big\{ \langle p_1 - p_2, f\rangle : |\langle f,\psi\rangle| \le |I(\psi)|^{\alpha+\frac{1}{2}} \ \forall\psi \Big\} \\ &= D_2(p_1,p_2). \end{aligned}$$

The first line follows from Theorem 3, the second from Theorem 1, and the third from Lemma 7.

For the reverse we have, using Theorem 2:

$$\begin{aligned} D_2(p_1,p_2) &= \sup\Big\{ \int_\Omega (p_1(x)-p_2(x))f(x)dx : f \text{ s.t. } |\langle f,\psi\rangle| \le |I(\psi)|^{\alpha+\frac{1}{2}} \Big\} \\ &\le \sup\Big\{ \int_\Omega (p_1(x)-p_2(x))f(x)dx : f \text{ s.t. } C_H(f) < \frac{2(1-B_L)}{B_L(1-B_U^\alpha)} \Big\} \\ &= \frac{2(1-B_L)}{B_L(1-B_U^\alpha)} D_1(p_1,p_2). \end{aligned}$$

We have shown:

Theorem 4. *The metrics D_1 and D_2 are equivalent; more specifically,*

$$\frac{B_L(1-B_U^\alpha)}{2(1-B_L)} D_2(p_1,p_2) \le D_1(p_1,p_2) \le D_2(p_1,p_2).$$

Note that if Ω has N points, there is an $\mathcal{O}(N)$ algorithm analogous to the classical fast Haar transform for computing all the wavelet coefficients of a function f. Consequently, the distance $D(p_1, p_2)$ can be computed in linear time.

5 Third Metric: Averages on Each Folder

We now define the third metric D_3 by

$$D_3(p_1, p_2) = \sum_{I \in \mathcal{T}} |I|^{\alpha+1} |m(p_1 - p_2, I)| = \sum_{l \geq 0} \sum_{I \in \mathcal{P}_l} |I|^{\alpha} \int_I |\mathbb{E}_l(p_1 - p_2)|. \tag{7}$$

In the case of a perfectly balanced $M-$ary tree, where each folder has M subfolders of the same size (so in particular, the size of each folder at level l is M^{-l}), we can write D_3 as a weighted sum L_1 norms of the expectations of $p_1 - p_2$ at each level:

$$\begin{aligned} D_3(p_1, p_2) &= \sum_{l \geq 0} \sum_{I \in \mathcal{P}_l} |I|^{\alpha} \int_I |\mathbb{E}_l(p_1 - p_2)| = \sum_{l \geq 0} M^{-l\alpha} \sum_{I \in \mathcal{P}_l} \int_I |\mathbb{E}_l(p_1 - p_2)| \\ &= \sum_{l \geq 0} M^{-l\alpha} ||E_l(p_1 - p_2)||_1. \end{aligned} \tag{8}$$

In this special case, D_3 is similar to the approximate EMD introduced in [9]; however, D_3 is defined in the more abstract setting of hierarchical partition trees, and furthermore will be shown to be equivalent to EMD with constants of equivalence not dependent on the number of points in Ω. In particular, in our case the ratio of the distances between any two pairs of points can be arbitrarily large, whereas the constants in [9] depend logarithmically on the ratio of the diameter of the space to the minimum distance.

For the metric D_3 given by (7) on any tree, we will prove:

Theorem 5. *The metric D_3 is equivalent to the metrics D_1 and D_2.*

We introduce some constructions that will be used in the proof of Theorem 5. For each level l of the tree, define the pseudo-metric

$$d_l(x, y) = \begin{cases} |I|^{\alpha} + |J|^{\alpha} \text{ if } x \in I,\ y \in J, \text{ and } I \neq J \text{ are in level } l \text{ partition} \\ 0 \text{ if } x, y \text{ are in same level } l \text{ folder} \end{cases}$$

Though we will not use the fact in the proof of Theorem 5, we show that each d_l is a pseudo-metric.

Lemma 9. *$d_l(x, y)$ is a pseudo-metric (that is, it satisfies all axioms of a metric except $d_l(x, y) = 0 \Rightarrow x = y$)*

Proof. Non-negativity and symmetry are obvious, as is $x = y \Rightarrow d_l(x, y) = 0$. As for the triangle inequality, take three points x, y, z. We want to show $d_l(x, z) \leq d_l(x, y) + d_l(y, z)$. This is trivial if x, z are in the same level l folder (since the left side is zero), so suppose

otherwise. If y is in the same folder as x, then $d_l(x,y) = 0$ and $d_l(y,z) = d_l(x,z)$, so there is equality. Similarly if y is in the same folder as z. If all three points are in separate folders I_1, I_2, I_3, then the left side is $|I_1|^\alpha + |I_3|^\alpha$, while the right side is $|I_1|^\alpha + 2|I_2|^\alpha + |I_3|^\alpha$, which is bigger. □

We then define the metric

$$\rho_\alpha(x,y) = \sum_{l=1}^{\infty} d_l(x,y).$$

Lemma 10. *Let $d(x,y)$ denote the usual tree metric. Then the metrics $\rho_\alpha(x,y)$ and $d(x,y)^\alpha$ are equivalent.*

Proof. For each x, y let l be the first level at which x and y are in different folders, and let $I \in \mathcal{P}_{l-1}$ be the folder in the previous partition containing both x and y. Then by definition $d(x,y)^\alpha = |I|^\alpha$; furthermore, for every $l' \geq l$, if I_x and I_y are the level l' folders containing x and y, respectively, then $d_{l'}(x,y) = |I_x|^\alpha + |I_y|^\alpha$; the tree balance condition (1) gives us that

$$2B_L^{\alpha(l'-l+1)}|I|^\alpha \leq |I_x|^\alpha + |I_y|^\alpha \leq 2B_U^{\alpha(l'-l+1)}|I|^\alpha$$

i.e.

$$2B_L^{\alpha(l'-l+1)}|I|^\alpha \leq d_{l'}(x,y) \leq 2B_U^{\alpha(l'-l+1)}|I|^\alpha.$$

Furthermore, $d_{l'}(x,y) = 0$ if $l' < l$. Consequently, we have

$$\begin{aligned}
\rho_\alpha(x,y) &= \sum_{l' \geq l} d_{l'}(x,y) \geq \sum_{l' \geq l} 2B_L^{\alpha(l'-l+1)}|I|^\alpha \\
&= |I|^\alpha 2 \sum_{l'=1}^{\infty} B_L^{\alpha l'} = |I|^\alpha \frac{2B_L^\alpha}{1 - B_L^\alpha} \\
&= \frac{2B_L^\alpha}{1 - B_L^\alpha} d(x,y)^\alpha
\end{aligned}$$

and similarly

$$\rho_\alpha(x,y) \leq \frac{2B_U^\alpha}{1 - B_U^\alpha} d(x,y)^\alpha.$$

□

Next, suppose we fix a level l of the tree and take any probability distribution π on $\Omega \times \Omega$ with difference of marginals $p_1 - p_2$, as in condition (6). We then define probability mass functions $\tilde{\pi}$ on $\mathcal{P}_l \times \mathcal{P}_l$, and $\tilde{p}_1$ and $\tilde{p}_2$ on $\mathcal{P}_l$, by

$$\begin{aligned}
\tilde{\pi}(I,J) &= \pi(I \times J) \\
\tilde{p}_1(I) &= \int_I p_1(x)dx \\
\tilde{p}_2(I) &= \int_I p_2(x)dx
\end{aligned}$$

for all $I, J \in \mathcal{P}_l$.

