Appl. Comput. Harmon. Anal. • • • (• • • •) • • •–• • •

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

The mixed Lipschitz space and its dual for tree metrics

William Leeb [a,b,*]

[a] *Department of Mathematics, Yale University, New Haven, CT 06511, United States*
[b] *PACM, Princeton University, Princeton, NJ 08544, United States*

ARTICLE INFO

Article history:
Received 25 January 2015
Received in revised form 5 November 2015
Accepted 23 June 2016
Available online xxxx
Communicated by Dominique Picard

MSC:
68W01
05C05
26A16
54E35
46M05

Keywords:
Tree metric
Lipschitz
Mixed Lipschitz
Dual space
Earth Mover's Distance
EMD
Martingale difference

ABSTRACT

This paper develops a theory of harmonic analysis on spaces with tree metrics, extending previous work in this direction by Gavish, Nadler and Coifman (2010) [30] and Gavish and Coifman (2011, 2012) [28,29]. We show how a natural system of martingales and martingale differences induced by a partition tree leads to simple and effective characterizations of the Lipschitz norm and its dual for functions on a single tree metric space. The restrictions we place on the tree metrics are far more general than those considered in previous work. As the dual norm is equal to the Earth Mover's Distance (EMD) between two probability distributions, we recover a simple formula for EMD with respect to tree distances presented by Charikar (2002) [36].

We also consider the situation where an arbitrary metric is approximated by the average of a family of dominating tree metrics. We show that the Lipschitz norm and its dual for the tree metrics can be combined to yield an approximation to the corresponding norms for the underlying metric.

The main contributions of this paper, however, are the generalizations of the aforementioned results to the setting of the product of two or more tree metric spaces. For functions on a product space, the notion of regularity we consider is not the Lipschitz condition, but rather the mixed Lipschitz condition that controls the size of a function's mixed difference quotient. This condition is extremely natural for datasets that can be described as a product of metric spaces, such as word-document databases. We develop effective formulas for norms equivalent to the mixed Lipschitz norm and its dual, and extend our results on combining pairs of trees.

1. Introduction

This paper develops a theory of harmonic analysis on spaces endowed with tree metrics, which are distances that arise naturally throughout pure and applied mathematics. We are concerned primarily with spaces of Lipschitz and mixed Lipschitz functions and their duals, and in particular, simple and computable characterizations of the norms on these spaces.

* Correspondence to: PACM, Princeton University, Princeton, NJ 08544, United States.
E-mail address: wleeb@math.princeton.edu.

http://dx.doi.org/10.1016/j.acha.2016.06.008
1063-5203/

1.1. The Lipschitz space, the dual space and Earth Mover's Distance

Given a metric space (X, d), a natural way of measuring the variation of a function f defined on X is its Lipschitz norm, defined by

$$\sup_{x \neq y} \frac{f(x) - f(y)}{d(x, y)}. \tag{1}$$

Oftentimes, it is convenient to define the Lipschitz norm to be the sum or maximum of (1) and $\|f\|_\infty$; we will consider both versions in this paper, though for the purposes of this introductory section we will restrict our attention to (1). If f is a differentiable function on $\mathbb{R}$, the Lipschitz norm (1) is equal to $\|f'\|_\infty$, the supremum of f's derivative. Expression (1), however, is defined for non-differentiable functions and makes sense in the abstract setting of any metric space.

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 (where the underlying metric space is $\mathbb{R}$ and the distance is defined as $d(x, y) = |x-y|^\alpha$ for some $0 < \alpha < 1$) or a more general regularity class [1,2]. 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.

In the Euclidean setting of classical analysis, where we consider the space $\mathbb{R}$ equipped with the distance $d(x, y) = |x - y|^\alpha$ for some $0 < \alpha < 1$, (1) (which is then referred to as the Hölder norm of f) can be shown to be equivalent in size to a number of other expressions that look at differences of averages of f over different scales. For example, if we take a sufficiently nice wavelet basis $\{\psi_{j,k}\}$ of $\mathbb{R}^n$ (where $j \in \mathbb{Z}$ indexes the dyadic scale 2^j and $k \in \mathbb{Z}$ the location), then the expression

$$\sup_{j,k} 2^{-j(\alpha+1/2)} |\langle f, \psi_{j,k} \rangle| \tag{2}$$

is equivalent in size to (1), which is to say that the ratio of the two quantities is bounded above and below by finite constants not depending on f [3]. The wavelet coefficients $\langle f, \psi_{j,k} \rangle$ can be thought of as measuring f's variation across scales.

In a discrete setting, where we only have f sampled on a grid of k points in $\mathbb{R}^n$, computing the Lipschitz norm (1) directly would require $O(k^2)$ operations, as all pairs of points need to be accounted for. However, using the fast wavelet transform (see, for instance, [4]), the expression (2) can be computed with only $O(k)$ operations. In addition to their computational tractability, the simple characterization of the Hölder norm of a function via its wavelet coefficients gives rise to efficient statistical procedures for signal recovery in the nonparametric setting; see [5].

Also of interest is the space dual to Lipschitz. Given any normed space $(\mathcal{X}, \|\cdot\|)$, one defines its dual space as the collection of linear functionals on $\mathcal{X}$, equipped with the norm

$$\|T\|_* = \sup_{f \in \mathcal{X} : \|f\| \leq 1} \langle f, T \rangle. \tag{3}$$

When $\mathcal{X}$ is the space of Lipschitz functions over a metric space (X, d), the dual norm (3) has another interpretation, described by the Kantorovich–Rubinstein Theorem [6,7]. If μ and ν are two probability measures over X, we define their Earth Mover's Distance (EMD) to be

$$\mathrm{EMD}(\mu, \nu) = \inf_{\pi} \int_{\Omega \times \Omega} d(x, y) d\pi(x, y). \tag{4}$$

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

where the infimum is over measures π on $X \times X$ satisfying the following equality-of-marginals condition with respect to μ and ν:

$$\begin{aligned} \pi(X, E) &= \mu(E) \\ \pi(E, X) &= \nu(E) \end{aligned} \tag{5}$$

for all (measurable) sets $E \subset X$.

The Earth Mover's Distance between μ and ν has the following interpretation. We view each measure π satisfying the equality-of-marginals condition (5) with respect to μ and ν as a transport between the measures μ and ν; that is, for any two measurable sets $A, B \subset X$, $\pi(A, B)$ is interpreted as the amount of mass moved from set A to set B. The equality-of-marginals condition (5) guarantees that the transport rearranges the mass distribution described by ν to end up with the distribution described by μ. If the distance $d(x, y)$ is the cost-per-mass of moving mass from location x to location y, then $\mathrm{EMD}(\mu, \nu)$ is the minimal cost over all transports; in other words, it is the cheapest way of rearranging mass distributed like ν to get mass distributed like μ.

The Kantorovich–Rubinstein Theorem states that $\mathrm{EMD}(\mu, \nu)$ is equal to $\|\mu - \nu\|_*$, the norm of $\mu - \nu$ in the space dual to Lipschitz functions. Due to the way it exploits the geometry of the metric space on which the two probability distributions are defined, EMD has many desirable properties that make it a natural choice of metric for many problems in machine learning [8–11].

