Many databases can be described using a tensor product of metric spaces. The mixed Lipschitz condition is a natural notion of function regularity in this context, and the norm dual to the mixed Lipschitz space is a natural distance between measures. In this paper, we consider the tensor product of spaces equipped with tree metrics and give effective formulas for the mixed Lipschitz norm and its dual. We also show that these norms behave well when approximating an arbitrary metric by tree metrics.

The mixed Lipschitz space and its dual for tree metrics

William Leeb

Technical Report YALEU/DCS/TR-1502
November 21, 2014

Dept. of Mathematics, Yale University, New Haven CT 06511

Approved for public release: distribution is unlimited.
Keywords: *tree metric, tensor product, Lipschitz, mixed Lipschitz, dual space, Earth Mover's Distance, EMD*

1 Introduction

The space of Lipschitz functions defined on a metric space arises naturally in many areas of machine learning and statistics. For example, standard models in non-parametric statistics posit that unknown signals lie in a Hölder space or a more general regularity class [5, 16]. Extrapolating a function value to new points, or inferring its values from noisy samples, can only be achieved if some kind of regularity on the function is assumed, the Lipschitz condition being a natural kind of regularity.

Also of interest is the space dual to Lipschitz. The dual norm of the difference between two probability measures is equal to their Earth Mover's Distance (EMD) [17], a popular metric in areas such as image processing [15]. The dual distance provides a robust way of comparing two measures on a dataset that is insensitive to perturbations, a desirable property for many tasks.

The Lipschitz space and its dual are defined with respect to a single metric space. Many datasets, however, are not modeled well by one metric space, but rather the tensor product of several metric spaces. Consider, for example, a word-document matrix, with rows indexed by documents and columns indexed by words. The documents and words are best described by two separate geometries, each with its own metric. The natural notion of regularity for a function on the product of metric spaces is the mixed Lipschitz condition, which requires f to have bounded mixed difference quotients. We define the mixed Lipschitz space in Section 4.

The space dual to mixed Lipschitz functions is also of interest, as its norm provides a robust distance between measures on the product of metric spaces. In this paper, we study the mixed Lipschitz space and its dual when the underlying metrics are *tree metrics*, defined in Section 2. In particular, in Sections 4 and 5 we develop norms equivalent to the mixed Lipschitz norm and its dual that can be computed in linear time. In Section 6, we relate the space of mixed Lipschitz functions and its dual for tree metrics to the corresponding spaces for metrics approximated by dominating tree metrics [1].

2 Tree metrics, martingales, and martingale differences

In this section we introduce the basic notation and definitions that we will be using throughout this paper. X will denote a finite set that is equipped with a partition tree $\mathcal{T}$; that is, $\mathcal{T}$ is a collection of subsets of X such that for any two such subsets I and J, either $I \subset J$, $J \subset I$, or I and J are disjoint. We will assume that the entire set X is one of the folders in $\mathcal{T}$, as are all the singletons $\{x\}$ for $x \in X$.

We can view each folder in the tree, including the singletons, as a point in a graph, where an edge is placed between folders I and J if I is a *child* of J (we will also say J is I's *parent*); that is, $I \subset J$ and there are no folders in between I and J. In this sense, we can view the set X as being the leaves of a graph-theoretic rooted tree, the folder X being the root. Of course, given any set of leaves X of any rooted tree, we can build a partition tree by assigning to each node of the tree the folder of all leaves that branch off from that node; so graph-theoretic trees with X as the leaves and partition trees describe the same structure on X.

These two different ways of viewing trees give rise to two different notions of a tree metric, one of which is a special case of the other. In Subsection 2.1 we define these tree metrics. In Subsection 2.2, we introduce the the martingale and martingale difference operators on trees, and prove some of their basic properties. In Subsection 2.3, we generalize these operators to the product of trees.

2.1 Tree metrics

Throughout the paper, we will be considering two kinds of tree metrics, one of which is a special case of the other. The first metric arises by viewing the points X as the leaves of a graph-theoretic tree, where each edge of the tree has some positive weight attached to it. If I is a point in the tree, we will denote by e_I the weight on the edge connecting I to its parent. The distance between any two points in the graph is then the geodesic distance, which in the simple case of the tree is just the sum of the edge weights on the unique path connecting the two points. In particular, this gives rise to a distance on X.

Given two points x and y, let $\mathcal{S}_{x,y}$ denote the set of all folders that contain exactly one of x or y. The following expression for $d(x, y)$ is immediate from the definitions:

Lemma 1. *For any two points x, y in X,*

$$d(x,y) = \sum_{I \in \mathcal{S}_{x,y}} e_I.$$

Another kind of tree metric arises by taking weights not on the edges of the tree but rather on the nodes of the trees themselves - that is, on the folders in $\mathcal{T}$. We will denote by $w(I)$ the weight on the folder I. We think of $w(I)$ as being the diameter of the set I, and in the metric we define this will be the case. We therefore require that if $I \subset J$, then $w(I) \leq w(J)$. With this, we define the distance between any two points x and y to be $w(I_{x,y})$, where $I_{x,y}$ denotes the smallest folder containing both x and y.

The following lemma is also trivial:

Lemma 2. *Let $\mathcal{T}$ be a partition tree on X, and $w(I)$ be any collection of folders weights satisfying $w(I) < w(J)$ for $I \subsetneq J$, and $w(\{x\}) = 0$. Then the collection of edge weights $e_I = \frac{1}{2}(w(I') - w(I))$, where I' denotes the parent folder of I, gives rise to the same metric on X as the folder weights $w(I)$.*

In this paper, we will not discuss the important question of how to construct a partition tree $\mathcal{T}$ with edge weights e_I. The choice of tree will depend on the task at hand. For instance, much work has been done in constructing tree distances that approximate a given metric on X [1, 6, 11]; we will have more to say about this subject in Section 6. Other methods include clustering the data at different scales using a family of diffusion operators and taking the weight on a folder to be a power of its measure [7, 8]. For the remainder of this paper, we will view all trees as given and not concern ourselves with where they come from.

2.2 Martingales and martingale differences

In this section, we suppose that X is also equipped with a measure, and that every singleton $\{x\}$ has positive measure. We define the martingale and martingale difference operators and prove some of their basic properties. Given a function f and a folder I, we let $m_I f$ denote the function whose value on I is the mean value of f, and is zero outside I; that is,

$$m_I f(x) = \left(\frac{1}{|I|}\int_I f(y)dy\right)\chi_I(x).$$

We denote by

$$m_I f(I) = \frac{1}{|I|}\int_I f(y)dy$$

the unique value that the function $m_I f$ achieves on the folder I.

Also define the martingale difference operator $\Delta_I f$ by

$$\Delta_I f(x) = \sum_{J \text{ child of } I} m_J f(x) - m_I f(x).$$

Note that $\Delta_I f$ is constant on the child folders of I. If J is a child of I, we will denote by $\Delta_I f(J)$ the unique value that $\Delta_I f$ takes on J.

We prove some basic properties of the operators m_I and Δ_I that will be useful in Section 5.