Then $\tilde{\pi}$ satisfies the difference of marginals condition with $\tilde{p}_1$ and $\tilde{p}_2$, since for each $I \in \mathcal{P}_l$,

$$\begin{aligned}\sum_J (\tilde{\pi}(I,J) - \tilde{\pi}(J,I)) &= \pi(I \times \Omega) - \pi(\Omega \times I) \\ &= \int_I p_1(x)dx - \int_I p_2(x)dx \\ &= \tilde{p}_1(I) - \tilde{p}_1(I)\end{aligned} \tag{9}$$

where the second equality follows from the difference of marginals condition (6) that π is assumed to satisfy.

Also define the metric $\tilde{d}_l$ on $\mathcal{P}_l$ by

$$\tilde{d}_l(I,J) = |I|^\alpha + |J|^\alpha;$$

the proof that this is a metric on $\mathcal{P}_l$ is identical to the proof that d_l is a pseudo-metric on Ω. We then have:

$$\begin{aligned}\int_\Omega \int_\Omega d_l(x,y)d\pi(x,y) &= \sum_I \sum_J \int_{I\times J} d_l(x,y)d\pi(x,y) \\ &= \sum_{I \neq J} \int_{I\times J} (|I|^\alpha + |J|^\alpha) d\pi(x,y) \\ &= \sum_{I \neq J} \tilde{\pi}(I,J)(|I|^\alpha + |J|^\alpha) \\ &= \sum_I \sum_J \tilde{\pi}(I,J)\tilde{d}_l(I,J).\end{aligned}$$

Furthermore, using the difference of marginals condition satisfied by $\tilde{\pi}$ we have the inequality:

$$\begin{aligned}\sum_{I\in\mathcal{P}_l} |I|^\alpha |\tilde{p}_1(I) - \tilde{p}_2(I)| &= \sum_{I\in\mathcal{P}_l} |I|^\alpha \left| \sum_J \tilde{\pi}(I,J) - \sum_J \tilde{\pi}(J,I) \right| \\ &= \sum_{I\in\mathcal{P}_l} |I|^\alpha \left| \sum_{J\neq I} \tilde{\pi}(I,J) - \sum_{J\neq I} \tilde{\pi}(J,I) \right|\end{aligned}$$

(since the terms with $\tilde{\pi}(I,I)$ cancel from both sums)

$$\begin{aligned}&\leq \sum_{I\in\mathcal{P}_l} |I|^\alpha \sum_{J\neq I} \tilde{\pi}(I,J) + \sum_{I\in\mathcal{P}_l} |I|^\alpha \sum_{J\neq I} \tilde{\pi}(J,I) \\ &= \sum_{I\in\mathcal{P}_l} |I|^\alpha \sum_{J\neq I} \tilde{\pi}(I,J) + \sum_{J\in\mathcal{P}_l} |J|^\alpha \sum_{I\neq J} \tilde{\pi}(I,J)\end{aligned}$$

(by a change of variable)

$$\begin{aligned}
&= \sum_{I\in\mathcal{P}_l} |I|^\alpha \sum_{J\neq I} \tilde{\pi}(I,J) + \sum_{I\in\mathcal{P}_l} |J|^\alpha \sum_{J\neq I} \tilde{\pi}(I,J) \\
&= \sum_{I\neq J} \tilde{\pi}(I,J)(|I|^\alpha + |J|^\alpha) \\
&= \sum_{I\in\mathcal{P}_l} \sum_{J\in\mathcal{P}_l} \tilde{\pi}(I,J)\tilde{d}_l(I,J).
\end{aligned}$$

Consequently, we have

$$\begin{aligned}
&\inf\left\{ \int_\Omega \int_\Omega d_l(x,y)d\pi(x,y) : \pi \text{ satisfying (6)} \right\} \\
&\qquad \geq \inf\left\{ \sum_{I\in\mathcal{P}_l} \sum_{J\in\mathcal{P}_l} \tilde{\pi}(I,J)\tilde{d}_l(I,J) : \tilde{\pi} \text{ satisfying (9)} \right\} \\
&\qquad \geq \sum_{I\in\mathcal{P}_l} |I|^\alpha |\tilde{p}_1(I) - \tilde{p}_2(I)| \\
&\qquad = \sum_{I\in\mathcal{P}_l} |I|^{\alpha+1} |m(p_1 - p_2, I)|.
\end{aligned}$$

Now sum each side over all $l \geq 1$. Using the equivalence of the metrics ρ_α and d^α, we get, ignoring constant factors:

$$\begin{aligned}
D_1(p_1,p_2) &= \inf\left\{ \int_\Omega \int_\Omega d(x,y)^\alpha d\pi(x,y) : \pi \text{ satisfying (6)} \right\} \\
&\gtrsim \inf\left\{ \int_\Omega \int_\Omega \rho_\alpha(x,y) d\pi(x,y) : \pi \text{ satisfying (6)} \right\} \\
&= \inf\left\{ \int_\Omega \int_\Omega \sum_{l=1}^\infty d_l(x,y) d\pi(x,y) : \pi \text{ satisfying (6)} \right\} \\
&\geq \sum_{l=1}^\infty \inf\left\{ \int_\Omega \int_\Omega d_l(x,y) d\pi(x,y) : \pi \text{ satisfying (6)} \right\} \\
&\geq \sum_{l=1}^\infty \sum_{I\in\mathcal{P}_l} |I|^{\alpha+1} |m(p_1 - p_2, I)| \\
&= \sum_{I\in\mathcal{T}} |I|^{\alpha+1} |m(p_1 - p_2, I)| \\
&= D_3(p_1,p_2)
\end{aligned}$$

(note we can start exclude the term $I = \Omega$ since $m(p_1 - p_2, \Omega) = 0$). So we have shown $D_3(p_1,p_2) \lesssim D_1(p_1,p_2)$.

For the reverse direction, we will prove $D_2(p_1, p_2) \lesssim D_3(p_1, p_2)$; since Theorem 4 tells us D_1 and D_2 are equivalent, this will show all three metrics are equivalent.

For a folder $J \in \mathcal{P}_{l-1}$, denote by $sub(J)$ the set of folders $I \subset J$ contained in $\mathcal{P}_l$, i.e. the set of J's children. With this notation:

$$\begin{aligned}\sum_{I\in\mathcal{P}_l} |I|^{\alpha+1}|m(p_1-p_2,I)| &= \sum_{J\in\mathcal{P}_{l-1}} \sum_{I\in sub(J)} |I|^{\alpha+1}|m(p_1-p_2,I)| \\ &\geq \sum_{J\in\mathcal{P}_{l-1}} B_L^\alpha |J|^\alpha \sum_{I\in sub(J)} |I||m(p_1-p_2,I)|\end{aligned}$$

Furthermore, we have

$$\begin{aligned}\int_J \big|\mathbb{E}_l(p_1) - \mathbb{E}_l(p_2)\big| &= \sum_{I\in sub(J)} \int_I \big|\mathbb{E}_l(p_1) - \mathbb{E}_l(p_2)\big| \\ &= \sum_{I\in sub(J)} \int_I \left|\frac{1}{|I|}\int_I p_1(x)dx - \frac{1}{|I|}\int_I p_2(x)dx\right| dx' \\ &= \sum_{I\in sub(J)} \left|\int_I p_1 - \int_I p_2\right| \\ &= \sum_{I\in sub(J)} |I||m(p_1-p_2,I)|\end{aligned}$$

and so

$$\sum_{I\in\mathcal{P}_l} |I|^{\alpha+1}|m(p_1-p_2,I)| \geq \sum_{J\in\mathcal{P}_{l-1}} B_L^\alpha |J|^\alpha \int_J \big|\mathbb{E}_l(p_1) - \mathbb{E}_l(p_2)\big|. \tag{10}$$