Part of the reason why the dual distance (in the form of EMD) is a popular metric in these applications is that it provides a robust way of comparing two measures on a dataset that is insensitive to perturbations, a desirable property for many tasks. This property is formalized in the following result, whose proof can be found in [12]:

Theorem 1. *Suppose F is an L^1 function on a measure space X equipped with metric $d(x, y)$, and $h : X \to X$ is a perturbation of the identity, by which we mean h is an absolutely continuous bijection and $d(x, h(x)) \leq \epsilon$ for all $x \in X$, where $0 < \epsilon < 1$. Define*

$$G(x) = F(h(x))\frac{dh}{dx}(x) \tag{6}$$

where $\frac{dh}{dx}$ denotes the Radon–Nikodym derivative of h. Then $\|F - G\|_* \leq \epsilon \|F\|_1$.

1.2. The mixed Lipschitz space and its dual

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 product of several metric spaces. Such datasets arise naturally in a variety of applications. For example, in the theory of transposable arrays [13,14], both the rows and columns of a dataset are studied. Similarly, methods of co-clustering [15,16] search for a clustering of both the row and column sets of a data matrix, and so fits into the same framework.

For ease of exposition, we restrict attention to the product of two spaces, say (X, d_X) and (Y, d_Y). The notion of regularity we consider for a function f defined on $X \times Y$ is the mixed Lipschitz condition, which requires f to have bounded mixed difference quotients; that is,

$$\sup_{x \neq x', y \neq y'} \frac{f(x, y) - f(x, y') - f(x', y) + f(x', y')}{d_X(x, x') d_Y(y, y')} \tag{7}$$

must be finite. We give a more formal definition of the mixed Lipschitz space in Section 4. The expression (7) for smooth functions on $\mathbb{R}^2$ is equivalent in size to $\|\partial^2_{x,y} f\|_\infty$.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

If f is defined on, say, $[0,1]\times[0,1]$ and has bounded mixed difference quotients, many desirable properties follow. For instance, such functions can be reconstructed to precision ϵ using only $O((1/\epsilon)\log(1/\epsilon))$ samples; these samples form what is known as a "sparse grid" [17,18]. This compares favorably with the $\Omega(1/\epsilon^2)$ points that would be needed if f were only known to be Lipschitz. Similarly, only $O((1/\epsilon)\log(1/\epsilon))$ coefficients from a suitable wavelet basis are needed to reconstruct such a function f to precision ϵ [19]. Relatedly, statistical estimators of such an f from noisy samples achieve higher minimax rates than estimators of functions that are merely assumed to be Lipschitz [20,21].

However, the expression (7) is unnatural in many applications in the Euclidean setting, as it depends heavily on the choice of x- and y-axes. In particular, it is not rotationally invariant; for example, consider a function of the form $f(x,y)=g(x)h(y)$, for smooth g and h. Then

$$\partial^2_{xy} f(x,y) = g'(x)h'(y) \tag{8}$$

whereas

$$\partial^2_{xy} f\left(\frac{x+y}{\sqrt{2}},\frac{x-y}{\sqrt{2}}\right) = \frac{1}{2}g''\left(\frac{x+y}{\sqrt{2}}\right)h\left(\frac{x-y}{\sqrt{2}}\right) - \frac{1}{2}g\left(\frac{x+y}{\sqrt{2}}\right)h''\left(\frac{x-y}{\sqrt{2}}\right) \tag{9}$$

and by constructing functions g and h with enormous first derivatives but small second and zeroth derivatives, one easily shows that the size of the quantity (7) depends heavily on the coordinate system, a severe limitation in many physical settings where the axes are chosen arbitrarily. Therefore, despite the nice properties of the space of functions with bounded mixed differences, their applicability is limited in settings where the choice of axes is not meaningful.

By contrast, in many data-analysis problems the axes are not arbitrarily chosen but rather intrinsic to the problem itself; consider, for example, the word/document axes of a word-document database [22], or the time/frequency axes of a spectrogram [20,21]. It is reasonable, therefore, to look at norms, like the mixed Lipschitz norm, that depend on the choice of axes; in fact, such norms make the most sense in this context.

We also consider the space dual to mixed Lipschitz functions; the norm on this space is defined by (3), where $\mathcal{X}$ is now the space of mixed Lipschitz functions. As with Earth Mover's Distance, discussed in Subsection 1.1, the dual norm to the mixed Lipschitz space provides a robust distance between measures on the product of metric spaces. In fact, as defined in Section 4 every mixed Lipschitz function is also Lipschitz, and so Theorem 1 trivially applies in the mixed Lipschitz setting with two-dimensional perturbations of the identity. We are currently exploring the use of this norm as a distance for comparing two-dimensional databases as well as its use in certain statistical applications, and plan to report on these results in future publications.

1.3. Tree metrics

The discussion in Subsections 1.1 and 1.2 is applicable to any metric space. In this paper we restrict our attention to a specific class of metrics, known as *tree metrics*, which will be rigorously defined in Subsection 2.1. A tree metric is derived from a collection of nested subsets of X, referred to as *folders*, that form a tree graph with the elements in X as its leaves. The tree metric is induced from numerical weights placed on the edges connecting each folder to its parent.

Due to their simple structure, tree metrics are amenable to computation and show up throughout pure and applied mathematics. Tree metrics yield fast algorithms for metric tasks from computer science, such as nearest neighbor searches, the k-server problem, distributed paging, and the vehicle routing problem, among others [23,24]. Trees also exhibit good metric embedding properties into l_p spaces [25,26], and are also used in the construction of embeddings of more general metric spaces [27].

The launching point for the present work can be found in [28–30], where a certain restricted family of tree metrics is studied with an eye toward their application in machine learning problems. In these papers, the

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

diameter of a tree's folder is given by its volume raised to some power. Furthermore, the volumes of folders must decay at a very controlled rate as the folders shrink in size. Under these restrictions, an elegant theory of harmonic analysis is developed which parallels the classical theory and permits us to adapt classical signal processing techniques – including sparse grids and wavelet decompositions – to more general data analysis problems. A cornerstone of the theory are the simple characterizations of the Lipschitz and mixed Lipschitz spaces by a wavelet-like basis of functions (the Haar-like basis).

The current paper extends this analysis in several ways. First, we give not only characterizations of the Lipschitz and mixed Lipschitz spaces, but also of their duals, which, as we indicated in Subsections 1.1 and 1.2, provide robust distances for comparing distributions on the data. Second, by measuring the local variation of f via *martingale differences* (to be defined in Subsection 2.2) rather than Haar coefficients, we are able to remove many of the restrictions found in [28–30]. For the characterizations of the Lipschitz and mixed Lipschitz spaces, we do not require folder diameters to be volume-based, nor do we require a lower bound on the decay rate of the folder weights, but only an upper bound; that is, the trees must be hierarchically well-separated in the sense of [23,24]. For the dual norms, our results apply to arbitrary tree metrics, and the constants of equivalence we derive are universal (they do not depend on the tree at all).

In addition to the increased generality of our results, we maintain the computational efficiency of the equivalent norms we derive. In particular, all norms we derive can be computed at cost proportional to the number of points in the space.

