We introduce a family of bounded, multiscale distances on any space equipped with an operator semigroup. In many examples, these distances are equivalent to a snowflake of an intrinsic distance on the space. Under weak regularity assumptions on the kernels defining the semigroup, we derive simple characterizations of the Lipschitz norm and its dual with respect to these distances. As the dual norm between the difference of two probability measures is the Earth Mover's Distance (EMD) between these measures, our characterizations give simple formulas for a metric equivalent to EMD. We extend these results to the mixed Lipschitz norm and its dual on the product of spaces, each of which is equipped with its own semigroup. Additionally, we derive an approximation theorem for mixed Lipschitz functions in this setting.

Lipschitz norms, mixed Lipschitz norms and their duals on spaces with semigroups, with applications to Earth Mover's Distance

William Leeb and Ronald Coifman

Technical Report YALEU/DCS/TR-1499
September 30, 2014

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

Approved for public release: distribution is unlimited.
Keywords: *Lipschitz, mixed Lipschitz, Besov space, dual space, tensor product, Earth Mover's Distance, EMD, diffusion distance, semigroup, snowflake metric, heat kernel*

1 Introduction

Operator semigroups are ubiquitous objects in pure and applied mathematics. It is well-known that many basic function spaces, such as the Besov spaces on $\mathbb{R}^n$, can be characterized using, for example, the heat kernel [2, 23]. Recent work has generalized these results to Besov spaces in more general domains; see, for instance, the papers [1] and [11]. Even more abstractly, a substantial amount of classical harmonic analysis in $\mathbb{R}^n$ can be pushed through to the setting of a measure space equipped with a diffusion semigroup, as found in Stein's book [22]. A limiting aspect of the theory developed in this book is the absence of an explicit geometry; though the statement of maximal theorems, Littlewood-Paley theorems and interpolation theorems make sense in this general setting, basic notions such as Lipschitz functions cannot be defined as there is no distance on the space.

A natural way of overcoming this lack of geometry is to use the semigroup itself to define a distance. This is the approach taken in the theory of *diffusion maps* [5]. If the kernel of the semigroup at time t is denoted by $a_t(x, y)$, then the *diffusion distance* at time t is defined as $\|a_t(x, \cdot) - a_t(y, \cdot)\|_2$. This conceptually meaningful distance has found wide application in machine learning, where the kernel $a_t(x, y)$ is a power of an affinity matrix measuring the relationship between two points in a data set.

The idea of letting the semigroup itself define a distance on the data is the starting point for the present work. However, the ground distances $D_\alpha(x, y)$ we introduce in Section 2 are not defined at a fixed scale, but incorporate all scales at once; the parameter α controls the weight placed at each scale, but all scales are present. In a variety of examples, we will show that this distance is equivalent to a "snowflake" of the the intrinsic distance $\rho(x, y)$ on the underlying space; that is, the distance $\rho(x, y)$ raised to a power less than 1 [12]. However, the distance we define makes sense even when there is no externally given distance.

Next, we consider the space of functions on the measure space that are Lipschitz with respect to the distance $D_\alpha(x, y)$. In the settings where $D_\alpha(x, y)$ is a snowflake of an intrinsic distance $\rho(x, y)$, Lipschitz functions are Hölder with respect to $\rho(x, y)$. We will give simple characterizations of the Lipschitz norm using the semigroup itself, analogous to corresponding formulas from classical harmonic analysis. As in the classical setting, the basic principle is that the size of a function's variation across scales, as defined by the semigroup, is equivalent to the size of its variation across space with respect to the distance $D_\alpha(x, y)$.

Finally, we consider the space of measures that can be integrated against Lipschitz functions – that is, the space dual to the Lipschitz space. We will give simple characterizations of the norm on this space that generalize formulas from classical analysis. This is of particular interest as the dual norm of the difference between two probability measures is equal to the *Earth Mover's Distance (EMD)* between the two probability measures. We will recall the definition and some basic properties of EMD in Section 6; we note here that it is a popular tool in machine learning applications that suffers from high computational cost, and that the equivalent metrics we develop provide, in many situations, a fast way of approximating it.

We will play close attention to the regularity conditions we impose on the semigroup needed to make our theory hold. The conditions are highly non-restrictive, and we will show that they hold for a very broad class of semigroups. Examples include heat kernels on closed Riemannian manifolds, heat kernels on certain fractals, and subordinated heat kernels in $\mathbb{R}^n$, as well as the non-symmetric example of shifted heat kernels on $\mathbb{R}^n$. In addition, we will show that if the theory holds for some finite collection of semigroups on different spaces, then it holds for their product on the cross-product of these spaces. This subsumes the theory of mixed-homogeneity distances.