Recall that $I(\psi)$ denotes the folder supporting the wavelet ψ; if $J \in \mathcal{P}_{l-1}$ and $I(\psi) = J$, then ψ is constant on each subfolder of J, and it is easy to see that for any function f, $\langle \mathbb{E}_l f, \psi\rangle = \langle f, \psi\rangle$; we therefore have

$$\begin{aligned}\sum_{\psi:I(\psi)=J} \big|\langle p_1 - p_2, \psi\rangle\big| &= \sum_{\psi:I(\psi)=J} \big|\langle \mathbb{E}_l(p_1) - \mathbb{E}_l(p_2), \psi\rangle\big| \\ &\leq \sum_{\psi:I(\psi)=J} ||\psi||_\infty \int_J \big|\mathbb{E}_l(p_1) - \mathbb{E}_l(p_2)\big| \\ &\lesssim |J|^{-1/2} \int_J \big|\mathbb{E}_l(p_1) - \mathbb{E}_l(p_2)\big|\end{aligned}$$

where the last inequality (in which we have suppressed the constant) follows from the L^∞ bounds for the wavelets and the fact that the number of wavelets supported on J

is no more than $1/B_L - 1$. Multiplying the inequality by $|J|^{\alpha+1/2}$, summing over all $J \in \mathcal{P}_l$, and using (10) then yields

$$\begin{aligned}\sum_{J\in\mathcal{P}_{l-1}} |J|^{\alpha+1/2} \sum_{\psi:I(\psi)=J} \big|\langle p_1 - p_2, \psi\rangle\big| &\lesssim \sum_{J\in\mathcal{P}_{l-1}} |J|^{\alpha} \int_J \big|\mathbb{E}_l(p_1) - \mathbb{E}_l(p_2)\big| \\ &\lesssim \sum_{I\in\mathcal{P}_l} |I|^{\alpha+1} |m(p_1 - p_2, I)|.\end{aligned}$$

Now we sum this inequality over all $l \geq 1$ to get

$$\begin{aligned}D_2(p_1,p_2) = \sum_{\psi} |I(\psi)|^{\alpha+\frac{1}{2}} |\langle p_1 - p_2, \psi\rangle| &= \sum_{l=1}^{\infty} \sum_{J\in\mathcal{P}_{l-1}} |J|^{\alpha+1/2} \sum_{\psi:I(\psi)=J} |\langle p_1 - p_2, \psi\rangle| \\ \lesssim \sum_{l=1}^{\infty} \sum_{I\in\mathcal{P}_l} |I|^{\alpha+1} |m(p_1 - p_2, I)| &= D_3(p_1,p_2)\end{aligned}$$

completing the proof.

Like D_2, the metric D_3 is extremely simple to compute in practice; simply take the average of $p_1 - p_2$ on each folder I, multiply by $|I|^{\alpha+1}$, and add up over the folders. Furthermore, if $\#\Omega = N$, it is easy to see that computing all the averages requires only $\mathcal{O}(N)$ operations (the constant factor depends on the balance constants of the tree), as to get the average of $p_1 - p_2$ on a folder I, one only needs to take a weighted average of the averages on the subfolders of I (where the weight on subfolder $J \in sub(I)$ is $|J|/|I|$).

6 Fourth Metric: Difference of Averages on Each Folder

We now introduce another metric equivalent to the three already introduced. Recall that $m(f, I)$ denotes the average of a function f over the set I, and for a folder $I \in \mathcal{T}$, $sub(I)$ denotes the set of its immediate subfolders. We then define

$$D_4(p_1,p_2) = \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} |m(p_1 - p_2, I) - m(p_1 - p_2, J)| \tag{11}$$

We first note that in the case of perfectly balanced binary trees, i.e. trees where $B_U = B_L = 1/2$, this metric is equal to exactly twice the wavelet metric D_2. To see this, we will write the unique Haar function supported on folder I as $h_I(x) = \frac{1}{|I|^{1/2}}(\chi_{I_+}(x) - \chi_{I_-}(x))$, where I_+ and I_- are the two subfolders of I (the choice of sign makes no difference to the definition of the metric D_2). For any function f, we have

$$\begin{aligned}\int_I f(x)dx - 2\int_{I_+} f(x)dx &= \int_{I_+} f(x)dx + \int_{I_-} f(x)dx - 2\int_{I_+} f(x)dx \\ &= \int_{I_-} f(x)dx - \int_{I_+} f(x)dx \\ &= -|I|^{1/2}\langle f, h_I\rangle\end{aligned}$$

and similarly

$$\int_I f(x)dx - 2\int_{I_-} f(x)dx = |I|^{1/2}\langle f, h_I\rangle.$$

Consequently, in this case we have

$$\begin{aligned}
D_4(p_1,p_2) &= \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} |m(p_1-p_2,I) - m(p_1-p_2,J)| \\
&= \sum_{l=0}^{\infty} 2^{-l(\alpha+1)} \sum_{I\in\mathcal{P}_l} \left(\left| \frac{1}{|I|}\int_I p_1(x)-p_2(x)dx - \frac{1}{|I_+|}\int_{I_+} p_1(x)-p_2(x)dx \right| + \right. \\
&\qquad \left. + \left| \frac{1}{|I|}\int_I p_1(x)-p_2(x)dx - \frac{1}{|I_-|}\int_{I_-} p_1(x)-p_2(x)dx \right| \right) \\
&= \sum_{l=0}^{\infty} 2^{-l\alpha} \sum_{I\in\mathcal{P}_l} \left(\left| \int_I p_1(x)-p_2(x)dx - 2\int_{I_+} p_1(x)-p_2(x)dx \right| + \right. \\
&\qquad \left. + \left| \int_I p_1(x)-p_2(x)dx - 2\int_{I_-} p_1(x)-p_2(x)dx \right| \right)
\end{aligned}$$

(since each folder I at level l has size 2^{-l})

$$= \sum_{l=0}^{\infty} 2^{-l\alpha} \sum_{I\in\mathcal{P}_l} 2|I|^{1/2}|\langle p_1-p_2, h_I\rangle|$$

(by the computation above)

$$\begin{aligned}
&= 2\sum_{l=0}^{\infty}\sum_{I\in\mathcal{P}_l} |I|^{\alpha+1/2}|\langle p_1-p_2, h_I\rangle| \\
&= 2D_2(p_1,p_2).
\end{aligned}$$

Furthermore, for any perfectly balanced $M-$ary tree, i.e. any tree with $B_U = B_L = 1/M$, we can derive another expression for D_4 in terms of the martingale difference

operators Δ_l. Let $p(x) = p_1(x) - p_2(x)$. We then have

$$\begin{aligned}
D_4(p_1,p_2) &= \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} |m(p_1-p_2,I) - m(p_1-p_2,J)| \\
&= \sum_{l=0}^{\infty} M^{-l(\alpha+1)} \sum_{I\in\mathcal{P}_l} \sum_{J\in sub(I)} |\Delta_l p(J)| \\
&= M \sum_{l=0}^{\infty} M^{-l\alpha} \sum_{I\in\mathcal{P}_l} \sum_{J\in sub(I)} |J||\Delta_l p(J)| \\
&= M \sum_{l=0}^{\infty} M^{-l\alpha} \sum_{I\in\mathcal{P}_l} \sum_{J\in sub(I)} \int_J |\Delta_l p| \\
&= M \sum_{l=0}^{\infty} M^{-l\alpha} \sum_{I\in\mathcal{P}_l} \int_I |\Delta_l p| \\
&= M \sum_{l=0}^{\infty} M^{-l\alpha} ||\Delta_l(p_1-p_2)||_1.
\end{aligned}$$

Compare this expression to expression (8) for D_3 in the case of $M-$ary trees.