Lemma 3. *For any function f and any (non-singleton) folder $I \in \mathcal{T}$,*

$$\int_X \Delta_I f(x)dx = \int_I \Delta_I f(x)dx = 0.$$

Proof. By definition, we have

$$\begin{aligned}
\int_X \Delta_I f(x)dx &= \int_X \sum_{J \text{ child of } I} m_J f(x)dx - \int_X m_I f(x)dx \\
&= \sum_{J \text{ child of } I} |J| m_J f(J) - |I| m_I f(I) \\
&= \sum_{J \text{ child of } I} \int_J f(x)dx - \int_I f(x)dx = 0 \\
&= \int_I f(x)dx - \int_I f(x)dx = 0.
\end{aligned}$$

□

Corollary 1. *For folders $I \neq J$ and any functions f, g, we have $\langle \Delta_I f, \Delta_J g\rangle = 0$.*

Proof. Clearly, if $I \cap J = \emptyset$, the supports of $\Delta_I f$ and $\Delta_J g$ are disjoint, and consequently their inner product is 0. Otherwise, suppose without loss of generality that $I \subsetneq J$. Then I is contained in (or perhaps equal to) a proper subfolder of J, and so $\Delta_J g$ is constant on the support of $\Delta_I f$. Since $\int_X \Delta_I f(x)dx = 0$, the result follows. □

Lemma 4. *The operators m_I are self-adjoint; that is,*

$$\langle m_I f, g\rangle = \langle f, m_I g\rangle.$$

Proof. We have

$$\langle m_I f, g\rangle = \int_X \left(\frac{1}{|I|}\int_I f(y)dy\right)\chi_I(x)g(x)dx = \frac{1}{|I|}\int_X\int_X f(y)\chi_I(y)\chi_I(x)g(x)dxdy.$$

Since the expression on the right is symmetric in f and g, the result follows. □

Corollary 2. *The operators Δ_I are self-adjoint; that is,*

$$\langle \Delta_I f, g\rangle = \langle f, \Delta_I g\rangle.$$

It is also easy to see the following:

Lemma 5. *For every folder $I \in \mathcal{T}$, $m_I^2 f = m_I f$.*

Proof. By definition,

$$\begin{aligned} m_I^2 f(x) &= \left(\frac{1}{|I|}\int_I m_I f(y)dy\right)\chi_I(x) = \left(\frac{1}{|I|}\int_I m_I f(I)dy\right)\chi_I(x) = m_I f(I)\chi_I(x) \\ &= \left(\frac{1}{|I|}\int_I f(y)dy\right)\chi_I(x) = m_I f(x). \end{aligned}$$

□

2.3 Product of trees

The primary concern of this paper is the product of spaces, each of which is equipped with its own partition tree. For simplicity, we will consider the case of two spaces, X and Y, with trees $\mathcal{T}_X$ and $\mathcal{T}_Y$ and edge weights e_I^X and e_J^Y, respectively.

We define the operators

$$m_{X,I} f(x,y) = \left(\frac{1}{|I|}\int_I f(x',y)dx'\right)\chi_I(x)$$

and

$$m_{Y,J} f(x,y) = \left(\frac{1}{|J|}\int_J f(x,y')dy'\right)\chi_J(y).$$

We will denote $m_{X,X}$ and $m_{Y,Y}$ by m_X and m_Y, respectively. Note that $m_X f$ is a function of the y variable alone, and $m_Y f$ is a function of the x variable alone; we will therefore also write $m_X f(y) = m_X f(x,y)$ and $m_Y f(x) = m_Y f(x,y)$.

We also define

$$\Delta_{X,I} f(x,y) = \sum_{I' \text{ child of } I} m_{X,I'} f(x,y) - m_{X,I} f(x,y)$$

and

$$\Delta_{Y,J} f(x,y) = \sum_{J' \text{ child of } J} m_{Y,J'} f(x,y) - m_{Y,J} f(x,y).$$

As for a single tree, these martingale and martingale difference operators are self-adjoint. The functions $\Delta_{X,I} f$ and $\Delta_{Y,J} f$ are also mean-zero. Furthermore, we have the identities $m_{X,I}^2 = m_{X,I}$ and $m_{Y,J}^2 = m_{Y,J}$, and

$$\langle \Delta_{X,I} f, \Delta_{X,I'} g \rangle = \langle \Delta_{Y,J} f, \Delta_{Y,J'} g \rangle = 0$$

whenever $I \neq I'$ and $J \neq J'$. The proofs of these statements are nearly identical to the corresponding results for a single tree.

3 The Lipschitz class and its dual

In this section we develop characterizations for the Lipschitz norm and its dual with respect to an arbitrary tree metric on X. Let $d(x,y)$ be a tree metric on X, with edge weights e_I. Define the L^∞ variation of a function f on X with respect to the metric $d(x,y)$ by

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

We define the Lipschitz norm of f to be

$$\|f\|_\Lambda = \max\{\|f\|_d, \|m_X f\|_\infty\}.$$

We also define the norm on the space dual to mean-zero functions of bounded L^∞ variation by

$$\|T\|_d^* = \sup_{\|f\|_d \leq 1, m_X f = 0} \langle f, T \rangle,$$

and we define the dual norm to the space of Lipschitz functions as

$$\|T\|_{\Lambda^*} = \sup_{\|f\|_\Lambda \leq 1} \langle f, T \rangle.$$

We have the following simple lemma:

Lemma 6. *For every T,*

$$\|T\|_{\Lambda^*} = \|T\|_d^* + \|m_X T\|_1.$$

Proof. Define $f_1 = f - m_X f$ and $f_2 = m_X f$. Then $f = f_1 + f_2$, and

$$\|f\|_\Lambda = \|f_1\|_d + \|m_X f_2\|_\infty.$$

We then have

$$\begin{aligned} \|T\|_{\Lambda^*} &= \sup_{\|f\|_\Lambda \leq 1} \langle f, T \rangle = \sup_{\|f_1\|_d \leq 1, |m_X f_2| \leq 1} \{\langle f_1, T \rangle + \langle f_2, T \rangle\} \\ &= \sup_{\|f\|_d \leq 1, m_X f = 0} \langle f, T \rangle + \sup_{|m_X f| \leq 1, f \text{ constant}} \langle f, T \rangle \\ &= \|T\|_d^* + \|m_X T\|_1 \end{aligned}$$

as claimed. □

Our goal in this section is to develop simple formulas for the dual norms $\|T\|_{d^*}$ and $\|T\|_{\Lambda^*}$. We will do this by use of the following formula for the Lipschitz norm $\|f\|_d$.

Theorem 1. *For any function f on X, let $\mathcal{A}_f$ denote the set of all sequences of coefficients $\{a_I\}_{I\in\mathcal{T}}$ such that*

$$f(x)=\sum_I a_I\chi_I(x).$$

We then have the following expression for $\|f\|_d$:

$$\|f\|_d=\inf_{\{a_I\}\in\mathcal{A}_f}\sup_{I\neq X}\frac{|a_I|}{e_I}.$$

Proof. Let $C_f=\inf_{\{a_I\}\in\mathcal{A}_f}\sup_{I\neq X}\frac{|a_I|}{e_I}$. Suppose first that we have written $f=\sum_I a_I\chi_I$. Take any two points x and y in X, and denote by $I_{x,y}$ the smallest folder containing both points. Then $\chi_I(x)=\chi_I(y)$ if either $I\supset I_{x,y}$ or I is disjoint from $I_{x,y}$; consequently,

$$\begin{aligned}f(x)-f(y)&=\sum_{I\subsetneq I_{x,y};x\in I}a_I-\sum_{I\subsetneq I_{x,y};y\in I}a_I\\&\le C_f\Bigg\{\sum_{I\subsetneq I_{x,y};x\in I}e_I+\sum_{I\subsetneq I_{x,y};y\in I}e_I\Bigg\}=C_f d(x,y)\end{aligned}$$

which shows that $\|f\|_d\le C_f$.

For the other direction, let $\bar{f}$ denote any extension of f to *all* nodes of the tree (that is, $\bar{f}$ is a function on the set of all folders in $\mathcal{T}$) that has the same variation as f; in other words, $\|\bar{f}\|_d=\|f\|_d$, where

$$\|\bar{f}\|_d=\sup_{I\neq J}\frac{\bar{f}(I)-\bar{f}(J)}{d(I,J)}$$

the supremum being over all distinct folders I and J in the tree. A simple formula for one choice of $\bar{f}$ is given in the paper [12].

Then if we let I' denote the parent of the folder I, we can write f as the telescopic sum

$$f=\sum_{I\neq X}(\bar{f}(I)-\bar{f}(I'))\chi_I+\bar{f}(X)\equiv\sum_{I\neq X}a_I\chi_I+\bar{f}(X).$$

Since $\|\bar{f}\|_d=\|f\|_d$, $|a_I|=|\bar{f}(I)-\bar{f}(I')|\le e_I$ for all $I\neq X$, which shows $C_f\le\|f\|_d$ and completes the proof. □

Corollary 3. *We have the following upper and lower bounds for $\|f\|_d$:*

$$\sup_I\frac{\|\Delta_I f\|_\infty}{\mathrm{diam}(I)}\le\|f\|_d\le\sup_{I\neq X}\frac{|\Delta_{I'}f(I)|}{e_I}.$$

The supremum on the left is over all non-singleton folders I.