Another contribution of this paper is to generalize the aforementioned results to the setting of tensor products of two or more measure spaces, each equipped with its own semigroup. We derive characterizations of the norms on the space of mixed Lipschitz functions and its dual. The application areas to which this topic applies is the comparison of two databases that can be viewed as functions on a product of two spaces, each with its own natural geometry defined by a semigroup. For example, we might wish to compare the spectrograms of two signals. The time and frequency domains each have their own natural geometry, independent of the other. The metric we derive provides a natural way of comparing the spectrograms by the maximum difference in their response to a class of sensors, the mixed Lipschitz functions.

1.1 Notation

By "$A \lesssim B$" or "$B \gtrsim A$" we mean inequalities up to positive constants; that is, there is a constant $C > 0$ such that $A \leq C \cdot B$. Similarly, by "$A \simeq B$" we mean there are constants $c, C > 0$ such that $c \cdot A \leq B \leq C \cdot A$. What is meant by C being a "constant" will be clear in each instance. The other notation used throughout is specific to each section and will be defined as it is encountered.

2 Multiscale diffusion distance

Our setting throughout the paper will be an abstract measure space $\mathcal{X}$, whose measure we will denote dx. We suppose that $\mathcal{X}$ is equipped with a family of kernels $a_t(x,y)$, $t > 0$, in L^1. Defining the operators

$$A_t f(x) = \int_{\mathcal{X}} a_t(x,y) f(y) dy$$

we assume the following conditions:

(S) (The semigroup property.) For all $t, s > 0$, $A_t A_s = A_{t+s}$. This property can be expressed in terms of the kernels $a_t(x,y)$ as

$$a_{t+s}(x,y) = \int_{\mathcal{X}} a_t(x,w) a_s(w,y) dw.$$

(C) (The conservation property.) If $\mathbf{1}$ is the constant function 1 on $\mathcal{X}$, then for all $t > 0$, $A_t\mathbf{1} = \mathbf{1}$. This property can be expressed in terms of the kernels $a_t(x, y)$ as

$$\int_{\mathcal{X}} a_t(x, y)dy = 1.$$

(I) (The integrability property.) There is a constant $C > 0$ such that for all $t > 0$ and $x \in \mathcal{X}$,

$$\int_{\mathcal{X}} |a_t(x, y)|dy \leq C.$$

(R) (The regularity property.) There are constants $C > 0$ and $\alpha > 0$ such that for every $1 \geq s \geq t > 0$ and every $x \in \mathcal{X}$,

$$\int_{\mathcal{X}} |a_t(x, y)| \cdot \|a_s(x, \cdot) - a_s(y, \cdot)\|_1 \, dy \leq C\left(\frac{t}{s}\right)^{\alpha}.$$

We will have more to say about the regularity condition (R) later in this section. In particular, we will give an alternate characterization that reveals its geometric content. In Section 3, we will show that condition (R) holds for a wide variety of examples. In Section 4 we will show that this condition also implies convergence to the identity for a class of suitably regular functions.

As noted in the Introduction, we do not assume any external geometry on the space $\mathcal{X}$. Rather, we will use the kernels $a_t(x, y)$ to *define* a geometry from scratch. This approach is inspired by the paper [5]. In this work, each time t defines a *diffusion distance*: the time t diffusion distance between x and y is the L^2 distance between $a_t(x, \cdot)$ and $a_t(y, \cdot)$. These distances also have the feature that they can be approximately embedded into a low-dimensional Euclidean space. Each distance captures the geometry of the space at a particular scale.

In the present work, however, we consider a single distance that incorporates all scales at once. Also, we use the L^1 distance between kernels at each scale, rather than the L^2 distance. This avoids use of the spectral theory of the operators A_t present in [5]. Furthermore, though there are not Euclidean embeddings of the distance we define as with L^2 diffusion distance, for the application areas we have in mind there will usually be no need to explicitly compute our distance for all pairs of points.