In the case of a general tree, we will prove

Theorem 6. *The metric D_4 is equivalent to the metrics D_1, D_2, and D_3.*

Showing $D_4 \lesssim D_3$ is straightforward. We have

$$\begin{aligned}
D_4(p_1,p_2) &= \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} |m(p_1-p_2,I) - m(p_1-p_2,J)| \\
&\le \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} (|m(p_1-p_2,I)| + |m(p_1-p_2,J)|) \\
&= \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} |m(p_1-p_2,I)| + \sum_{I\in\mathcal{T}} |I|^{\alpha+1} \sum_{J\in sub(I)} |m(p_1-p_2,J)| \\
&\le \frac{1}{B_L} \sum_{I\in\mathcal{T}} |I|^{\alpha+1} |m(p_1-p_2,I)| + \frac{1}{B_L^{\alpha+1}} \sum_{I\in\mathcal{T}} \sum_{J\in sub(I)} |J|^{\alpha+1} |m(p_1-p_2,J)|
\end{aligned}$$

(since the maximum number of subfolders of any folder is $\frac{1}{B_L}$.)

$$\begin{aligned}
&\le \left(\frac{1}{B_L} + \frac{1}{B_L^{\alpha+1}}\right) \sum_{I\in\mathcal{T}} |I|^{\alpha+1} |m(p_1-p_2,I)| \\
&= \left(\frac{1}{B_L} + \frac{1}{B_L^{\alpha+1}}\right) D_3(p_1,p_2)
\end{aligned}$$

Take any f that has Hölder constant 1. We can write f as a telescopic series

$$f(x) - \int_\Omega f = \sum_{l=0}^{\infty} \Delta_l f(x)$$

Let $p(x) = p_2(x) - p_1(x)$. It follows from the aforementioned properties of Δ_l that

$$\begin{aligned}\int_\Omega f(x)p(x)dx &= \int_\Omega (f(x) - \textstyle\int f)p(x)dx = \sum_{l=0}^{\infty} \int_\Omega \Delta_l f(x)p(x)dx \\ &= \sum_{l=0}^{\infty} \int_\Omega \Delta_l^2 f(x)p(x)dx = \sum_{l=0}^{\infty} \int_\Omega \Delta_l f(x)\Delta_l p(x)dx\end{aligned}$$

(using the identity $\Delta^2 = \Delta$ and the self-adjointness of Δ)

$$\begin{aligned}&\leq \sum_{l=0}^{\infty} \sum_{J\in\mathcal{P}_l} |J|^\alpha \int_J |\Delta_l p(x)|dx \\ &= \sum_{l=0}^{\infty} \sum_{J\in\mathcal{P}_l} |J|^\alpha \sum_{I\in sub(J)} |I||m(p,I) - m(p,J)|\end{aligned}$$

(from Lemma 5)

$$\begin{aligned}&\leq B_U^\alpha \sum_{l=0}^{\infty} \sum_{J\in\mathcal{P}_l} |J|^{\alpha+1} \sum_{I\in sub(J)} |m(p,I) - m(p,J)| \\ &= B_U^\alpha D_4(p_1,p_2).\end{aligned}$$

Using the Kantorovich-Rubinstein Theorem (Theorem 3), it follows that

$$D_1(p_1,p_2) = \sup\left\{ \int_\Omega f(x)p(x)dx : f \text{ with Hölder constant } 1 \right\} \leq B_U^\alpha D_4(p_1,p_2)$$

completing the proof that D_4 is equivalent to the other three metrics.

As with D_2 and D_3, the metric D_4 can be computed in time $\mathcal{O}(N)$ if Ω is a finite set with N elements.

7 Construction of a Near-Optimal Transport

We return to the metric D_1 defined as the minimal transport cost, when transforming the probability measure p_1 into p_2. We show how to construct a non-negative function π on $\Omega \times \Omega$ whose cost is within a constant factor of the minimal cost. In fact, we will show

$$cost(\pi) \lesssim D_4(p_1,p_2)$$

where

$$cost(\pi) = \int_\Omega \int_\Omega d(x,y)^\alpha \pi(x,y)dxdy.$$

The key building block of the construction are measures on $\Omega \times \Omega$ that we will denote π_F, where $F \in \mathcal{T}$ is a folder. These measures describe the movement of mass between the subfolders of F. These can be defined in any number of ways, so long as they satisfy the following two conditions:

- Let $p = p_1 - p_2$. For any folder F, divide the set $sub(F)$ of subfolders of F into two sets: $\{I_1^{(F)}, \ldots, I_m^{(F)}\}$, the folders on which $\Delta_l p$ is nonnegative; and $\{J_1^{(F)}, \ldots, J_n^{(F)}\}$, the folders on which $\Delta_l p$ is negative (since $\Delta_l p$ is constant on the level $l+1$ folders, this decomposition is well-defined). Then we require that π_F be supported on the union of the $I_i^{(F)} \times J_j^{(F)}$, and constant on each $I_i^{(F)} \times J_j^{(F)}$. In particular, π_F is supported on $F \times F$.

- For all $x \in F$, $\int_\Omega \pi_F(x,y)dy - \int_\Omega \pi_F(y,x)dy = \Delta_l p(x)$.

We construct a particular choice of π_F, where $F \in \mathcal{P}_l$, by the formula

$$\pi_F(I_i^{(F)} \times J_j^{(F)}) = \frac{(\Delta_l p)(J_j^{(F)})}{\sum_{j'}(\Delta_l p)(J_{j'}^{(F)})|J_{j'}^{(F)}|}(\Delta_l p)(I_i^{(F)}) = \frac{-(\Delta_l p)(I_i^{(F)})}{\sum_{i'}(\Delta_l p)(I_{i'}^{(F)})|I_{i'}^{(F)}|}(\Delta_l p)(J_j^{(F)}) \tag{12}$$

where $(\Delta_l p)(I_i^{(F)})$ is the unique value $\Delta_l p$ takes on folder I_i, and similarly for $J_j^{(F)}$. For pairs of points taken from other pairs of folders, set π_F to be zero. The equality of the two definitions follows from the following lemma:

Lemma 11. *For any function f on Ω, if $F \in \mathcal{P}_l$, then $\int_F \Delta_l f(x)dx = 0$.*

Proof. By definition $\Delta_l f = \mathbb{E}_{l+1} f - \mathbb{E}_l f$. Both $\mathbb{E}_{l+1} f$ and $\mathbb{E}_l f$ are constant on the subfolders $I \in sub(F)$, with values $\frac{1}{|I|}\int_I f$ and $\frac{1}{|F|}\int_F f$, respectively. Therefore

$$\begin{aligned}\int_F \Delta_l f(x)dx &= \sum_{I \in sub(F)} \int_I \left(\frac{1}{|I|}\int_I f - \frac{1}{|F|}\int_F f\right)dx = \sum_{I \in sub(F)} \left(\int_I f - \frac{|I|}{|F|}\int_F f\right) \\ &= \int_F f - \sum_{I \in sub(F)} \frac{|I|}{|F|}\int_F f = 0.\end{aligned}$$