1.4. Combining the output from multiple trees

As we indicated in Subsection 1.3, tree metrics have many properties that make them especially easy to work with. However, a limitation of tree metrics encountered in applications is that any single tree usually arises from breaking a continuous geometry into a discrete one, and artificial discontinuities can result.

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 and similar structures appear. For instance, tree-based regression algorithms in statistics are augmented by the use of "random forests" [31], 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 [32].

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

$$d(x, y) \leq d_{\mathcal{T}}(x, y) \tag{10}$$

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

$$\mathbb{E}_{\mathcal{T}} d_{\mathcal{T}}(x, y) \leq K d(x, y) \tag{11}$$

for some constant $K \geq 1$. See [33,23,24].

In Section 6, we will show how to combine the regularity norms (Lipschitz and mixed Lipschitz) and their duals for a family of tree metrics to approximate the corresponding norms for a metric approximated by these tree metrics in the sense of (10) and (11).

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

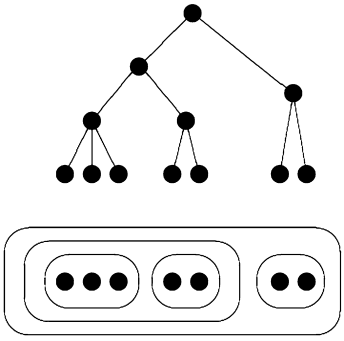

Fig. 1. Bottom: The folders (as ovals) of a tree on a 7 point set. Top: The same tree, with folders represented as nodes in a graph; the original points are the leaves on the bottom level.

2. Preliminaries

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, which we will refer to as *folders*, such that for any two folders 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. We will denote by $\mathrm{sub}(I)$ the set of all children of a folder I.

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, if X is the set of leaves 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. This correspondence between the set-theoretic view and the graph-theoretic view of partition trees is illustrated by Fig. 1.

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 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.

Remark 1. We note that the restriction that the spaces X and Y be finite sets is not critical for most if not all of the results in this paper to hold. We make this assumption because the application areas we have in mind do not require infinite sets, and because doing so allows us to circumvent measure-theoretic technicalities encountered in the study of infinite dimensional function spaces that would only distract from the results we present. Note that whenever we prove that two norms of functions defined on X or $X \times Y$ are equivalent, the constants of equivalence do not depend on the number of points in X or Y; consequently, extending our results to infinite spaces is fairly straightforward.

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.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

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 $w(I) \leq w(J)$; and for any $x \in X$, $w(\{x\}) = 0$. 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 straightforward:

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 [23,24,34]; 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 [28–30]. We will have more to say about this subject in Section 6. 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 will let $|S|$ denote the measure of a subset $S \subset X$. All integrals encountered in this paper are defined with respect to this same measure on X.

We define the martingale and martingale difference operators with respect to the measure on X, 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

$$\Delta_I f(x) = \sum_{J \in \mathrm{sub}(I)} m_J f(x) - m_I f(x).$$

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

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

$$\begin{aligned}\int_X \Delta_I f(x)dx &= \int_X \sum_{J\in \mathrm{sub}(I)} m_J f(x)dx - \int_X m_I f(x)dx \\ &= \sum_{J\in \mathrm{sub}(I)} |J| m_J f(J) - |I| m_I f(I) \\ &= \sum_{J\in \mathrm{sub}(I)} \int_J f(x)dx - \int_I f(x)dx = 0 \\ &= \int_I f(x)dx - \int_I f(x)dx = 0. \quad \square\end{aligned}$$

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. $\square$

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. $\square$

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,

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

$$m_I^2 f(x) = \left(\frac{1}{|I|}\int_I m_I f(y)dy\right)\chi_I(x) = \left(\frac{1}{|I|}\int_I m_I f(I)dy\right)\chi_I(x) = m_I f(I)\chi_I(x)$$
$$= \left(\frac{1}{|I|}\int_I f(y)dy\right)\chi_I(x) = m_I f(x). \quad \square$$

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 will adopt the language of probability theory and refer to $m_X f$ and $m_Y f$ as the *marginals* of f.

We also define

$$\Delta_{X,I} f(x,y) = \sum_{I' \in \mathrm{sub}(I)} m_{X,I'} f(x,y) - m_{X,I} f(x,y)$$

and

$$\Delta_{Y,J} f(x,y) = \sum_{J' \in \mathrm{sub}(J)} m_{Y,J'} f(x,y) - m_{Y,J} f(x,y).$$

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 positive edge weights e_I. Define the L^∞ variation of a function f on X with respect to the metric $d(x,y)$ by

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

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

We define the Lipschitz norm of f to be

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

Note that $\|m_X f\|_\infty$ is nothing more than $|m_X f(X)|$, as $m_X f$ is a constant function.

The norm dual to $\|\cdot\|_d$ is given by

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

and the norm dual to $\|\cdot\|_\Lambda$ is given by

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

We have the following simple lemma:

Lemma 6. *For every T,*

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

Note that $\|m_X T\|_1$ is nothing more than $|X||m_X T(X)|$, since $m_X T$ is a constant.

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

$$\|f\|_\Lambda = \max\{\|f_1\|_d, \|m_X f_2\|_\infty\}.$$

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. □

In this section we derive simple and effectively computable 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 2. *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*

$$f(x) = \sum_{I \in \mathcal{T}} a_I \chi_I(x).$$

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

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

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

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,

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

where we have used Lemma 1 in the last equality. This shows that $\|f\|_d \leq C_f$.

For the other direction, let $\bar{f}(I) = \sup_{x\in X}(f(x) - \|f\|_d d(x,I))$, where $d(x,I)$ denotes the distance between the point x and the folder I in the tree (the sum of the weights on the edges connecting them). It is shown in [35] that $\bar{f}$ extends f (that is, $\bar{f}(x) = f(x)$ whenever $x \in X$; this is obvious, since $d(y,x)$ is minimized when $y = x$) and that $\bar{f}$ 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. This follows immediately from the fact that $\bar{f}$ extends f, and the inequality

$$\bar{f}(I) - \bar{f}(J) \leq \sup_{x\in X}(f(x) - \|f\|_d d(x,I) - f(x) + \|f\|_d d(x,J)) \leq \|f\|_d d(I,J).$$

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

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

$$\sup_I \frac{\|\Delta_I f\|_\infty}{\operatorname{diam}(I)} \leq \|f\|_d \leq \sup_{I\neq X} \frac{|\Delta_{I'} f(I)|}{e_I} \tag{12}$$

where 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

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

$$\leq \frac{1}{|I'|}\frac{1}{|I|}\int_{I'}\int_{I} \|f\|_d \operatorname{diam}(I')dxdy = \|f\|_d \operatorname{diam}(I').$$

Dividing each side by $\operatorname{diam}(I')$ and taking the supremum over all I gives the leftmost inequality of (12).

For the other side, we make use of Theorem 2. For each folder $I \neq X$, let I' denote its parent, and define $a_I = \Delta_{I'} f(I)$. We have the telescopic sum

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

and consequently Theorem 2 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 2 to derive a very simple formula for $\|T\|_d^*$.