Proof. Take any folder I and let I' denote its parent; then for any $x, y \in I'$, we have $|f(x) - f(y)| \leq \|f\|_d \operatorname{diam}(I')$. Therefore

$$\begin{aligned}
|m_I(f) - m_{I'}(f)| &= \left| \frac{1}{|I|} \int_I f(x)dx - \frac{1}{|I'|} \int_{I'} f(y)dy \right| \\
&= \left| \frac{1}{|I'|} \int_{I'} \frac{1}{|I|} \int_I f(x)dxdy - \frac{1}{|I|} \int_I \frac{1}{|I'|} \int_{I'} f(y)dydx \right| \\
&= \left| \frac{1}{|I'|} \int_{I'} \frac{1}{|I|} \int_I (f(x) - f(y))dxdy \right| \\
&\leq \frac{1}{|I'|} \frac{1}{|I|} \int_{I'} \int_I \|f\|_d \operatorname{diam}(I')dxdy = \|f\|_d \operatorname{diam}(I').
\end{aligned}$$

Dividing each side by $\operatorname{diam}(I')$ and taking the supremum over all I gives the leftmost inequality.

For the other side, we make use of Theorem 1. For each folder $I \neq X$, let I' denote its parent, and define $a_I = \Delta_{I'} f(I)$. It is easy to see that, up to an additive constant,

$$f = \sum_{I \neq X} a_I \chi_I$$

and consequently Theorem 1 yields

$$\|f\|_d \leq \sup_{I \neq X} \frac{|a_I|}{e_I} \leq \sup_{I \neq X} \frac{|\Delta_{I'} f(I)|}{e_I}$$

completing the proof. □

We can use the expression for $\|f\|_d$ from Theorem 1 to derive a very simple formula for $\|T\|_d^*$.

Theorem 2. *For any L^1 measure T, we have*

$$\|T\|_d^* = \sum_{I \neq X} e_I |\langle \chi_I, T \rangle|. \tag{1}$$

Note that $|\langle \chi_I, T \rangle| = |I|(m_I T)(I)$.

Proof. Take any function f with $\|f\|_d \leq 1$. By the previous theorem, we can write

$$f = \sum_I a_I \chi_I$$

where $1 \geq \|f\|_d = \sup_{I \neq X} |a_I|/e_I$. Since f has mean zero, we can assume without loss of generality that T has total measure zero when taking the inner product. Therefore, we have

$$|\langle f, T \rangle| = \left| \sum_{I \neq X} a_I \langle \chi_I, T \rangle \right| \leq \sum_{I \neq X} \frac{|a_I|}{e_I} e_I |\langle \chi_I, T \rangle| \leq \sum_{I \neq X} e_I |\langle \chi_I, T \rangle|$$

and taking the supremum over all f yields $\|T\|_d^* \leq \sum_{I\neq X} e_I|\langle \chi_I, T\rangle|$.

For the other inequality, define the function $\tilde{f}$ by

$$\tilde{f} = \sum_{I\neq X} e_I \operatorname{sgn}(\langle \chi_I, T\rangle)\chi_I + K$$

where K ensures that $\tilde{f}$ has mean zero. The previous theorem shows that $\|\tilde{f}\|_d = 1$. Again, since $\tilde{f}$ has mean zero, we can assume T also has total measure zero as well when taking the inner product. Therefore,

$$\|T\|_d^* \geq \langle \tilde{f}, T\rangle = \sum_{I\neq X} e_I \operatorname{sgn}(\langle \chi_I, T\rangle)\langle \chi_I, T\rangle = \sum_{I\neq X} e_I|\langle \chi_I, T\rangle|$$

which completes the proof. □

Corollary 4. *For every L^1 measure T on X, its dual Lipschitz norm $\|T\|_{\Lambda^*}$ is equal to*

$$\|T\|_{\Lambda^*} = \sum_{I\neq X} e_I|\langle \chi_I, T\rangle| + \|m_X T\|_1.$$

Remark 1. Theorem 2 can be easily derived from the formula for Earth Mover's Distance given in [3], using the fact that when T is the difference of two probability measures, $\|T\|_d^*$ is equal to the Earth Mover's Distance between them; this is the content of the Kantorovich-Rubinstein Theorem [17]. The proof we give here, however, appears to be new. Furthermore, we will use Theorem 1 in Section 5 to derive equivalent formulas for the Lipschitz and mixed Lipschitz norms on a special class of trees.

Remark 2. The formula for $\|T\|_d^*$ from Theorem 2 can be computed in cost proportional to the size of X. To see this, first observe that the number of folders in any partition tree $\mathcal{T}$ on a set of size N cannot exceed $2N-1$. Furthermore, to compute each term $\langle \chi_I, T\rangle$ that appears on the right side of (1), we need to compute the integral of T on I; we can do this by simply adding up its integral over each of the children of I. Consequently, we can compute all the terms $\langle \chi_I, T\rangle$ by starting with the N integrals of T over the singletons, and then recursively computing the integral of T over a folder I by adding up its integral over the children of I. Each folder is only touched once, in the computation of its parent's integral; and so the total cost is linear in N.

4 The mixed Lipschitz space and its dual for general trees

The characterizations of the Lipschitz space and its dual can be extended to characterizations of the space of mixed Lipschitz functions and its dual. Our setting here is the product of two spaces, X and Y, each equipped with its own partition tree $\mathcal{T}_X$ and $\mathcal{T}_Y$ with weights e_I^X, e_J^Y and corresponding metrics $d_X(x,x'), d_Y(y,y')$, respectively.

The mixed variation $\|f\|_{d_X,d_Y}$ of a function f on $X\times Y$ is defined by

$$\|f\|_{d_X,d_Y} = \sup_{x\neq x', y\neq y'} \frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{d_X(x,x')d_Y(y,y')}.$$

We then define the mixed Lipschitz norm of f to be

$$\|f\|_{\Lambda_{X,Y}} = \max\{\|f\|_{d_X,d_Y}, \|m_X f\|_{d_Y}, \|m_Y f\|_{d_X}, \|m_X m_Y f\|_\infty\}.$$

Note that, since $m_Y f$ is a function on X alone and $m_X f$ is a function on Y alone, the notation we use is sensible.

We define the corresponding dual norms. First, we consider the norm dual to the space of functions of bounded mixed difference quotients and zero marginals:

$$\|T\|^*_{d_X,d_Y} = \sup\{\langle f, T\rangle : \|f\|_{d_X,d_Y} \le 1, m_Y f = 0, m_X f = 0\}$$

The dual norm of an L^1 measure T to the space of mixed Lipschitz functions is defined as

$$\|T\|_{\Lambda^*_{X,Y}} = \sup_{\|f\|_{\Lambda_{X,Y}} \le 1} \langle f, T\rangle.$$

We then have the following lemma:

Lemma 7. *For any distribution T on $X \times Y$,*

$$\|T\|_{\Lambda^*_{X,Y}} = \|T\|^*_{d_X,d_Y} + \|m_Y T\|^*_{d_X} + \|m_X T\|^*_{d_Y} + \|m_X m_Y T\|_1.$$

Proof. For any function f, let $f_1 = f - m_X f - m_Y f + m_X m_Y f$, $f_2 = (m_Y - m_X m_Y) f$, $f_3 = (m_X - m_Y m_X) f$, and $f_4 = m_X m_Y f$. It is easy to see that $f = f_1 + f_2 + f_3 + f_4$, and that

$$\|f\|_{\Lambda_{X,Y}} = \max\{\|f_1\|_{d_X,d_Y}, \|f_2\|_{d_X}, \|f_3\|_{d_Y}, \|m_X m_Y f_4\|_\infty\}.$$

Consequently, we can write

$$\begin{aligned}
&\sup_{\|f\|_{\Lambda_{X,Y}} \le 1} \langle f, T\rangle \\
=&\sup_{\|f_1\|_{d_X,d_Y} \le 1} \langle f_1, T\rangle + \sup_{\|f_2\|_{d_X} \le 1} \langle f_2, T\rangle + \sup_{\|f_3\|_{d_Y} \le 1} \langle f_3, T\rangle + \sup_{|m_X m_Y f_4| \le 1} \langle f_4, T\rangle \\
=&\|T\|^*_{d_X,d_Y} + \|m_Y T\|^*_{d_X} + \|m_X T\|^*_{d_Y} + \|m_X m_Y T\|_1
\end{aligned}$$

which is the desired equality. □

From Section 3, we have formulas for $\|m_Y T\|^*_{d_X}$, $\|m_X T\|^*_{d_Y}$ and $\|m_X m_Y T\|_1$ that can be computed at cost proportional to the size of $X \times Y$. We now turn to the computation of $\|T\|^*_{d_X,d_Y}$. We give a formula that approximates $\|T\|^*_{d_X,d_Y}$ and that can be computed in linear time as well, and whose distortion is bounded by a universal constant independent of T or the tree. As in the proof of Theorem 2, which depended on a formula for the Lipschitz norm of mean zero functions, this formula is derived from a characterization of mixed Lipschitz functions with zero marginals, which we present now.