□

Corollary 1. *For any folder $F \in \mathcal{P}_l$,*

$$\sum_j (\Delta_l f)(J_j^{(F)})|J_j^{(F)}| = -\sum_i (\Delta_l f)(I_i^{(F)})|I_i^{(F)}|.$$

Proof. Every subfolder of F is either one of the $I_i^{(F)}$ or one of the $J_j^{(F)}$. Consequently

$$\begin{aligned}0 = \int_F \Delta_l f(x)dx &= \sum_i \int_{I_i^{(F)}} \Delta_l f(x)dx + \sum_j \int_{J_j^{(F)}} \Delta_l f(x)dx \\ &= \sum_i (\Delta_l f)(I_i^{(F)})|I_i^{(F)}| + \sum_j (\Delta_l f)(J_j^{(F)})|J_j^{(F)}|.\end{aligned}$$

□

We check that our choice of π_F satisfies the two conditions given above. The first condition is true by definition. For the second condition, suppose $x \in F$, where $F \in \mathcal{P}_l$ is a folder at level l. If $x \in I_i^{(F)}$, then $\pi(y,x) = 0$ for all y, and

$$\begin{aligned}\int_\Omega \pi_F(x,y)dy - \int_\Omega \pi_F(y,x)dy &= \sum_j \int_{J_j^{(F)}} \pi_F(x,y)dy \\ &= \sum_j \int_{J_j^{(F)}} \frac{(\Delta_l p)(J_j^{(F)})}{\sum_{j'}(\Delta_l p)(J_{j'}^{(F)})|J_{j'}^{(F)}|}(\Delta_l p)(I_i^{(F)})dy \\ &= (\Delta_l p)(I_i^{(F)}) \sum_j \frac{(\Delta_l p)(J_j^{(F)})}{\sum_{j'}(\Delta_l p)(J_{j'}^{(F)})|J_{j'}^{(F)}|}|J_{j'}^{(F)}| \\ &= (\Delta_l p)(I_i^{(F)}) = \Delta_l p(x).\end{aligned}$$

A nearly identical proof holds if $x \in J_j^{(F)}$.

With π_F defined for each folder $F \in \mathcal{T}$, we define the transport π by

$$\pi(x,y) = \sum_{F \in \mathcal{T}} \pi_F(x,y). \tag{13}$$

We check that π satisfies the difference of marginals condition (6), and upper bound its cost by $D_4(p_1, p_2)$.

Take any $x \in \Omega$ and suppose $x \in F$, where $F \in \mathcal{P}_l$ is a folder at level l. Then $\pi_{F'}(x,y) = \pi_{F'}(y,x) = 0$ for all y if $F' \neq F$, $F' \in \mathcal{P}_l$. Consequently,

$$\sum_{F' \in \mathcal{P}_l} \left(\int \pi_{F'}(x,y)dy - \int \pi_{F'}(y,x)dy \right) = \int \pi_F(x,y)dy - \int \pi_F(y,x)dy = \Delta_l p(x)$$

from which it follows

$$\int_\Omega \pi(x,y)dy - \int_\Omega \pi(y,x)dy = \sum_{l \geq 0} \Delta_l p(x) = p_1(x) - p_2(x).$$

As for the cost of π, for each folder F, $\pi_F(x,y)$ is only non-zero if $x \in I_i^{(F)}$ and $y \in J_j^{(F)}$, and consequently in this case $d(x,y) = |F|$ by definition of the tree distance (since F is the smallest folder containing both x and y). Furthermore, for every $x \in I_i^{(F)}$, $\pi_F(y,x) = 0$ for all y; consequently, the second defining condition of π_F gives that

$$\int \pi_F(x,y)dy = \Delta_l p(x)$$

and so

$$\int_{I_i^{(F)}} \int_\Omega \pi(x,y)dxdy = \Delta_l p(I_i^{(F)})|I_i^{(F)}| = |\Delta_l p(I_i^{(F)})||I_i^{(F)}|.$$

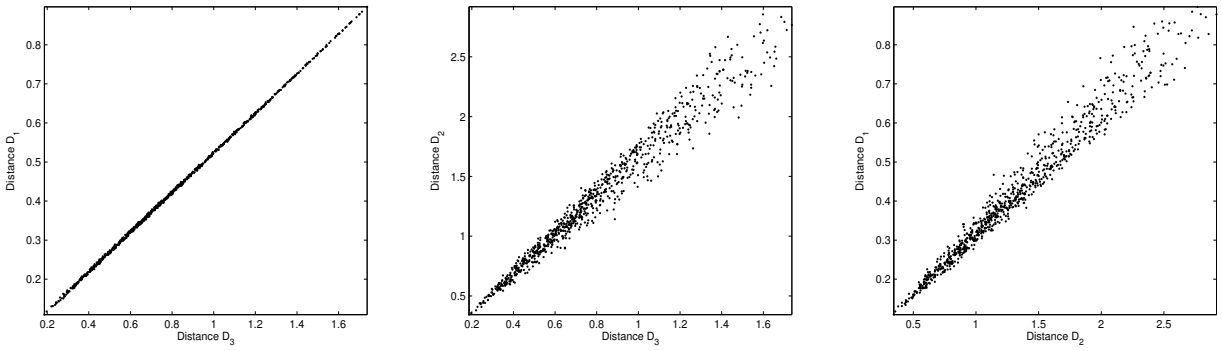

Figure 1: Pairwise comparisons of the metrics on random probability distributions, with $\alpha = 1$. From left to right, the metrics D_1 to D_3, D_2 to D_3, and D_1 to D_2

It follows that for each F

$$\begin{aligned}\int_\Omega \int_\Omega \pi_F(x,y)dxdy &= \sum_i \int_{I_i^{(F)}} \int_\Omega \pi_F(x,y)dxdy \\ &= \sum_i |\Delta_l p(I_i^{(F)})||I_i^{(F)}| \\ &= \frac{1}{2}\sum_{I \in sub(F)} |I||\Delta_l p(I)|\end{aligned}$$

where we have used Corollary 1 for the last line. Consequently

$$\begin{aligned}cost(\pi) &= \int_\Omega \int_\Omega d(x,y)^\alpha \pi_F(x,y)dxdy = \sum_{F\in\mathcal{T}} \int_\Omega \int_\Omega d(x,y)^\alpha \pi_F(x,y)dxdy \\ &= \sum_{F\in\mathcal{T}} |F|^\alpha \int_\Omega \int_\Omega \pi_F(x,y)dxdy = \sum_{F\in\mathcal{T}} |F|^\alpha \frac{1}{2} \sum_{I\in sub(F)} |I||\Delta_l p(I)| \\ &\leq \frac{B_U^\alpha}{2} \sum_{F\in\mathcal{T}} |F|^{\alpha+1} \sum_{I\in sub(F)} |\Delta_l p(I)| \\ &= \frac{B_U^\alpha}{2} \sum_{F\in\mathcal{T}} |F|^{\alpha+1} \sum_{I\in sub(F)} |m(p_1-p_2,I) - m(p_1-p_2,F)| \\ &= \frac{B_U^\alpha}{2} D_4(p_1,p_2).\end{aligned}$$