Theorem 3. *For every T,*

$$\|T\|_d^* = \sum_{I \neq X} e_I |\langle \chi_I, T\rangle|. \tag{13}$$

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

Proof. Take any function f with $\|f\|_d \leq 1$ and $m_X f = 0$. By Theorem 2, we can write

$$f = \sum_{I \in \mathcal{T}} a_I \chi_I$$

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 integral 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. Theorem 2 shows that $\|\tilde{f}\|_d = 1$. Again, since $\tilde{f}$ has mean zero, we can assume T also has total integral 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. □

Combining Theorem 3 and Lemma 6, we get:

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

Corollary 4. *For every T, 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 2. Theorem 3 can be easily derived from the formula for Earth Mover's Distance given in [36], 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 [6,7], discussed in the introduction. However, Theorem 2 is new, as is the proof we give of Theorem 3 that uses the formula for $\|f\|_d$ from Theorem 2. Furthermore, this proof will generalize to the setting of product spaces, as we will see in Section 4. In Section 5 we will make further use of Theorem 2 to derive equivalent formulas for the Lipschitz and mixed Lipschitz norms on a special class of trees.

The formula (13) for $\|T\|_d^*$ from Theorem 3 can be computed in cost proportional to the size of X. In fact, we have the following algorithm:

Algorithm 1 *(Computation of $\|T\|_d^*$).*

I. Compute all terms $\langle \chi_I, T\rangle$:
   1. Evaluate T at all singleton folders $\{x\}$.
   2. For every non-singleton folder $I \in \mathcal{T}$ whose child integrals have already been computed, recursively compute $\langle \chi_I, T\rangle$ using the formula

$$\langle \chi_I, T\rangle = \sum_{I' \in \mathrm{sub}(I)} \langle \chi_{I'}, T\rangle \tag{14}$$

II. Multiply each $\langle \chi_I, T\rangle$ by e_I and add them together.

If we let M denote the number of points in X, the number of folders in $\mathcal{T}$ cannot exceed $2M-1$. Step I1 requires M operations, while Step I2 for a folder I whose child integrals have already been computed requires $|\,\mathrm{sub}(I)|$ operations. Since

$$\sum_{\text{non-singleton } I} |\,\mathrm{sub}(I)| = O(|\mathcal{T}|) = O(M) \tag{15}$$

the total operation count for Step I is $O(M)$. Step II clearly requires only $O(M)$ additional operations, bringing the total operation count to $O(M)$.

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, which we define presently. 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

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

Note that we can add to f any function of the form $g(x) + h(y)$ without changing the value of $\|f\|_{d_X,d_Y}$.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

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 ($\|m_X f\|_{d_Y}$ and $\|m_Y f\|_{d_X}$) is sensible. Note too that $\|m_X m_Y f\|_\infty$ is nothing more than the unique value of $|m_X m_Y f|$, as this is a constant function.

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

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

Note that we can add to T any function of the form $g(x) + h(y)$ without changing $\|T\|^*_{d_X,d_Y}$.

The dual norm of T acting on 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 T on $X \times Y$,*

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

Note that $\|m_X m_Y T\|_1$ is simply the unique value of $|X||Y||m_X m_Y T|$, since this is a constant function.

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

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

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, specifically Theorem 3 and Algorithm 1, $\|m_Y T\|^*_{d_X} + \|m_X T\|^*_{d_Y} + \|m_X m_Y T\|_1$ 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 our proof of Theorem 3, which employed the formula from Theorem 2 for the Lipschitz norm of mean zero functions, our formula for approximating $\|T\|^*_{d_X,d_Y}$ is derived from a similar characterization of mixed Lipschitz functions with zero marginals given by the following theorem:

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

Theorem 4. *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

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

Proof. Take any $\{a_{I\times J}\} \in \mathcal{A}_f$, and any $x, x' \in X$, $y, y' \in Y$. Recall that $\mathcal{S}_{x,x'}$ denotes the set of folders in $\mathcal{T}_X$ that contain exactly one of x or x'; then for all $I \in \mathcal{S}_{x,x'}$, $|\chi_I(x) - \chi_I(x')| = 1$. Similar remarks apply to $\mathcal{S}_{y,y'}$.

We have:

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

where we have made use of Lemma 1 in the last equality. This proves that $\|f\|_{d_X,d_Y} \le \inf_{\{a_{I\times J}\}\in\mathcal{A}_f} \sup_{I\neq X,J\neq Y} \frac{|a_{I\times J}|}{e_I^X e_J^Y}$.

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

$$|\bar{f}(I,J) - \bar{f}(I,J') - \bar{f}(I',J) + \bar{f}(I',J')| \le C\|f\|_{d_X,d_Y} e_I^X e_J^Y \tag{16}$$

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

If we had such an extension, we would be finished since we could expand f as

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

Taking $\tilde{a}_{I\times J} = \bar{f}(I,J) - \bar{f}(I,J') - \bar{f}(I',J) + \bar{f}(I',J')$, (16) shows $|\tilde{a}_{I\times J}| \le C\|f\|_{d_X,d_Y} e_I^X e_J^Y$; consequently,

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

which is the desired result.

We now show how to prove the existence of such an extension $\bar{f}$. First, by fixing $y_0 \in Y$ and replacing f with $f(x,y) - f(x,y_0)$ we can assume without loss of generality that $f(x,y_0) = 0$ for all x. We now interpret the function f as a map not from $X \times Y$ into $\mathbb{R}$, but rather from X into the space $\mathrm{Lip}_0(Y)$ of Lipschitz

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

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 from [37] 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 $\bar{f}(I,y) = \bar{F}(I)(y)$ (in case the reader finds this notation confusing, note that $\bar{F}(I)$ 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 4 to derive a semi-norm equivalent to the dual semi-norm $\|T\|^*_{d_X,d_Y}$.

Theorem 5. *Let C be the same universal constant from Theorem 4. Then for any T on $X \times Y$,*

$$\frac{1}{C}\|T\|^*_{d_X,d_Y} \leq \sum_{I \neq X} \sum_{J \neq Y} e_I^X e_J^Y |\langle \chi_I \chi_J, T\rangle| \leq \|T\|^*_{d_X,d_Y}. \tag{17}$$

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 4, we can write

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

where $1 \geq \|f\|_{d_X,d_Y} \geq \frac{1}{C} \sup_{I \neq X, J \neq Y} |a_{I\times J}|/(e_I^X e_J^Y)$. Since the marginals of f are all zero, replacing T by $T - m_Y T - m_X T + m_X m_Y T$ does not change the value of $\|T\|^*_{d_X,d_Y}$, and ensures that $\langle \chi_I \chi_J, T\rangle = 0$ whenever $I = X$ or $J = Y$. Therefore, we have

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

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

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

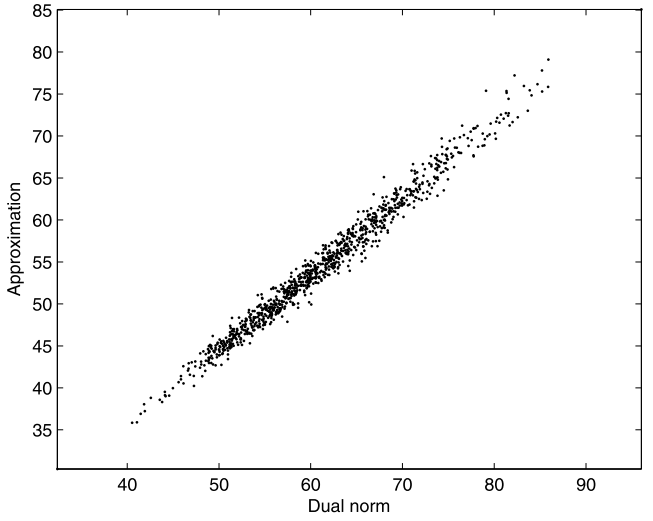

Fig. 2. The dual norm and its approximation for 1000 random vectors.