Theorem 3. *For any function f on $X \times Y$, let $\mathcal{A}_f$ denote the collection of the sets of all coefficients $a_{I\times J}$ such that*

$$f(x,y) = \sum_{I\in\mathcal{T}_X}\sum_{J\in\mathcal{T}_Y} a_{I\times J}\chi_I(x)\chi_J(y).$$

Then there is a universal constant C, independent of the trees $\mathcal{T}_X$ and $\mathcal{T}_Y$ and the function f, such that

$$\|f\|_{d_X,d_Y} \leq \inf_{\{a_{I\times J}\}\in\mathcal{A}_f}\ \sup_{I\neq X, J\neq Y} \frac{|a_{I\times J}|}{e_I^X e_J^Y} \leq C\|f\|_{d_X,d_Y}.$$

Proof. Take any $\{a_{I\times J}\}\in\mathcal{A}_f$. Then

$$\begin{aligned}
&f(x,y) - f(x,y') - f(x',y) + f(x',y') \\
&= \sum_{I\in\mathcal{T}_X}\sum_{J\in\mathcal{T}_Y} a_{I\times J}(\chi_I(x)\chi_J(y) - \chi_I(x)\chi_J(y') - \chi_I(x')\chi_J(y) + \chi_I(x')\chi_J(y')) \\
&= \sum_{I\in\mathcal{S}_{x,x'}}\sum_{J\in\mathcal{S}_{y,y'}} a_{I\times J}(\chi_I(x) - \chi_I(x'))(\chi_J(y) - \chi_J(y')) \\
&\leq \sup_{I\neq X, J\neq Y}\frac{|a_{I\times J}|}{e_I^X e_J^Y}\sum_{I\in\mathcal{S}_{x,x'}} e_I^X \sum_{J\in\mathcal{S}_{y,y'}} e_J^Y = \sup_{I\neq X, J\neq Y}\frac{|a_{I\times J}|}{e_I^X e_J^Y} d_X(x,x')d_Y(y,y').
\end{aligned}$$

This proves that $\|f\|_{d_X,d_Y} \leq \inf_{\{a_{I\times J}\}\in\mathcal{A}_f}\sup_{I\neq X,J\neq Y}\frac{|a_{I\times J}|}{e_I^X e_J^Y}$.

For the other inequality, we will show that we can extend the function f defined on $X\times Y$ to a function $\bar{f}$ defined on $\mathcal{T}_X \times \mathcal{T}_Y$, where the mixed variation of $\bar{f}$ is no more than C times $\|f\|_{d_X,d_Y}$. In other words, the function $\bar{f}$ will satisfy

$$|\bar{f}(I,J) - \bar{f}(I,J') - \bar{f}(I',J) + \bar{f}(I',J')| \leq C\|f\|_{d_X,d_Y} e_I^X e_J^Y \tag{2}$$

where I' denotes the parent of I, and J' the parent of J.

If we had such an extension, we would be finished, for after adjusting the marginals of f (which do not affect the norm $\|f\|_{d_X,d_Y}$), we could expand f in the double telescopic sum

$$f(x,y) = \sum_{I\neq X}\sum_{J\neq Y}(\bar{f}(I,J) - \bar{f}(I,J') - \bar{f}(I',J) + \bar{f}(I',J'))\chi_I(x)\chi_J(y).$$

Taking $\tilde{a}_{I\times J} = \bar{f}(I,J) - \bar{f}(I,J') - \bar{f}(I',J) + \bar{f}(I',J')$, (2) shows $|\tilde{a}_{I\times J}| \leq C\|f\|_{d_X,d_Y} e_I^X e_J^Y$; consequently,

$$\inf_{\{a_{I\times J}\}\in\mathcal{A}_f}\ \sup_{I\neq X, J\neq Y}\frac{|a_{I\times J}|}{e_I^X e_J^Y} \leq C\|f\|_{d_X,d_Y}$$

which is the desired result.

We now show how to prove the existence of such an extension $\bar{f}$. First, by adjusting the marginals of f we can assume without loss of generality that there is a point $y_0 \in Y$ such that $f(x,y_0) = 0$ for all x. We now interpret the function f as a map not from

$X \times Y$ in to $\mathbb{R}$, but rather from X into the space $\mathrm{Lip}_0(Y)$ of Lipschitz functions g on Y that are zero at y_0, equipped with the Lipschitz norm $\|g\|_{d_Y}$. More formally, for any $x \in X$, define the function $f_x(y) = f(x, y)$ and the map

$$F : X \to \mathrm{Lip}_0(Y), \quad x \mapsto f_x.$$

The key observation is that the mixed Lipschitz norm $\|f\|_{d_X,d_Y}$ of f is an upper bound on the Lipschitz norm of F, since by definition

$$\begin{aligned}\|F(x) - F(x')\|_{d_Y} &= \sup_{y \neq y'} \frac{(f_x - f_{x'})(y) - (f_x - f_{x'})(y')}{d_Y(y, y')} \\ &= \sup_{y \neq y'} \frac{(f(x, y) - f(x', y)) - (f(x, y') - f(x', y'))}{d_Y(y, y')} \\ &\leq \|f\|_{d_X,d_Y} d_X(x, x').\end{aligned}$$

We now quote the result of the result from [13] that any function on a subspace of a metric tree to a Banach space can be extended to the entire tree without distorting the Lipschitz constant by more than a universal constant C_1. Let $\bar{F}$ denote the extension of F to the entire tree $\mathcal{T}_X$. Define $\bar{f}(I, y) = \bar{F}(I)(y)$ (this notation may be confusing; $\bar{F}(I)$ is a function on Y, and $\bar{F}(I)(y)$ denotes its value at y). Then the mixed Lipschitz constant of $\bar{f}$ is no more than $C_1\|f\|_{d_X,d_Y}$.

This gives us the desired extension of f to $\mathcal{T}_X \times Y$. Observe that this argument required only that X be a subspace of a metric tree; we did not exploit anything about the metric properties of Y. We can therefore repeat the same argument, taking Y in place of X and $\mathcal{T}_X$ in place of Y. This yields an extension $\bar{f}$ to all of $\mathcal{T}_X \times \mathcal{T}_Y$ at the loss of another factor C_1. Consequently, we have found the desired extension $\bar{f}$, with distortion no more than $C \equiv C_1^2$. □

We will now use the formula from Theorem 3 to derive a semi-norm equivalent to the dual semi-norm $\|T\|^*_{d_X,d_Y}$.