8 Experimental Results

8.1 Dyadic Tree

We ran the following experiment to see how the metrics compare in practice. We take Ω to be a set with 32 points, and the tree $\mathcal{T}$ to be a perfectly balanced binary tree. Our

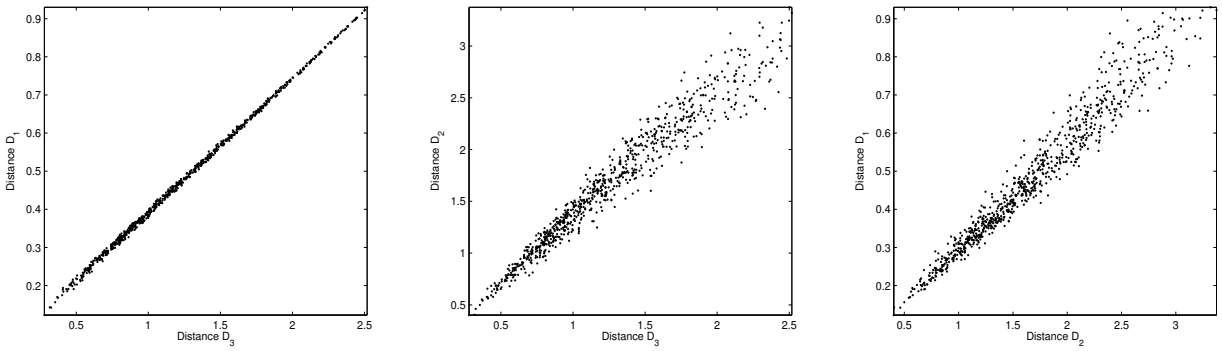

Figure 2: Pairwise comparisons of the metrics on random probability distributions, with $\alpha = .75$. From left to right, the metrics D_1 to D_3, D_2 to D_3, and D_1 to D_2

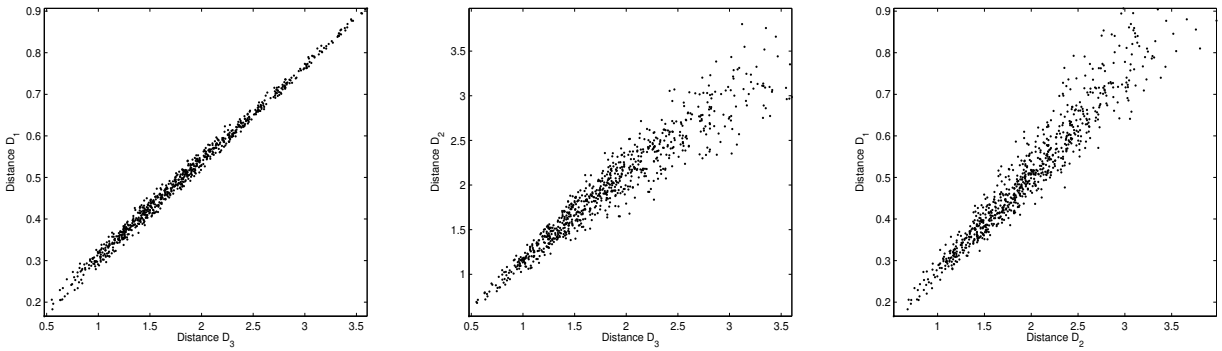

Figure 3: Pairwise comparisons of the metrics on random probability distributions, with $\alpha = .5$. From left to right, the metrics D_1 to D_3, D_2 to D_3, and D_1 to D_2

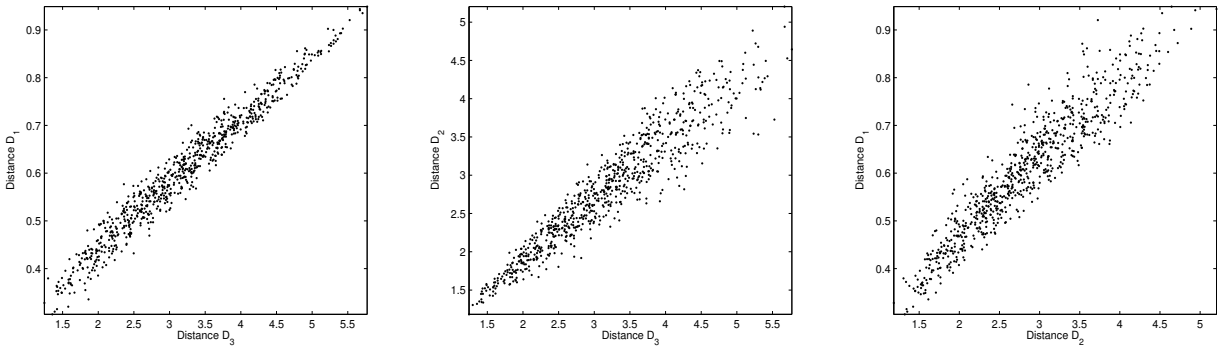

Figure 4: Pairwise comparisons of the metrics on random probability distributions, with $\alpha = .25$. From left to right, the metrics D_1 to D_3, D_2 to D_3, and D_1 to D_2

measure is normalized counting measure. It is easy to see that any probability measure on Ω can be written as a product of the form

$$p(x) = \prod_{I \in \mathcal{T}} (1 + a_I H_I(x))$$

where H_I is the L^∞ normalized Haar function on I, that is,

$$H_I(x) = \chi_{I_+}(x) - \chi_{I_-}(x)$$

and

$$-1 \leq a_I \leq 1.$$

See, for example, the paper [11].

We generated a collection of probability measures by randomly choosing the coefficients a_I uniformly from $[-1, 1]$ in the product above and computing their distance to a fixed source random measure p_1 under the metrics D_1, D_2 and D_3 (note that in this case, D_4 and D_2 are equal, up to a constant multiple, as we showed earlier). We repeated this experiment for several values of α. In the case of binary trees, the Haar functions in the definition of D_2 are unique up to sign, which does not change the definition of the metric. In the graphs, shown in Figures 1 to 4, we show scatterplots of the values of all three pairs of metrics for the different values of α.

We make several superficial observations based on the graphs. First, it appears that for $\alpha = 1$, the metrics D_1, the EMD, and the metric D_3 are nearly identical. Indeed for all values of α, these two metrics appear to be closer than any of the other pairs. However, all pairs of metrics for all values of α appear to be highly correlated, as our theory suggests they should be.

Second, the constants of equivalence get larger as α shrinks to zero, which is consistent with the bounds we derive above. For example, with the metrics D_1 (the EMD) and D_2, it follows from Theorem 4 that when $B_L = B_U = 1/2$

$$D_1(p_1, p_2) \leq D_2(p_1, p_2) \leq \frac{2}{1 - 2^{-\alpha}} D_1(p_1, p_2).$$

Since $\frac{2}{1-2^{-\alpha}}$ approaches ∞ as α goes to 0, for small α we expect the metrics to display weaker correlation.

8.2 Matrix Organization

An application area for the theory of tree hierarchical partition trees is 'coherent matrix organization' [5]. Here, the goal is to *build* trees on a data set so that certain functions we wish to predict or compress are as smooth as possible with respect to the tree metric. This is often done by simultaneously organizing the rows and columns of a matrix. Several heuristic methods are proposed in [5] for performing this organization. We do not go into details here as to our approach; we plan to present more complete results and analysis in forthcoming work. The basic idea, however, is as follows.