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 4 shows that $\|\tilde{f}\|_{d_X,d_Y} \leq 1$. Again, since $\tilde{f}$ has zero marginals, we can assume T does too when taking their inner product. Therefore,

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

which completes the proof. □

Combining Theorem 5 and Lemma 7, we get:

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

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

Unfortunately, we cannot take the constant C from Theorem 5 to be 1. However, numerical evidence suggests that the constant may not be too much bigger than 1. Fig. 2 shows a scatter plot of $\|T\|^*_{d_X,d_Y}$ and the approximation from (17) 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.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

As with the formula (13) from Theorem 5 for the norm dual to Lipschitz, the formula from (17) that is equivalent to $\|T\|^*_{d_X,d_Y}$ can be computed at cost proportional to the size of $X \times Y$. In fact, we have the following algorithm:

Algorithm 2 *(Computation of approximation to $\|T\|^*_{d_X,d_Y}$).*

I. Compute all terms $\langle \chi_I \chi_J, T\rangle$:
   1. Evaluate T at all singleton product folders $\{x\} \times \{y\}$.
   2. For every non-singleton product $I \times J$ ($I \in \mathcal{T}_X, J \in \mathcal{T}_Y$) whose child integrals have already been computed, recursively compute $\langle \chi_I \chi_J, T\rangle$ using the formula

$$\langle \chi_I \chi_J, T\rangle = \sum_{I' \in \mathrm{sub}(I)} \sum_{J' \in \mathrm{sub}(J)} \langle \chi_{I'} \chi_{J'}, T\rangle \tag{18}$$

II. Multiply each $\langle \chi_I \chi_J, T\rangle$ by $e_I e_J$ and add them together.

If we let M denote the number of points in X, and N the number of points in Y, the number of product folders in $\mathcal{T}_X \times \mathcal{T}_Y$ cannot exceed $(2M-1)(2N-1)$. Step I1 requires MN operations, while Step I2 for a product folder $I \times J$ whose child integrals have already been computed requires $|\mathrm{sub}(I)||\mathrm{sub}(J)|$ operations. Since

$$\sum_{I \in \mathcal{T}_X} \sum_{J \in \mathcal{T}_Y} |\mathrm{sub}(I)||\mathrm{sub}(J)| = O(|\mathcal{T}_X||\mathcal{T}_Y|) = O(MN) \tag{19}$$

the total operation count for Step I is $O(MN)$ (note that in the summations (18) and (19), for singletons we define $\mathrm{sub}(\{x\}) = \{x\}$ and $\mathrm{sub}(\{y\}) = \{y\}$). Step II clearly requires only $O(MN)$ additional operations, bringing the total operation count to $O(MN)$.

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

In Theorems 3 and 5, 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. Furthermore, the formulas can be computed in linear time.

The characterizations we gave of Lipschitz and mixed Lipschitz functions themselves in Theorems 2 and 4, 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 $I \subsetneq J$,

$$w(I) \leq A w(J). \tag{20}$$

Note that this family of trees includes k-hierarchically well-separated trees [23]. Note too that our assumptions are still far less restrictive than those found in [28–30], 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 (20). 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 of Haar-like functions. The equivalent norms we give use the martingale difference operators in place of the Haar functions, which allows us to prove results for a greater variety of trees.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

As in [28–30], and unlike Theorems 3 and 5 for the dual norms, the constants of distortion are not universal, but rather depend on the decay constant A from (20).

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 (20). Then we can approximate $\|f\|_d$ as follows:*

$$\frac{1-A}{2}\|f\|_d \le \sup_I \frac{\|\Delta_I f\|_\infty}{w(I)} \le \|f\|_d \tag{21}$$

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 from Lemma 2 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 $w(I) \le Aw(I')$, we have $e_I \ge \frac{1}{2}(1-A)w(I')$, and consequently from Corollary 3

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

The approximation to $\|f\|_d$ given by Proposition 1 can be computed at cost proportional to the size of X, as described in the following algorithm.

Algorithm 3 *(Computation of approximation to $\|f\|_d$).*

I. Compute all terms $m_I f(I)$:
   1. Evaluate f at all singleton folders $\{x\}$, and divide by the measure of $\{x\}$.
   2. For every non-singleton folder $I \in \mathcal{T}$ whose child averages have already been computed, recursively compute $m_I f(I)$ using the formula

$$m_I f(I) = \frac{1}{|I|} \sum_{I' \in \operatorname{sub}(I)} |I'| m_{I'} f(I') \tag{22}$$

II. Compute all terms $\|\Delta_I f(I)\|_\infty$ by the formula

$$\|\Delta_I f(I)\|_\infty = \max_{I' \in \operatorname{sub}(I)} |m_{I'} f(I') - m_I f(I)|. \tag{23}$$

III. Compute the maximum of $\|\Delta_I f(I)\|_\infty / w(I)$.

If we let M denote the number of points in X, the number of folders in $\mathcal{T}$ cannot exceed $2M-1$. Step I1 requires $O(M)$ operations, while Step I2 for a folder I whose child averages have already been computed requires $O(|\operatorname{sub}(I)|)$ operations. Since

$$\sum_{\text{non-singleton } I} |\operatorname{sub}(I)| = O(|\mathcal{T}|) = O(M) \tag{24}$$

the total operation count for Step I is $O(M)$. Since Step II requires only $O(|\operatorname{sub}(I)|)$ additional operations for each folder I, its total cost over all folders is also $O(M)$. Finally, Step III clearly only requires an additional $O(M)$ operations, bringing the total operation count to $O(M)$.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

Theorem 6. *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 (20). 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:*

$$\frac{(1-A)^2}{4}\|f\|_{d_X,d_Y} \leq \sup_{I\in\mathcal{T}_X}\sup_{J\in\mathcal{T}_Y}\frac{\|\Delta_{X,I}\Delta_{Y,J}f\|_\infty}{w_X(I)w_Y(J)} \leq \|f\|_{d_X,d_Y} \tag{25}$$

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

$$\begin{aligned}\sup_{I\in\mathcal{T}_X}\sup_{x\in X}\frac{|\Delta_{X,I}\Delta_{Y,J}f(x,y)|}{w_X(I)w_Y(J)} &\leq \sup_{x\neq x'}\frac{\Delta_{Y,J}f(x,y)-\Delta_{Y,J}f(x',y)}{d_X(x,x')w_Y(J)}\\ &= \sup_{x\neq x'}\frac{\Delta_{Y,J}[f(x,\cdot)-f(x',\cdot)](y)}{d_X(x,x')w_Y(J)}.\end{aligned} \tag{26}$$