Theorem 4. *Let C be the same universal constant from Theorem 3. Then for any distribution T on $X \times Y$,*

$$\frac{1}{C}\|T\|^*_{d_X,d_Y} \leq \sum_{I \neq X} \sum_{J \neq Y} e^X_I e^Y_J |\langle \chi_I \chi_J, T \rangle| \leq \|T\|^*_{d_X,d_Y}. \tag{3}$$

Proof. Take any function f with $\|f\|_{d_X,d_Y} \leq 1$ and zero marginals (that is, $m_X f = m_Y f = 0$). By Theorem 3, we can write

$$f(x, y) = \sum_{I \neq X} \sum_{J \neq Y} a_{I \times J} \chi_I(x) \chi_J(y)$$

where $1 \geq \|f\|_{d_X,d_Y} \geq \frac{1}{C} \sup_{I \neq X, J \neq Y} |a_{I \times J}|/(e^X_I e^Y_J)$. Since the marginals of f are all zero, we can assume without loss of generality that the same is true for T when taking the inner product. Therefore, we have

$$|\langle f,T\rangle| = \left|\sum_{I\neq X,J\neq Y} a_{I\times J}\langle \chi_I\chi_J,T\rangle\right| \leq \sum_{I\neq X,J\neq Y} \frac{|a_{I\times J}|}{e_I^X e_J^Y} e_I^X e_J^Y |\langle \chi_I\chi_J,T\rangle|$$
$$\leq C \sum_{I\neq X,J\neq Y} e_I^X e_J^Y |\langle \chi_I\chi_J,T\rangle|$$

and taking the supremum over all f yields $\|T\|^*_{d_X,d_Y} \leq C\sum_{I\neq X,J\neq Y} e_I^X e_J^Y|\langle \chi_I\chi_J,T\rangle|$.

For the other inequality, define the function $\tilde{f}$ by

$$\tilde{f}(x,y) = \sum_{I\neq X,J\neq Y} e_I^X e_J^Y \operatorname{sgn}(\langle \chi_I\chi_J,T\rangle)\chi_I(x)\chi_J(y) + g(x) + h(y)$$

where the functions g and h are taken to ensure that $\tilde{f}$ has zero marginals. Theorem 3 shows that $\|\tilde{f}\|_d = 1$. Again, since $\tilde{f}$ has zero marginals, we can assume T does too when taking their inner product. Therefore,

$$\|T\|^*_{d_X,d_Y} \geq \langle \tilde{f},T\rangle = \sum_{I\neq X,J\neq Y} e_I^X e_J^Y \operatorname{sgn}(\langle \chi_I\chi_J,T\rangle)\langle \chi_I\chi_J,T\rangle$$
$$= \sum_{I\neq X,J\neq Y} e_I^X e_J^Y |\langle \chi_I\chi_J,T\rangle|$$

which completes the proof. □

Corollary 5. *Let C be the same universal constant from Theorems 3 and 4. Then for any T in the space dual to Lipschitz on $X \times Y$,*

$$\frac{1}{C}\|T\|^*_{\Lambda_{X,Y}} \leq \sum_{I\neq X}\sum_{J\neq Y} e_I^X e_J^Y |\langle \chi_I\chi_J,T\rangle| + \sum_{I\neq X} e_I^X\langle \chi_I, m_Y T\rangle$$
$$+ \sum_{J\neq Y} e_J^Y\langle \chi_J, m_X T\rangle + \|m_X m_Y T\|_1 \leq \|T\|^*_{\Lambda_{X,Y}}.$$

Unfortunately, we cannot take the constant C from Theorem 4 to be 1. However, numerical evidence suggests that the constant may not be too much bigger than 1. Figure 1 shows a scatter plot of $\|T\|^*_{d_X,d_Y}$ and the approximation from (3) for 1000 random vectors T on the product of two spaces with 8 points each. The trees on each space are binary trees with all edges equal to 1; that is, $e_I^X = e_J^Y = 1$. The largest ratio of the true norm over its approximation was about 1.203, and the minimum ratio about 1.044. Other experiments have given similar results.

Remark 3. As with the formula (1) from Theorem 4 for the dual norm to Lipschitz, the formula from (3) that is equivalent to $\|T\|^*_{d_X,d_Y}$ can be computed at cost proportional to the size of $X \times Y$. The argument is the same as for one tree. If X has M points and Y has N points, then the total number of products $I \times J \in \mathcal{T}_X \times \mathcal{T}_Y$ is of the order $\mathcal{O}(M \cdot N)$. To compute the integral of T over $I \times J$, we need only add its integral over

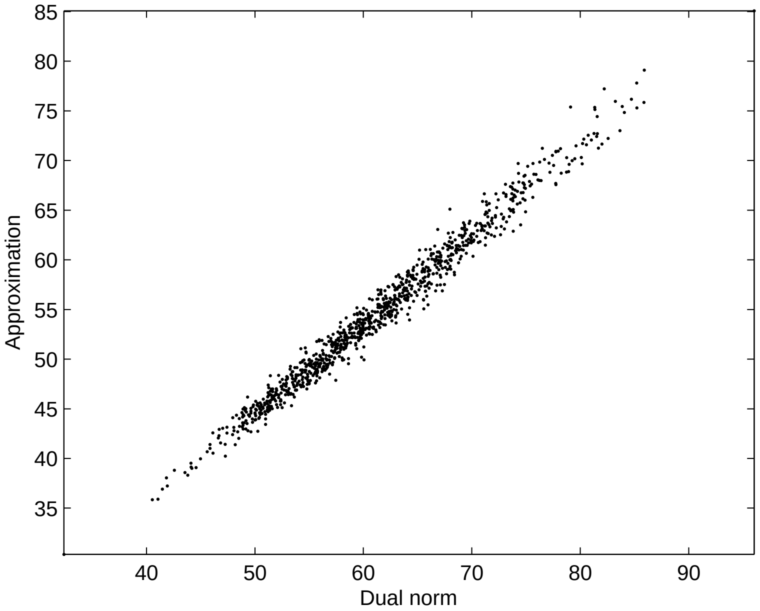

Figure 1: The dual norm and its approximation for 1000 random vectors

the folder $I' \times J'$, where I' ranges over the children of I, and J' over the children of J. Consequently, we can compute all the terms $\langle \chi_I \chi_J, T \rangle$ by starting with the $M \cdot N$ integrals of T over the singletons, and then recursively computing the integral of T over the product $I \times J$ by adding up its integral over the products $I' \times J'$, where I' is a child of I and J' a child of J. Each product of folders is only touched once, in the computation of the integral over the product of their parents; and so the total cost is linear in $M \cdot N$.

5 Lipschitz and mixed Lipschitz functions for trees with geometrically decaying folder weights

In Theorems 2 and 4, we derived simple formulas for the norms dual to the Lipschitz and mixed Lipschitz spaces. In all cases, the distortion guaranteed by these formulas does not depend on any features of the tree; the choice of edge weights can be arbitrary.

The characterizations we gave of Lipschitz and mixed Lipschitz functions themselves in Theorems 1 and 3, however, are not as directly useful, as they cannot be computed any more rapidly than the original definitions via difference quotients. In this section, we address this problem for the special class of tree metrics defined by folder weights $w(I)$, rather than edge weights e_I, as defined in Subsection 2.1.

We will assume geometric decay of the folder weights. More precisely, we assume that there is a constant $0 < A < 1$ such that for any folders $I \subsetneq J$,

$$w(I) \leq Aw(J). \tag{4}$$

Note that this family of trees includes the theory of k-hierarchically well-separated trees [1]. Note too that our assumptions are far less restrictive than those found in the papers [7, 8], in which the weights $w(I)$ are taken to be a power of the measure of I, and the measure is assumed to satisfy a two-sided decay condition, rather than the one-sided condition (4). These papers find norms equivalent to $\|f\|_d$ and $\|f\|_{d_X,d_Y}$ that use the coefficients of f in a special orthonormal basis. The formulas we give use the martingale difference operators in place of the Haar functions.

As in [7, 8], and unlike Theorems 2 and 4 for the dual norms, the constants of distortion are not universal, but rather depend on the decay constant A from (4).

Proposition 1. *Suppose f is any function on X, a set equipped with a tree $\mathcal{T}$. Suppose the distance on X is defined using folder weights $w(I)$ satisfying the decay condition* (4)*. Then we can approximate $\|f\|_d$ as follows:*

$$\frac{1-A}{2}\|f\|_d \leq \sup_I \frac{\|\Delta_I f\|_\infty}{w(I)} \leq \|f\|_d \tag{5}$$

where the supremum is over all non-singleton folders I.