We will be concerned with dyadic scales $t \in (0, 1]$; that is, scales $t = 2^{-k}, k \geq 0$. To this end, define

$$P_k = A_{2^{-k}}$$

and

$$p_k(x, y) = a_{2^{-k}}(x, y).$$

Also, we define

$$D_k(x, y) = \|p_k(x, \cdot) - p_k(y, \cdot)\|_1 .$$

Define the multiscale distance

$$D_\alpha(x,y) = \sum_{k\geq 0} 2^{-k\alpha} D_k(x,y). \tag{1}$$

Note that the condition $\int_{\mathcal{X}} |a_t(x,y)|dy \leq C$ guarantees $D_\alpha(x,y)$ is uniformly bounded for all x, y; in particular, it is finite.

In Section 3 we will compute the distance $D_\alpha(x,y)$ for many examples of semigroups. Before doing so, however, it will be convenient to turn our attention to the regularity condition (R) we impose on the kernels $a_t(x,y)$. We reformulate condition (R) in geometric terms, where the geometry is defined by the distance $D_\alpha(x,y)$. To that end, define the geometric condition (G) by

(G) (The geometric property.) There are constants $C > 0$ and $\alpha > 0$ such that for all $k \geq 0$ and $x \in \mathcal{X}$,

$$\int_{\mathcal{X}} |p_k(x,y)| \cdot D_\alpha(x,y)dy \leq C 2^{-k\alpha}.$$

We show that conditions (R) and (G) are essentially equivalent. The following lemma will be convenient.

Lemma 1. *If there are constants $C > 0$ and $\alpha > 0$ such that for every $k, l \geq 0$ and $x \in \mathcal{X}$*

$$\int |p_k(x,y)| \sum_{l=0}^{k} \|p_l(x,\cdot) - p_l(y,\cdot)\|_1 \, dy \leq C 2^{-k\alpha}.$$

then (G) holds, for the same choice of α and a possibly different constant C.

Proof. By the integrability condition (I) the integrals $\int_{\mathcal{X}} |p_k(x,y)|dy$ are uniformly bounded. Therefore

$$\sum_{l \geq k+1} 2^{-l\alpha} \|p_k(x,\cdot) - p_k(y,\cdot)\|_1 \lesssim 2^{-k\alpha}$$

and so

$$\int_{\mathcal{X}} |p_k(x,y)| \sum_{l \geq k+1} 2^{-l\alpha} \|p_k(x,\cdot) - p_k(y,\cdot)\|_1 \, dy \lesssim 2^{-k\alpha} \int_{\mathcal{X}} |p_k(x,y)|dy \lesssim 2^{-k\alpha}$$

from which the result follows. □

Proposition 1. *Suppose that (R) holds for some $\alpha > 0$ and all dyadic times $s = 2^{-l}, t = 2^{-k}$, where $0 \leq l \leq k$. Then (G) holds for all $k \geq 0$ and for any $0 < \alpha' < \alpha$.*

Proof. For all x, we have

$$\int_{\mathcal{X}} |p_k(x,y)| 2^{-l\alpha'} \|p_l(x,\cdot) - p_l(y,\cdot)\|_1 \lesssim 2^{-k\alpha} 2^{l(\alpha-\alpha')}.$$

Summing over $l = 0, \ldots, k$ gives

$$\int_{\mathcal{X}} |p_k(x,y)| \sum_{l=0}^{k} 2^{-l\alpha'} \|p_l(x,\cdot) - p_l(y,\cdot)\|_1 \lesssim 2^{-k\alpha} \sum_{l=0}^{k} 2^{l(\alpha-\alpha')} \lesssim 2^{-k\alpha} 2^{k(\alpha-\alpha')} = 2^{-k\alpha'}.$$

By Lemma 1, we are done. □

Proposition 2. *Suppose condition (G) holds for some $\alpha > 0$. Then (R) holds for all dyadic times $s = 2^{-l}, t = 2^{-k}$, and for all $0 < \alpha' \leq \alpha$. In other words, for all $0 < \alpha' \leq \alpha$ there is a constant C such that for all $0 \leq l \leq k$ and $x \in \mathcal{X}$,*

$$\int_{\mathcal{X}} |p_k(x,y)| \cdot \|p_l(x,\cdot) - p_l(y,\cdot)\|_1 \, dy \leq C 2^{-(k-l)\alpha'}.$$

Proof. Since $2^{-l\alpha} \|p_l(x,\cdot) - p_l(y,\cdot)\|_1 \leq D_\alpha(x,y)$ for all $l \geq 0$, we have

$$\int_{\mathcal{X}} |p_k(x,y)| \cdot \|p_l(x,\cdot) - p_l(y,\cdot)\|_1 \, dy \leq 2^{l\alpha} \int_{\mathcal{X}} |p_k(x,y)| D_\alpha(x,y) dy \lesssim 2^{-(k-l)\alpha}.$$

Since $\alpha' \leq \alpha$, the result follows. □

We will find condition (G) to be a more useful statement of regularity than (R) going forward. We note too that to recover (G) we need only assume (R) for dyadic times s and t between 0 and 1.