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

$$\frac{\Delta_{Y,J}[f(x,\cdot)-f(x',\cdot)](y)}{d_X(x,x')w_Y(J)} \leq \sup_{y'\neq y''}\frac{f(x,y')-f(x',y')-f(x,y'')+f(x',y'')}{d_X(x,x')d_Y(y,y')}. \tag{27}$$

Combining (26) and (27) and 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

$$\begin{aligned}\frac{f(x,y)-f(x,y')-f(x',y)+f(x',y')}{d_X(x,x')d_Y(y,y')} &\leq \frac{2}{1-A}\sup_{I\in\mathcal{T}_X}\sup_{x\in X}\frac{|\Delta_{X,I}[f(x,y)-f(x,y')]|}{w_X(I)d_Y(y,y')}\\ &= \frac{2}{1-A}\sup_{I\in\mathcal{T}_X}\sup_{x\in X}\frac{|\Delta_{X,I}f(x,y)-\Delta_{X,I}f(x,y')|}{w_X(I)d_Y(y,y')}.\end{aligned} \tag{28}$$

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

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

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

to get the inequality

$$\frac{|\Delta_{X,I} f(x,y) - \Delta_{X,I} f(x,y')|}{w_X(I) d_Y(y,y')} \leq \frac{2}{1-A} \sup_{J \in \mathcal{T}_Y} \sup_{y \in Y} \frac{|\Delta_{Y,J} \Delta_{X,I} f(x,y'')|}{w_X(I) w_Y(J)}. \tag{29}$$

From (28) and (29), 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. □

The approximation to $\|f\|_{d_X,d_Y}$ given by Theorem 6 can be computed at cost proportional to the size of $X \times Y$, as described in the following algorithm.

Algorithm 4 *(Computation of approximation to $\|f\|_{d_X,d_Y}$).*

I. Compute all terms $m_I f(I)$:
   1. Evaluate f at all singleton product folders $\{x\} \times \{y\}$, and divide by the measure of $\{x\} \times \{y\}$.
   2. For every product folder $I \times J$ ($I \in \mathcal{T}_X, J \in \mathcal{T}_Y$) whose child averages have already been computed, recursively compute $m_{X,I} m_{Y,J} f(I,J)$ using the formula

$$m_{X,I} m_{Y,J} f(I,J) = \frac{1}{|I||J|} \sum_{I' \in \mathrm{sub}(I)} \sum_{J' \in \mathrm{sub}(J)} |I'||J'| m_{X,I'} m_{Y,J'} f(I',J') \tag{30}$$

II. Compute all terms $\|\Delta_{X,I} \Delta_{Y,J} f\|_\infty$ by the formula

$$\begin{aligned} \|\Delta_{X,I} \Delta_{Y,J} f\|_\infty = \max_{I' \in \mathrm{sub}(I), J' \in \mathrm{sub}(J)} & |m_{X,I'} m_{Y,J'} f(I',J') - m_{X,I} m_{Y,J'} f(I,J') \\ & - m_{X,I'} m_{Y,J} f(I',J) + m_{X,I} m_{Y,J} f(I,J)|. \end{aligned} \tag{31}$$

III. Compute the maximum of $\|\Delta_{X,I} \Delta_{Y,J} f\|_\infty / (w_X(I) w_Y(J))$.

If we let M denote the number of points in X and N the number of points in Y, the number of product folders in $\mathcal{T}_X \times \mathcal{T}_Y$ cannot exceed $(2M-1)(2N-1)$. Step I1 requires $O(MN)$ operations, while Step I2 for a product folder $I \times J$ whose child averages have already been computed requires $O(|\mathrm{sub}(I)||\mathrm{sub}(J)|)$ operations. Since

$$\sum_{I \in \mathcal{T}_X} \sum_{J \in \mathcal{T}_Y} |\mathrm{sub}(I)||\mathrm{sub}(J)| = O(|\mathcal{T}_X||\mathcal{T}_Y|) = O(MN) \tag{32}$$

the total operation count for Step I is $O(MN)$ (note that in the summations (30) and (32), for singletons we define $\mathrm{sub}(\{x\}) = \{x\}$ and $\mathrm{sub}(\{y\}) = \{y\}$). Since Step II requires only $O(|\mathrm{sub}(I)||\mathrm{sub}(J)|)$ additional operations for each product folder $I \times J$, its total cost over all folders is also $O(MN)$. Finally, Step III clearly only requires an additional $O(MN)$ operations, bringing the total operation count to $O(MN)$.

To illustrate the result of Theorem 6, we ran the following experiment. We took the product of two 16-point spaces with binary trees. For each choice of weight decay parameter $A = i/10$, $i = 1, \ldots, 9$ from (20), we compared the true value of $\|f\|_{d_X,d_Y}$ to the approximation from Theorem 6 for 200 random functions. Fig. 3 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

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

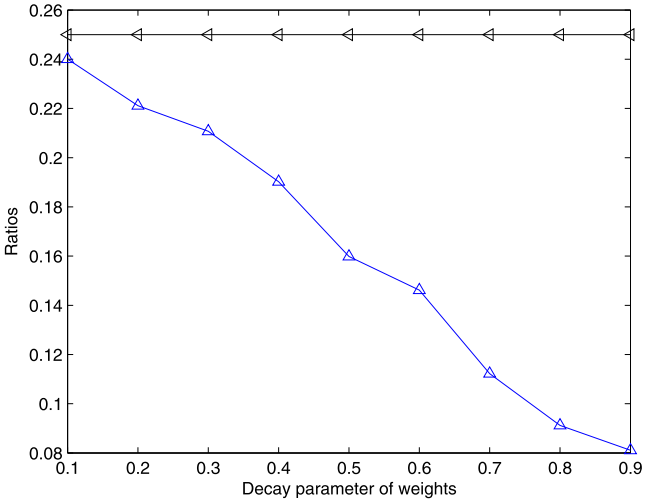

Fig. 3. The maximum ratio (black, backward arrows) and minimum ratio (blue, forward arrows) of the approximate mixed Lipschitz norm to the truth for different A. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

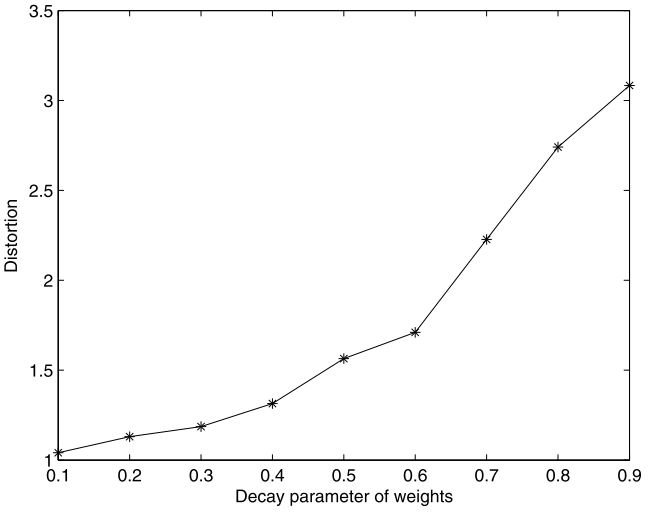

Fig. 4. The distortion of approximation for the mixed Lipschitz norm for different A.

decays as A gets bigger. In Fig. 4, 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.