Proof. Since $\operatorname{diam}(I) = w(I)$, the second inequality follows immediately from Corollary 3. For the other direction, recall that the distance $d(x,y)$ can be defined using the edge weights $e_I = \frac{1}{2}(w(I') - w(I))$, where I' is the parent of I. Since $w(I) \leq Aw(I')$, we have $e_I \geq \frac{1}{2}(1-A)w(I')$, and consequently from Corollary 3

$$\|f\|_d \leq \sup_{I \neq X} \frac{|\Delta_{I'} f(I)|}{e_I} \leq \frac{2}{1-A} \sup_I \frac{|\Delta_{I'} f(I)|}{w(I')} \leq \frac{2}{1-A} \sup_I \frac{\|\Delta_I f\|_\infty}{w(I)}.$$

□

Remark 4. As with the formula (1) from Theorem 4 for the dual norm to Lipschitz, the formula from (5) that is equivalent to $\|f\|_d$ can be computed at cost proportional to the size of X. In fact, to compute the middle term of (5), one needs to evaluate the martingale differences of f on every folder; to do this, it is sufficient to evaluate the average of f on every folder, and by Remark 2 all integrals, and hence averages, can be computed in linear time.

Theorem 5. *Suppose X and Y are two spaces equipped with trees $\mathcal{T}_X$ and $\mathcal{T}_Y$ with folder weights $w_X(I)$, $w_Y(J)$, respectively, each satisfying the decay condition* (4)*. Let d_X and d_Y be the metrics induced by these weights. Then for any function f on $X \times Y$, we can characterize its mixed Lipschitz semi-norm as follows:*

$$\frac{(1-A)^2}{4}\|f\|_{d_X,d_Y} \leq \sup_{I \in \mathcal{T}_X} \sup_{J \in \mathcal{T}_Y} \frac{\|\Delta_{X,I}\Delta_{Y,J} f\|_\infty}{w_X(I)w_Y(J)} \leq \|f\|_{d_X,d_Y} \tag{6}$$

where the supremums are over non-singleton folders only.

Proof. Fix any point $y \in Y$ and any folder $J \in \mathcal{T}_Y$, and consider the function

$$x \mapsto \frac{\Delta_{Y,J} f(x,y)}{w_Y(J)}.$$

Applying Proposition 1 to this function yields

$$\begin{aligned}\sup_{I \in \mathcal{T}_X} \sup_{x \in X} \frac{|\Delta_{X,I}\Delta_{Y,J} f(x,y)|}{w_X(I) w_Y(J)} &\leq \sup_{x \neq x'} \frac{\Delta_{Y,J} f(x,y) - \Delta_{Y,J} f(x',y)}{d_X(x,x') w_Y(J)} \\ &= \sup_{x \neq x'} \frac{\Delta_{Y,J}[f(x,\cdot) - f(x',\cdot)](y)}{d_X(x,x') w_Y(J)}.\end{aligned}$$

Temporarily fix two points $x \neq x'$. We apply Proposition 1 to the function

$$y \mapsto \frac{f(x,y) - f(x',y)}{d_X(x,x')}$$

to obtain the upper bound

$$\frac{\Delta_{Y,J}[f(x,\cdot) - f(x',\cdot)](y)}{d_X(x,x') w_Y(J)} \leq \sup_{y' \neq y''} \frac{f(x,y') - f(x',y') - f(x,y'') + f(x',y'')}{d_X(x,x') d_Y(y,y')}.$$

Taking the supremum over all J and y proves the inequality

$$\sup_{I \in \mathcal{T}_X} \sup_{J \in \mathcal{T}_Y} \frac{\|\Delta_{X,I}\Delta_{Y,J} f\|_\infty}{w_X(I) w_Y(J)} \leq \|f\|_{d_X, d_Y}.$$

To show the other direction, we apply the same method of reducing to Proposition 1, but going in the other direction. Fix any two points $y \neq y'$ in Y and apply Proposition 1 to the function

$$x \mapsto \frac{f(x,y) - f(x,y')}{d_Y(y,y')}.$$

This yields the inequality

$$\begin{aligned}\frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{d_X(x,x') d_Y(y,y')} &\leq \frac{2}{1-A} \sup_{I \in \mathcal{T}_X} \sup_{x \in X} \frac{|\Delta_{X,I}[f(x,y) - f(x,y')]|}{w_X(I) d_Y(y,y')} \\ &= \frac{2}{1-A} \sup_{I \in \mathcal{T}_X} \sup_{x \in X} \frac{|\Delta_{X,I} f(x,y) - \Delta_{X,I} f(x,y')|}{w_X(I) d_Y(y,y')}.\end{aligned}$$

Fixing any $x \in X$ and any $I \in \mathcal{T}_X$, we can apply Proposition 1 again to the function

$$y'' \mapsto \frac{\Delta_{X,I} f(x,y'')}{w_X(I)}$$

to get the inequality

$$\frac{|\Delta_{X,I} f(x,y) - \Delta_{X,I} f(x,y')|}{w_X(I) d_Y(y,y')} \leq \frac{2}{1-A} \sup_{J \in \mathcal{T}_Y} \sup_{y \in Y} \frac{|\Delta_{Y,J}\Delta_{X,I} f(x,y'')|}{w_X(I) w_Y(J)}.$$

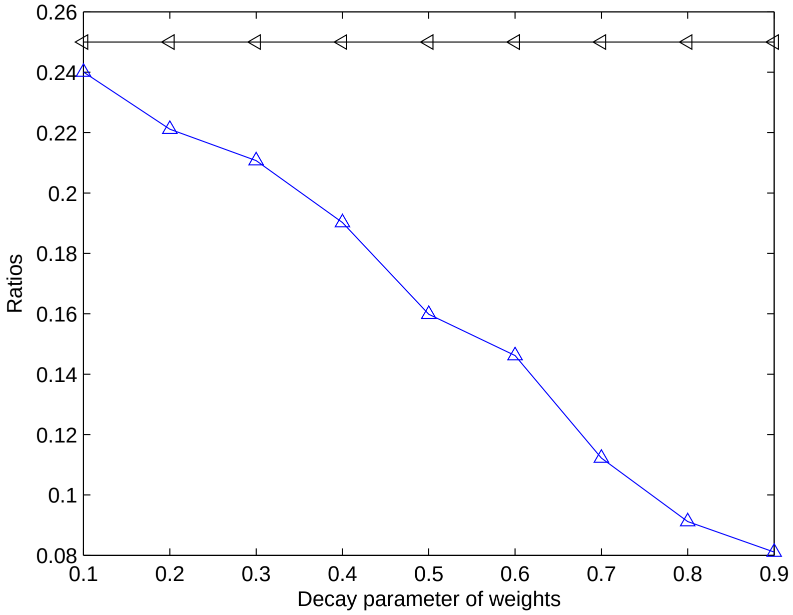

Figure 2: The maximum ratio (black, backward arrows) and minimum ratio (blue, forward arrows) of the approximate mixed Lipschitz norm to the truth for different A

It follows immediately that

$$\|f\|_{d_X,d_Y} \leq \frac{4}{(1-A)^2} \sup_{I\in\mathcal{T}_X} \sup_{J\in\mathcal{T}_Y} \frac{\|\Delta_{X,I}\Delta_{Y,J}f\|_\infty}{w_X(I)w_Y(J)}$$

and the result is proved. □

To illustrate the result of Theorem 5, we ran the following experiment. We took the product of two 16-point spaces with binary trees. For each choice of weight decay parameter $A = i/10, i = 1,\dots,9$ from (4), we compared the true value of $\|f\|_{d_X,d_Y}$ to the approximation from Theorem 5 for 200 random functions. Figure 2 shows the minimum and maximum ratios of the approximation divided by the true value, both as functions of A. As predicted by the theorem, the maximum stays more or less constant (its value is about .25, which is better than the worst-case value of 1 predicted by the theorem), while the minimum decays as A gets bigger. In Figure 3, we plot the ratios of the maximum ratio to the minimum ratio (the distortion) as a function of A. As expected, the distortion grows with A.