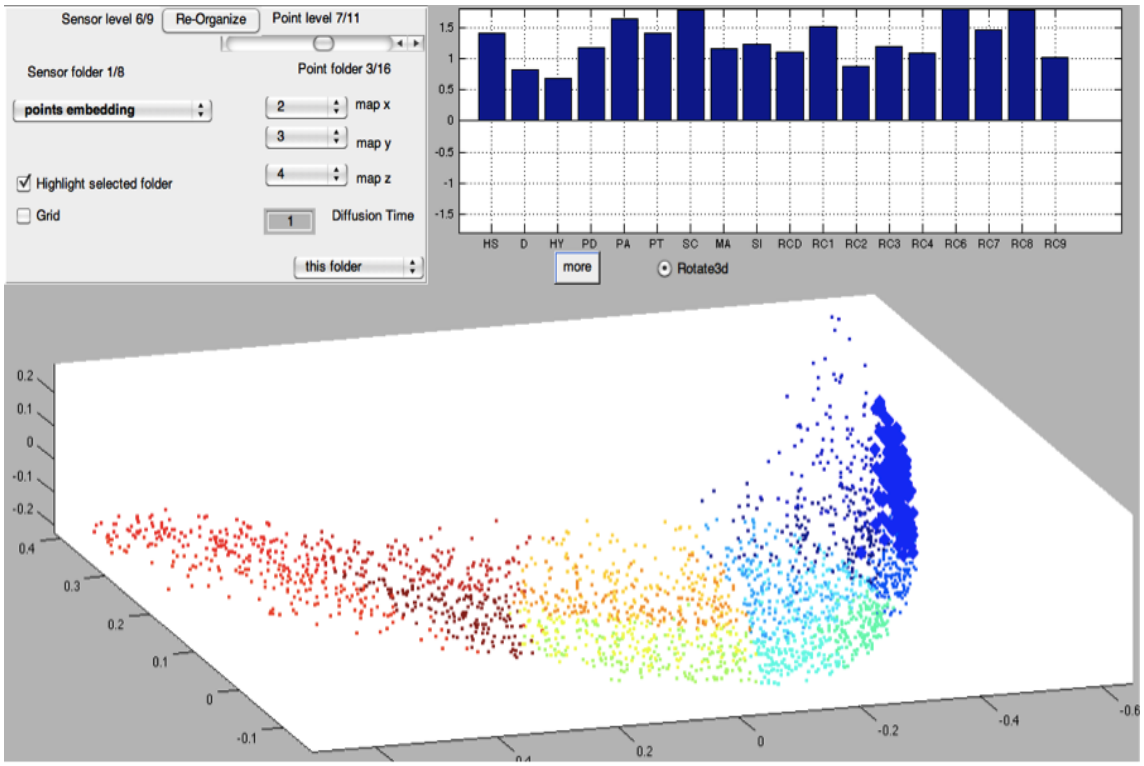

Figure 5: Snapshot of software displaying diffusion embedding of people based on responses to psychological questionnaire. The organization is derived from an affinity built using an EMD-like metric.

We suppose we are given a matrix M with an initial tree on its columns, which is either given a priori or built using a generic method. We then use this tree to organize the rows of M by viewing each row as a function on the columns, and measuring the distance between rows with an EMD-like metric. We use the distances on the rows to define affinities between the rows, which in turn are used to define a diffusion process on the rows, along the lines of [6]. The diffusion process can be used to define a tree on the rows. We then flip rows and columns and repeat the process.

We applied this process to a matrix of 1's and -1's whose entries represent yes/no responses of people (columns) to a list questions (rows) on a psychological questionnaire. In Figure 5 we show the diffusion embedding of the people after several iterations. It is apparent that an extraordinarily simple geometry is uncovered, and preliminary tests indicate that the people are positioned in this geometry according to their score on psychiatric evaluations.

9 Additional Theoretical Results

9.1 Averaging Over Trees

Suppose we have a family of trees $\mathcal{T}$, each with tree metric $d_{\mathcal{T}}(x, y)$. Consider the metric $d_{\mathcal{T}}(x, y)^\alpha$. We can define a new metric as the average over all choices of tree of these metrics, i.e. define

$$d_{ave}^{(\alpha)}(x, y) = \int_{\mathcal{T}} d_{\mathcal{T}}(x, y)^\alpha.$$

The reason for averaging over many trees is to decrease the impact that the artificial boundaries imposed by any one tree will have on the distance between two points. We trust a tree metric only when it tells us two points are close, i.e. when $d_{\mathcal{T}}(x, y)$ is small. However, if most tree metrics in a suitable family tells us that two points are far away, we will believe they are far away. Thus, we suppose that if there is a metric $\rho(x, y)$ which represents the 'true' distance between the points x and y, that there are constants $A_1, A_2 > 0$ such that $\rho(x, y)$ is bounded above by $A_2 d_{\mathcal{T}}(x, y)^\alpha$ for every tree $\mathcal{T}$, and bounded below by $A_1 d_{ave}^{(\alpha)} T(x, y)$ for every $\mathcal{T}$; i.e. we suppose that for all $\mathcal{T}$,

$$A_1 d_{ave}^{(\alpha)} T(x, y) \leq \rho(x, y) \leq A_2 d_{\mathcal{T}}(x, y)^\alpha. \tag{14}$$

One can show that this holds for the family of shifted dyadic trees on the circle, where $\rho(x, y) = |x - y|^\alpha$.

We define the earth mover's distance D_ρ with respect to the metric ρ the same way as before, but with ρ in place of the original $d = d_{\mathcal{T}}$. We can think of D_ρ as the 'true' EMD, since it employs the 'true' distance ρ as the cost function. Denote by $D_{\mathcal{T}}(p_1, p_2)$ the earth mover's distance with respect to the tree $D_{\mathcal{T}}$ (since we were only considering a single tree earlier, we suppressed the dependence of the metric on $\mathcal{T}$). We can also consider the averaging metric

$$D_{ave}(p_1, p_2) = \int_{\mathcal{T}} D_{\mathcal{T}}(p_1, p_2).$$

Under condition (14) above, the two metrics D_ρ and D_{ave} are equivalent.

Theorem 7. *The metrics D_ρ and D_{ave} are equivalent.*

Proof. By definition, $D_{\mathcal{T}}(p_1, p_2) \leq \int_\Omega \int_\Omega d_{\mathcal{T}}(x,y)^\alpha d\pi(x,y)$ for all probability measures π on $\Omega \times \Omega$ with difference of marginals $p_1 - p_2$, as in formula (6). Hence,

$$\begin{aligned} D_{ave}(p_1,p_2) &= \int_{\mathcal{T}} D_{\mathcal{T}}(p_1,p_2) \leq \int_{\mathcal{T}} \int_\Omega \int_\Omega d_{\mathcal{T}}(x,y)^\alpha d\pi(x,y) \\ &= \int_\Omega \int_\Omega \int_{\mathcal{T}} d_{\mathcal{T}}(x,y)^\alpha d\pi(x,y) \leq \int_\Omega \int_\Omega \frac{1}{A_1}\rho(x,y) d\pi(x,y) \\ &= \frac{1}{A_1} \int_\Omega \int_\Omega \rho(x,y) d\pi(x,y). \end{aligned}$$