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 applications tree metrics are often not the natural distance for a given problem; rather they are used only as a proxy for some other intrinsic metric. For example, in [28–30] tree metrics arise by clustering points according to their proximity with respect to a certain diffusion distance on the data, which is the L^2 distance between probability distributions on the data centered at each point [38]; higher levels of the tree group together points that are connected at larger scales in the diffusion process. However, the resulting tree metric only serves as an approximation to the true diffusion distance, which is the real quantity of interest.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

In computer science applications, similar issues arise all the time as well. As discussed in the introduction, many metric tasks in computer science such as nearest neighbor searches, the k-server problem, distributed paging, the vehicle routing problem, and many more, are quite simple when the metric is a tree metric. If a family of tree metrics approximates some intrinsic metric in an appropriate sense then the solution for the tree metrics can be combined to yield an approximate solution for the intrinsic metric [23,24].

As stated in the introduction, the formal problem of tree approximations of arbitrary metrics can be stated as follows: given an arbitrary finite metric space (X, d), we seek a random family of trees $\mathcal{T}$ on X so that the corresponding tree metrics $d_{\mathcal{T}}(x, y)$ satisfy

$$d(x, y) \leq d_{\mathcal{T}}(x, y) \tag{33}$$

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

$$\mathbb{E}_{\mathcal{T}} d_{\mathcal{T}}(x, y) \leq K d(x, y) \tag{34}$$

for some constant $K \geq 1$.

Bartal's paper [23] describes such an explicit distribution over trees, where the constant K is of size $\mathcal{O}(\log^2 M)$, if X contains M points (this result was sharpened to $K = O(\log M \log\log M)$ in [24]). The paper of Fakcharoenphol, Rao, and Talwar [34] describes a randomized construction of partition trees with $K = O(\log M)$. As there are metric spaces of size M for which no family of trees can achieve a distortion smaller than $\Omega(\log M)$ [23,34], this result is optimal in the general case. In a forthcoming paper [39] we show that the trees from [34] 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) (a quantity measuring the growth rate of metric balls) rather than the number of points in X.

We will adopt this formalism of tree metric approximations here. For the remainder of this section, we assume that we have a metric space (X, d) that is approximated by the average of dominating tree metrics. In other words, we have a family of trees $\mathcal{T}$, each with its own metric $d_{\mathcal{T}}(x, y)$ satisfying (33), and a distribution over these trees so that (34) holds as well.

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

$$\sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}} \leq \|f\|_d \leq K \mathbb{E}\left[\frac{1}{\|f\|_{d_{\mathcal{T}}}}\right]^{-1}.$$

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, for each tree $\mathcal{T}$, we have $f(x) - f(y) \leq \|f\|_{d_{\mathcal{T}}} d_{\mathcal{T}}(x, y)$. Dividing by $\|f\|_{d_{\mathcal{T}}}$ and taking expectations yields

$$(f(x) - f(y)) \mathbb{E}_{\mathcal{T}}\left[\frac{1}{\|f\|_{d_{\mathcal{T}}}}\right] \leq \mathbb{E}_{\mathcal{T}} d_{\mathcal{T}}(x, y) \leq K d(x, y)$$

which gives the desired result. □

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

Note that by Jensen's inequality, we also have:

$$\mathbb{E}_{\mathcal{T}}\left[\frac{1}{\|f\|_{d_{\mathcal{T}}}}\right]^{-1} \leq \mathbb{E}_{\mathcal{T}}[\|f\|_{d_{\mathcal{T}}}] \leq \sup_{\mathcal{T}} \|f\|_{d_{\mathcal{T}}}.$$

Proposition 3. *For any T 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 a mean zero function such that $\|f\|_{d_{\mathcal{T}}} \leq 1$ and

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

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

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

Since ϵ is arbitrary, the result follows. □

Proposition 3 can also be deduced from Charikar's paper [36], since the semi-norm $\|T\|^*_\rho$, when T is the difference of two probability distributions, is equal to the Earth Mover's Distance between these two distributions 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, which we turn to 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 trees, denoted $\mathcal{T}_X$ and $\mathcal{T}_Y$, respectively, and tree metrics $d_{\mathcal{T}_X}$, $d_{\mathcal{T}_Y}$, that approximate d_X and d_Y in the sense of (33) and (34). 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 values of $\|f\|_{d_{\mathcal{T}_X}, d_{\mathcal{T}_Y}}$ over all pairs of dominating trees $(\mathcal{T}_X, \mathcal{T}_Y)$. The proof is nearly identical to the proof of Proposition 2.

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

$$\sup_{\mathcal{T}_X, \mathcal{T}_Y} \|f\|_{d_{\mathcal{T}_X}, d_{\mathcal{T}_Y}} \leq \|f\|_{d_X, d_Y} \leq K^2 \mathbb{E}_{\mathcal{T}_X, \mathcal{T}_Y}\left[\frac{1}{\|f\|_{d_{\mathcal{T}_X}, d_{\mathcal{T}_Y}}}\right]^{-1}.$$

Proof. Take any $x \neq x'$ and $y \neq y'$. Then for any pair of trees $(\mathcal{T}_X, \mathcal{T}_Y)$, we have

$$\begin{aligned} f(x, y) - f(x, y') - f(x', y) + f(x', y') &\leq \|f\|_{d_X, d_Y} d_X(x, x') d_Y(y, y') \\ &\leq \|f\|_{d_X, d_Y} d_{\mathcal{T}_X}(x, x') d_{\mathcal{T}_Y}(y, y') \end{aligned}$$

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

which implies that

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

For the other inequality, for each pair of trees $(\mathcal{T}_X,\mathcal{T}_Y)$, we have

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

Dividing by $\|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}$, taking expectations, and using the independence of the trees yields

$$\begin{aligned}(f(x,y)-f(x,y')-f(x',y)+f(x',y'))\mathbb{E}_{\mathcal{T}}\left[\frac{1}{\|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}}\right] &\leq \mathbb{E}_{\mathcal{T}_X} d_{\mathcal{T}_X}(x,x')\mathbb{E}_{\mathcal{T}_Y} d_{\mathcal{T}_Y}(y,y') \\ &\leq K^2 d_X(x,x')d_Y(y,y')\end{aligned}$$

which gives the desired result. □

Again, note that by Jensen's inequality, we also have:

$$\mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y}\left[\frac{1}{\|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}}\right]^{-1} \leq \mathbb{E}_{\mathcal{T}_X,\mathcal{T}_Y}[\|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}] \leq \sup_{\mathcal{T}_X,\mathcal{T}_Y} \|f\|_{d_{\mathcal{T}_X},d_{\mathcal{T}_Y}}.$$

Proposition 5. *For any 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

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

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,

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

Since ϵ is arbitrary, we are done. □

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

7. Conclusion

In this paper we have developed a theory of harmonic analysis on tree metric spaces, extending the work of [28–30] to a more general collection of tree metrics. In particular, we have introduced computationally efficient approximations to the Lipschitz norm and its dual on a single tree metric space, and the mixed Lipschitz norm and its dual on a product of tree metric spaces. As discussed in the introduction, the mixed Lipschitz norm is a natural measure of regularity for functions on a product of spaces with meaningful axes. In [29], many desirable properties of mixed Lipschitz functions from the Euclidean setting are pushed through to a restricted class of tree metrics; an interesting topic for future research would be investigating how to adapt these same properties to the more general trees we have considered in this paper.