Remark 5. The formula from Theorem 5 that is equivalent to $\|f\|_{d_X,d_Y}$ can be computed at cost proportional to the size of $X \times Y$. To compute the middle term of (5), one needs to evaluate the double martingale differences of $\Delta_{X,I}\Delta_{Y,J}f$ on every pair of folders $I \times J$. To do this, it is sufficient to evaluate the integral of f on every pair of folders, as

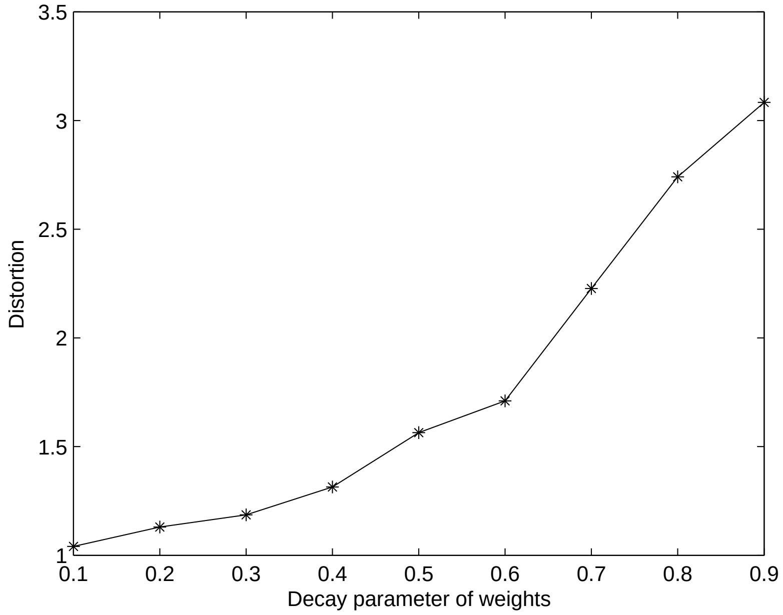

Figure 3: The distortion of approximation for the mixed Lipschitz norm for different A

$\Delta_{X,I}\Delta_{Y,J}f$ can be written as a linear combination of such integrals. We have already seen in Remark 3 that all these integrals can be computed in linear time.

6 Averaging Lipschitz norms and their duals over trees

Partition trees and tree metrics give rise to fast algorithms and simple formulas for the Lipschitz norm and their duals. However, in many applications from data analysis and machine learning, tree metrics are not refined enough to adequately capture the true geometry of the data. It is almost inevitable that any algorithm for constructing trees will separate points that are quite similar.

A standard way of overcoming this problem is to construct multiple trees on the same data set and combine the output from each tree. The hope is that the combination of many trees will "wash away" the artificial boundaries that any one tree will create. This idea has shown up in various places where trees appear. For instance, tree-based regression algorithms in statistics are augmented by the use of "random forests" [2], and in wavelet theory, Coifman and Donoho have proposed "spinning" the dyadic grid on $[0,1]$ to smooth out artifacts that would otherwise arise in tasks from signal processing such as filtering [4].

More relevant to the present work is the problem of approximating an arbitrary metric by the average of dominating tree metrics. More precisely, given an arbitrary finite metric space (X,d), the question is how to construct a random family of trees $\mathcal{T}$

on X so that the corresponding tree metrics $d_{\mathcal{T}}(x, y)$ satisfy

$$d(x, y) \leq d_{\mathcal{T}}(x, y) \tag{7}$$

for every tree $\mathcal{T}$, and in expectation we have the reverse inequality

$$\mathbb{E}_{\mathcal{T}} d_{\mathcal{T}}(x, y) \leq K d(x, y) \tag{8}$$

for some constant $K > 0$.

The problem dates back to Bartal [1]. In Fakcharoenphol et al [6], a construction is given of trees for which the expected distortion is $K = \mathcal{O}(\log |X|)$; this is the best one can obtain in general [14]. In [11], it is shown that the same trees constructed in [6] can be used to approximate the *snowflake metric* $d(x, y)^\alpha$, where $0 < \alpha < 1$, with a constant dependent only on the dimension of (X, d), and not on the number of points in X.

For the remainder of this Section, we will therefore assume that we have a metric space (X, d) that can be approximated by the average of dominating tree metrics. In other words, we assume we have a family of trees $\mathcal{T}$, each with its own metric $d_{\mathcal{T}}(x, y)$ that satisfies (7), and a distribution over these trees so that (8) holds as well.

Proposition 2. *For any function f on X,*

$$\sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}} \leq \|f\|_d \leq K \sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}}.$$

Proof. Since every tree metric $d_{\mathcal{T}}(x, y)$ dominates $d(x, y)$, we have

$$f(x) - f(y) \leq \|f\|_d d(x, y) \leq \|f\|_d d_{\mathcal{T}}(x, y)$$

which implies that $\|f\|_{d_{\mathcal{T}}} \leq \|f\|_d$ for all trees $\mathcal{T}$; consequently,

$$\sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}} \leq \|f\|_d.$$

For the other inequality, observe that for each tree $\mathcal{T}$, we have

$$f(x) - f(y) \leq \|f\|_{d_{\mathcal{T}}} d_{\mathcal{T}}(x, y) \leq \left(\sup_{\mathcal{T}'} \|f\|_{d_{\mathcal{T}'}} \right) d_{\mathcal{T}}(x, y).$$

Taking expectations of both sides gives

$$f(x) - f(y) \leq \left(\sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}} \right) \left(\mathbb{E}_{\mathcal{T}} d_{\mathcal{T}}(x, y) \right) \leq K \sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}} d(x, y)$$

implying that $\|f\|_d \leq K \sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}}$ and completing the proof. □

Proposition 3. *For any T in the space dual to Lipschitz functions on X,*

$$\frac{1}{K} \mathbb{E}_{\mathcal{T}} \|T\|_{d_{\mathcal{T}}}^* \leq \|T\|_d^* \leq \inf_{\mathcal{T}} \|T\|_{d_{\mathcal{T}}}^*$$

Proof. Since $\|f\|_{d_\mathcal{T}} \leq \|f\|_d$, we have

$$\|T\|_d^* = \sup_{\|f\|_d \leq 1} \langle f, T\rangle \leq \sup_{\|f\|_{d_\mathcal{T}} \leq 1} \langle f, T\rangle = \|T\|_{d_\mathcal{T}}^*$$

which yields the inequality $\|T\|_d^* \leq \inf_\mathcal{T} \|T\|_{d_\mathcal{T}}^*$.

For the other direction, fix any $\epsilon > 0$, and for each $\mathcal{T}$, let $f_\mathcal{T}$ be defined so that $\|f\|_{d_\mathcal{T}} \leq 1$ and

$$\|T\|_{d_\mathcal{T}}^* = \langle f_\mathcal{T}, T\rangle + \epsilon.$$

Now, since $f_\mathcal{T}(x) - f_\mathcal{T}(y) \leq d_\mathcal{T}(x, y)$, taking expectations gives

$$\mathbb{E}_\mathcal{T} f_\mathcal{T}(x) - \mathbb{E}_\mathcal{T} f_\mathcal{T}(y) \leq \mathbb{E}_\mathcal{T} d_\mathcal{T}(x, y) \leq K d(x, y)$$

or in other words $\|\mathbb{E}_\mathcal{T} f_\mathcal{T}\|_d \leq K$. Consequently, we have

$$\mathbb{E}_\mathcal{T} \|T\|_{d_\mathcal{T}}^* = \mathbb{E}_\mathcal{T} \langle f_\mathcal{T}, T\rangle + \epsilon = \langle \mathbb{E}_\mathcal{T} f_\mathcal{T}, T\rangle + \epsilon \leq K \sup_{\|f\|_d \leq 1} \langle f, T\rangle + \epsilon = K\|T\|_d^* + \epsilon.$$

Since ϵ is arbitrary, the result follows. □

Proposition 3 can also be deduced trivially from Charikar's paper [3], since the seminorm $\|T\|_\rho^*$, when T is the difference of two probability measures, is equal to the Earth Mover's Distance between these two measures with respect to the ground distance $\rho(x, y)$. However, the proof we have just given, which appears to be new, generalizes to the setting of mixed Lipschitz functions and their duals. We turn to this now.