Since this is true for all suitable π, taking the infimum we have

$$D_{ave}(p_1,p_2) \leq \frac{1}{A_1} D_\rho(p_1,p_2).$$

For the other direction, since $\rho(x,y) \leq A_2 d_{\mathcal{T}}(x,y)^\alpha$ for every tree $\mathcal{T}$ and all points $x, y \in \Omega$, for every joint distribution $\pi(x,y)$ satisfying (6)

$$\int_\Omega \int_\Omega \rho(x,y) d\pi(x,y) \leq A_2 \int_\Omega \int_\Omega d_{\mathcal{T}}(x,y)^\alpha d\pi(x,y).$$

Taking infimums over all appropriate π yields

$$D_\rho(p_1,p_2) \leq A_2 D_{\mathcal{T}}(p_1,p_2)$$

and then averaging over all $\mathcal{T}$ gives the inequality

$$D_\rho(p_1,p_2) \leq A_2 D_{ave}(p_1,p_2).$$

Putting this together with the other direction gives the full equivalence:

$$\frac{1}{A_2} D_\rho(p_1,p_2) \leq D_{ave}(p_1,p_2) \leq \frac{1}{A_1} D_\rho(p_1,p_2).$$

□

9.2 Wavelet Norm as Dual Norm to Space of Smooth Functions

Let $\mathcal{H}$ be the space of Hölder(α) functions on Ω, equipped with the semi-norm

$$||f||_{\mathcal{H}} = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x,y)^\alpha}.$$

By either restricting to mean zero functions, or declaring two functions equivalent if their difference is constant, we can regard $||\cdot||_{\mathcal{H}}$ as a norm on $\mathcal{H}$.

It can be easily shown, using results above, that this semi-norm is equivalent to the semi-norm

$$|||f||| = \sup_{\psi} |I(\psi)|^{-(\alpha+1/2)} |\langle f, \psi \rangle|.$$

Consider the subspace $\mathcal{S}$ of $\mathcal{H}$ defined by the condition

$$\lim_{l \to \infty} \sup_{\psi : I(\psi) \in \mathcal{P}_l} |I(\psi)|^{(\alpha+1/2)} |\langle f, \psi \rangle| = 0.$$

Let $\mathcal{S}^*$ denote the dual space of $\mathcal{S}$, that is, the space of continuous linear functionals on $\mathcal{S}$. We claim that the operator norm of a functional $\mu \in \mathcal{S}^*$, where we equip $\mathcal{S}$ with the norm $|||\cdot|||$, is given by

$$|||\mu|||_{op} = \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)|.$$

To prove this: first, pick any $L \geq 1$ and define

$$f_L(x) = \sum_{l=1}^{L} \sum_{\psi : I(\psi) \in \mathcal{P}_l} |I(\psi)|^{\alpha+1/2} sgn(\mu(\psi)) \psi(x).$$

It is easy to check that $f \in \mathcal{S}$, and that $|||f||| = 1$. We have

$$\begin{aligned} |||\mu|||_{op} \geq \mu(f) &= \sum_{l=1}^{L} \sum_{\psi : I(\psi) \in \mathcal{P}_l} |I(\psi)|^{\alpha+1/2} sgn(\mu(\psi)) \mu(\psi) \\ &= \sum_{l=1}^{L} \sum_{\psi : I(\psi) \in \mathcal{P}_l} |I(\psi)|^{\alpha+1/2} |\mu(\psi)|. \end{aligned}$$

Taking $L \to \infty$ proves $|||\mu|||_{op} \geq \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)|$.

For the reverse inequality: take any $f \in \mathcal{S}$. Write f as

$$\begin{aligned} f(x) &= \sum_{l=1}^{L} \sum_{\psi : I(\psi) \in \mathcal{P}_l} \langle f, \psi \rangle \psi(x) + \sum_{l=L+1}^{\infty} \sum_{\psi : I(\psi) \in \mathcal{P}_l} \langle f, \psi \rangle \psi(x) \\ &= \sum_{l=1}^{L} \sum_{\psi : I(\psi) \in \mathcal{P}_l} \langle f, \psi \rangle \psi(x) + R_L(x) \end{aligned}$$

where L is chosen large enough so that $\sup_{\psi : I(\psi) \in \mathcal{P}_l} |I(\psi)|^{(\alpha+1/2)} |\langle f, \psi, \rangle| \leq \epsilon$ for all

$l \geq L+1$, for arbitrary $\epsilon > 0$; such L exists by definition of $\mathcal{S}$. $|||R_L||| \leq \epsilon$, and we have

$$\begin{aligned}
|\mu(f)| &= \left| \sum_{l=1}^{L} \sum_{\psi: I(\psi) \in \mathcal{P}_l} \langle f, \psi \rangle \mu(\psi) + \mu(R_L) \right| \leq \sum_{l=1}^{L} \sum_{\psi: I(\psi) \in \mathcal{P}_l} |\langle f, \psi \rangle \mu(\psi)| + |\mu(R_L)| \\
&= \sum_{l=1}^{L} \sum_{\psi: I(\psi) \in \mathcal{P}_l} |I(\psi)|^{-(\alpha+1/2)} |\langle f, \psi \rangle| |I(\psi)|^{\alpha+1/2} |\mu(\psi)| + |\mu(R_L)| \\
&\leq \left(\sup_{\psi} \left\{ |I(\psi)|^{-(\alpha+1/2)} |\langle f, \psi \rangle| \right\} \right) \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)| + \epsilon |||\mu|||_{op} \\
&= |||f||| \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)| + \epsilon |||\mu|||_{op}
\end{aligned}$$

which, since ϵ is arbitrary, implies that

$$|\mu(f)| \leq |||f||| \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)|$$

and consequently that

$$|||\mu|||_{op} \leq \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)|$$

which proves

$$|||\mu|||_{op} = \sum_{\psi} |I(\psi)|^{\alpha+1/2} |\mu(\psi)|.$$

References

[1] Cha, S. *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions.* International Journal of Mathematical Models and Methods in Applied Sciences 2007, 1(4):300-307.

[2] Rubner,Y.,Tomasi, C., Guibas, L.J. *The Earth Mover's Distance as a Metric for Image Retrieval* International Journal of Computer Vision 2000; 40(2): 99-121.

[3] Gavish, M., Nadler, B., Coifman, R. *Multiscale wavelets on trees, graphs, and high-dimensional data.* Proceedings of ICML, 2010.

[4] Gavish, M., Nadler, B., Coifman, R. *Supplementary material for multiscale wavelets on trees, graphs, and high-dimensional data.* Proceedings of ICML, 2010.

[5] Gavish, M., Coifman, R.R. *Sampling, Denoising and Compression of Matrices by Coherent Matrix Organization.* Applied and Computational Harmonic Analysis 2012; 33(3): 354-369.

[6] Coifman, R., Lafon, S. *Diffusion maps.* Applied and Computational Harmonic Analysis 2006; 21(1):5-30.

[7] Shirdhonkar, S., Jacobs, D.W. *Approximate earth mover's distance in linear time.* IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[8] Villani, Cédric. *Topics in Optimal Transportation.* American Mathematical Society, 2003.

[9] Indyk, P., Thaper, N. *Fast Image Retrieval via Embeddings.* 3rd Workshop on Statistical and Computational Theories of Vision, Nice, France, 2003.

[10] Ling, H., Okada, K. *Diffusion Distance for Histogram Comparison.* IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[11] Fefferman, R.A., Kenig, C.E., Pipher, J. *The Theory of Weights and the Dirichlet Problem for Elliptic Equations.* Annals of Mathematics; 134(1): 1991.