The norm dual to mixed Lipschitz provides a distance between measures on a product of spaces that is analogous to Earth Mover's Distance (the dual norm to Lipschitz functions on a single space). As such, it exhibits many of the same properties as EMD, such as an insensitivity to perturbations of the data. However, since every mixed Lipschitz function is Lipschitz (with respect to the sum of the distances on the spaces), but the converse is not true, we expect that the norm dual to mixed Lipschitz should exhibit even stronger properties. This line of research, as well as the use of the dual norm as a distance in applications with real data, are currently being pursued. Finally, a question for further research is to obtain information on the size of the universal constant C from Theorem 5 that bounds the ratio of the norm dual to mixed Lipschitz and its approximation.

Acknowledgments

I thank my advisor Ronald Coifman for his guidance and enthusiastic support throughout this work. I also thank the anonymous reviewers for their insightful comments, which improved the exposition enormously and sharpened Propositions 2 and 4.

References

[1] D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc. 90 (432) (1995) 1200–1224.
[2] A.P. Korostelev, A.B. Tsybakov, Minimax Theory of Image Reconstruction, Springer, 1993.
[3] Y. Meyer, Wavelets and Operators, Cambridge University Press, 1992.
[4] S. Mallat, A Wavelet Tour of Signal Processing, 2nd edition, Academic Press, 1999.
[5] D.L. Donoho, I.M. Johnstone, Neo-classical minimax problems, thresholding and adaptive function estimation, Bernoulli 2 (1) (1996) 39–62.
[6] R.M. Dudley, Real Analysis and Probability, Cambridge University Press, 2002.
[7] C. Villani, Topics in Optimal Transportation, American Mathematical Society, 2003.
[8] S. Marinai, B. Miotti, G. Soda, Using Earth mover's distance in the bag-of-visual-words model for mathematical symbol retrieval, in: 2011 International Conference on Document Analysis and Recognition, 2011, pp. 1309–1313.
[9] Y. Rubner, C. Tomasi, L.J. Guibas, The Earth mover's distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2) (2000) 99–121.
[10] R. Sandler, M. Lindenbaum, Nonnegative matrix factorization with Earth mover's distance metric for image analysis, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1590–1602.
[11] X. Wan, A novel document similarity measure based on Earth mover's distance, Inform. Sci. 177 (18) (2007) 3718–3730.
[12] W. Leeb, R. Coifman, Hölder–Lipschitz norms and their duals on spaces with semigroups, with applications to Earth Mover's Distance, J. Fourier Anal. Appl. (2016), http://dx.doi.org/10.1007/s00041-015-9439-5.
[13] G.I. Allen, R. Tibshirani, Transposable regularized covariance models with an application to missing data imputation, Ann. Appl. Stat. 4 (2) (2010) 764–790.
[14] G.I. Allen, R. Tibshirani, Inference with transposable data: modelling the effects of row and column correlations, J. R. Stat. Soc. Ser. B. Stat. Methodol. 74 (4) (2012) 721–743.
[15] J.A. Hartigan, Direct clustering of a data matrix, J. Amer. Statist. Assoc. 67 (337) (1972) 123–129.
[16] I.S. Dhillon, S. Mallela, D.S. Modha, Information-theoretic co-clustering, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2003, pp. 89–98.
[17] S.A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Dokl. Akad. Nauk SSSR 4 (1963) 240–243.
[18] T. Gerstner, M. Griebel, Numerical integration using sparse grids, Numer. Algorithms 18 (1998) 209–232.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008

[19] J.-A. Strömberg, Computation with wavelets in higher dimensions, Doc. Math. 3 (1998) 523–532.
[20] M.H. Neumann, Multivariate wavelet thresholding in anisotropic function spaces, Statist. Sinica 10 (2) (2000) 399–432.
[21] M.H. Neumann, R. von Sachs, Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra, Ann. Statist. 25 (1) (1997) 38–76.
[22] G. Bisson, F. Hussain, χ-sim: a new similarity measure for the co-clustering task, in: Seventh International Conference on Machine Learning and Applications, IEEE, 2008, pp. 211–217.
[23] Y. Bartal, Probabilistic approximation of metric spaces and its algorithmic applications, in: 37th Annual Symposium on Foundations of Computer Science, IEEE, 1996, pp. 184–193.
[24] Y. Bartal, On approximating arbitrary metrics by tree metrics, in: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, ACM Press, 1998, pp. 161–168.
[25] J. Bourgain, The metrical interpretation of superreflexivity in Banach spaces, Israel J. Math. 56 (2) (1986) 222–230.
[26] J. Matoušek, On embedding trees into uniformly convex Banach spaces, Israel J. Math. 114 (1) (1999) 221–237.
[27] I. Abraham, Y. Bartal, O. Neiman, Advances in metric embedding theory, in: Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing, 2006, pp. 271–286.
[28] M. Gavish, R. Coifman, Harmonic analysis of digital databases, in: J. Cohen, A.I. Zayed (Eds.), Wavelets and Multiscale Analysis, Birkhäuser, 2011, pp. 161–197.
[29] M. Gavish, R.R. Coifman, Sampling, denoising and compression of matrices by coherent matrix organization, Appl. Comput. Harmon. Anal. 33 (3) (2012) 354–369.
[30] M. Gavish, B. Nadler, R.R. Coifman, 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, 2010, pp. 367–374.
[31] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[32] R.R. Coifman, D.L. Donoho, Translation-invariant de-noising, in: A. Antoniadis, G. Oppenheim (Eds.), Wavelets and Statistics, Springer, 1995, pp. 125–150.
[33] N. Alon, R.M. Karp, D. Peleg, D. West, A graph-theoretic game and its application to the k-server problem, SIAM J. Comput. 24 (1) (1995) 78–100.
[34] J. Fakcharoenphol, S. Rao, K. Talwar, A tight bound on approximating arbitrary metrics by tree metrics, in: Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing, ACM, 2003, pp. 448–455.
[35] E.J. McShane, Extension of range of functions, Bull. Amer. Math. Soc. 40 (12) (1934) 837–842.
[36] M.S. Charikar, Similarity estimation techniques from rounding algorithms, in: Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, 2002, pp. 380–388.
[37] J. Matoušek, Extension of Lipschitz mappings on metric trees, Comment. Math. Univ. Carolin. 31 (1) (1990) 99–104.
[38] R.R. Coifman, S. Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (1) (2006) 5–30.
[39] W. Leeb, Approximating snowflake metrics by trees. Submitted for publication.

Please cite this article in press as: W. Leeb, The mixed Lipschitz space and its dual for tree metrics, Appl. Comput. Harmon. Anal. (2016), http://dx.doi.org/10.1016/j.acha.2016.06.008