For the next two results, we assume that we have two metric spaces (X, d_X) and (Y, d_Y), each with a family of dominating tree metrics, denoted $\mathcal{T}_X$ and $\mathcal{T}_Y$, respectively, that approximate d_X and d_Y in the sense of (7) and (8). We assume too that the trees on X and Y are constructed independently.

We first show that we can approximate $\|f\|_{d_X, d_Y}$ by the maximum of $\|f\|_{d_{\mathcal{T}_X}, d_{\mathcal{T}_Y}}$ over all pairs of dominating trees $(\mathcal{T}_X, \mathcal{T}_Y)$. The proof follows from Proposition 2.

Proposition 4. *For any function f on $X \times Y$, we have*

$$\sup_{\mathcal{T}_X, \mathcal{T}_Y} \|f\|_{d_{\mathcal{T}_X}, d_{\mathcal{T}_Y}} \leq \|f\|_{d_X, d_Y} \leq K^2 \sup_{\mathcal{T}_X, \mathcal{T}_Y} \|f\|_{d_{\mathcal{T}_X}, d_{\mathcal{T}_Y}}.$$

Proof. For any $y \neq y'$, let $g_{y,y'}(x) = f(x, y) - f(x, y')$, and for any $x \neq x'$, let $h_{x,x'}(y) =$

$f(x,y) - f(x',y)$. Using Proposition 2, we then have

$$\begin{aligned}
\|f\|_{d_X,d_Y} &= \sup_{x\neq x', y\neq y'} \frac{f(x,y)-f(x,y')-f(x',y)+f(x',y')}{d_X(x,x')d_Y(y,y')} \\
&= \sup_{x\neq x'} \frac{1}{d_X(x,x')} \sup_{y\neq y'} \frac{h_{x,x'}(y)-h_{x,x'}(y')}{d_Y(y,y')} \\
&\leq K \sup_{x\neq x'} \frac{1}{d_X(x,x')} \sup_{\mathcal{T}_Y} \sup_{y\neq y'} \frac{h_{x,x'}(y)-h_{x,x'}(y')}{d_{\mathcal{T}_Y}(y,y')} \\
&= K \sup_{\mathcal{T}_Y} \sup_{y\neq y'} \frac{1}{d_{\mathcal{T}_Y}(y,y')} \sup_{x\neq x'} \frac{g_{y,y'}(x)-g_{y,y'}(x')}{d_X(x,x')} \\
&\leq K^2 \sup_{\mathcal{T}_Y} \sup_{y\neq y'} \frac{1}{d_{\mathcal{T}_Y}(y,y')} \sup_{\mathcal{T}_X} \sup_{x\neq x'} \frac{g_{y,y'}(x)-g_{y,y'}(x')}{d_{\mathcal{T}_X}(x,x')} \\
&= K^2 \sup_{\mathcal{T}_X,\mathcal{T}_Y} \|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}.
\end{aligned}$$

The other inequality is proved similarly. □

Proposition 5. *For any L^1 measure T on $X \times Y$, we have*

$$\frac{1}{K^2}\mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y}\|T\|^*_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}} \leq \|T\|^*_{d_X,d_Y} \leq \inf_{\mathcal{T}_X,\mathcal{T}_Y} \|T\|^*_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}.$$

Proof. We essentially repeat the one-dimensional proof of Proposition 3. Since $\|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}} \leq \|f\|_{d_X,d_Y}$, it follows that $\|T\|^*_{d_X,d_Y} \leq \|T\|^*_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}$, and consequently

$$\|T\|^*_{d_X,d_Y} \leq \inf_{\mathcal{T}_X,\mathcal{T}_Y} \|T\|^*_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}$$

For the other inequality, fix any $\epsilon > 0$. For any pair of trees $(\mathcal{T}_X, \mathcal{T}_Y)$, we can find a function $f_{\mathcal{T}_X,\mathcal{T}_Y}$ with $m_X f_{\mathcal{T}_X,\mathcal{T}_Y} = 0$, $m_Y f_{\mathcal{T}_X,\mathcal{T}_Y} = 0$, and $\|f_{\mathcal{T}_X,\mathcal{T}_Y}\|_{d_X,d_Y} \leq 1$, such that

$$\|T\|^*_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}} = \langle f_{\mathcal{T}_X,\mathcal{T}_Y}, T\rangle + \epsilon.$$

Then for any $x, x' \in X$ and $y, y' \in Y$, we have

$$f_{\mathcal{T}_X,\mathcal{T}_Y}(x,y) - f_{\mathcal{T}_X,\mathcal{T}_Y}(x,y') - f_{\mathcal{T}_X,\mathcal{T}_Y}(x',y) + f_{\mathcal{T}_X,\mathcal{T}_Y}(x',y') \leq d_{\mathcal{T}_X}(x,x')d_{\mathcal{T}_Y}(y,y').$$

Taking expectations of each side and using the fact that $\mathbb{E}_{\mathcal{T}_X} d_{\mathcal{T}_X}(x,x') \leq K d_X(x,x')$ and $\mathbb{E}_{\mathcal{T}_Y} d_{\mathcal{T}_Y}(y,y') \leq K d_Y(y,y')$, we can easily see that

$$\|\mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y} f_{\mathcal{T}_X,\mathcal{T}_Y}\|_{d_X,d_Y} \leq K^2.$$

Consequently,

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y}\|T\|^*_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}} &= \mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y}\langle f_{\mathcal{T}_X,\mathcal{T}_Y}, T\rangle + \epsilon \\
&= \langle \mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y} f_{\mathcal{T}_X,\mathcal{T}_Y}, T\rangle + \epsilon \\
&\leq K^2 \|T\|^*_{d_X,d_Y} + \epsilon.
\end{aligned}$$

Since ϵ is arbitrary, we are done.

□

References

[1] Bartal, Y. (1996, October). Probabilistic approximation of metric spaces and its algorithmic applications. In Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on (pp. 184-193). IEEE.

[2] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[3] Charikar, M. S. (2002, May). Similarity estimation techniques from rounding algorithms. In Proceedings of the Thirty-fourth Annual ACM Symposium on Theory of Computing (pp. 380-388). ACM.

[4] Coifman, R. R., & Donoho, D. L. (1995). Translation-invariant de-noising (pp. 125-150). Springer New York.

[5] Donoho, D. L., & Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. Journal of the American Statistical Association, 90(432), 1200-1224.

[6] Fakcharoenphol, J., Rao, S., and Talwar, K. (2003, June). A tight bound on approximating arbitrary metrics by tree metrics. In Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing (pp. 448-455). ACM.

[7] Gavish, M., and Coifman, R. R. (2012). Sampling, denoising and compression of matrices by coherent matrix organization. Applied and Computational Harmonic Analysis, 33(3), 354-369.

[8] Gavish, M., Nadler, B., and Coifman, R. R. (2010). Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In Proceedings of the 27th International Conference on Machine Learning (ICML-10) (pp. 367-374).

[9] Gu, C., and Taibleson, M. (1992). Besov spaces on non-homogeneous martingales (pp. 69-84). Springer US.

[10] Heinonen, J. (2001). *Lectures on Analysis on Metric Spaces.* Springer.

[11] Leeb, W. (2014). A note on approximating snowflake metrics by trees. Technical Report YALEU/DCS/TR-1501, Yale University.

[12] McShane, E. J. (1934). Extension of range of functions. Bulletin of the American Mathematical Society, 40(12), 837-842.

[13] Matoušek, J. (1990). Extension of Lipschitz mappings on metric trees. Commentationes Mathematicae Universitatis Carolinae, 31(1), 99-104.

[14] Rabinovich, Y., and Raz, R. (1998). Lower bounds on the distortion of embedding finite metric spaces in graphs. Discrete and Computational Geometry, 19(1), 79-94.

[15] Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 40(2), 99-121.

[16] Korostelev, A. P., & Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction.* Springer.

[17] Cédric Villani. (2003). *Topics in Optimal Transportation* (No. 58). American Mathematical Soc.