3 Examples of kernels satisfying our conditions

In this section, we show that the conditions we impose on the kernels $a_t(x,y)$ hold for a great diversity of semigroups arising in different settings. In all the examples we consider here, the analysis proceeds by obtaining an upper bound on the distance $D_\alpha(x,y)$. Although the metric $D_\alpha(x,y)$ given by equation (1) is defined from the kernels $a_t(x,y)$ themselves — that is, it is not given externally — there are many examples of spaces $\mathcal{X}$ on which there is already defined a natural distance $\rho(x,y)$ with respect to which the kernels $a_t(x,y)$ exhibit certain regularity. We will show that if $a_t(x,y)$ satisfies a certain Hölder continuity condition and a decay condition, then the distance $D_\alpha(x,y)$ is bounded above by a power of $\rho(x,y)$.

It will follow easily that condition (G) (and consequently (R) as well) is true for such kernels, and thus that the general results in this paper apply in this setting. Furthermore, by imposing even stronger assumptions on $a_t(x,y)$ we will show that the distance $D_\alpha(x,y)$ is in fact equivalent to a power of $\rho(x,y)$. We will follow up with a number of specific examples that satisfy one or both sets of conditions.

Throughout this section, we will always assume $0 < \alpha < 1$.

3.1 Hölder continuous kernels with decay

Suppose that there is a metric $\rho(x, y)$ on $\mathcal{X}$ and a measure μ such that $\mu(B(x, r)) \lesssim r^n$, where $n > 0$ is fixed. In addition to the conservation property (C) and the uniform L^1 bound (I) that we already assume, the kernel $a_t(x, y)$ is assumed to be symmetric, and the following two regularity conditions are imposed:

1. An upper bound on the kernel: there is a non-negative, monotonic decreasing function Φ on $[0, \infty)$ and a number $\beta > 0$ such that for any $\gamma < \beta$,
$$\int^{\infty} \tau^{n+\gamma} \Phi(\tau) \frac{d\tau}{\tau} < \infty$$
and
$$|a_t(x, y)| \leq \frac{1}{t^{n/\beta}} \Phi\left(\frac{\rho(x, y)}{t^{1/\beta}}\right).$$

2. The Hölder continuity estimate: there is some constant $\Theta > 0$ sufficiently small for such that for all $t \in (0, 1]$ and for all x, y, u with $\rho(x, y) \leq t^{1/\beta}$,
$$|a_t(x, u) - a_t(y, u)| \leq \left(\frac{\rho(x, y)}{t^{1/\beta}}\right)^{\Theta} \frac{1}{t^{n/\beta}} \Phi\left(\frac{\rho(x, y)}{t^{1/\beta}}\right).$$

These conditions are found in the paper [11]. As discussed there, examples of semigroups satisfying these estimates include the subordinated heat kernels in $\mathbb{R}^n$, the heat kernel on certain Riemanninan manifolds, the heat kernel on a variety of fractals such as the unbounded Sierpinksi Gasket, and the heat kernel of the semigroup e^{-tL} for certain elliptic operators L on $\mathbb{R}^n$.

We will show that if we assume conditions 1 and 2, then our geometric condition (G) is satisfied for all $\alpha < \min\{1, \Theta/\beta\}$. The first step in showing this is to prove that our distance $D_\alpha(x, y)$ defined from the semigroup is bounded above by a power of the distance $\rho(x, y)$.

Lemma 2. *For any $0 \leq \eta < 1$, there is a finite constant $C > 0$ such that for every $0 < t \leq 1$ and every $x \in \mathcal{X}$,*
$$\int_{\mathcal{X}} \rho(x, y)^{\beta\eta} \frac{1}{t^{n/\beta}} \Phi\left(\frac{\rho(x, y)}{t^{1/\beta}}\right) dy \leq C t^{\eta}.$$

Proof. Let $V_k = B(x, 2^{k+1} t^{1/\beta}) \setminus B(x, 2^k t^{1/\beta})$. The upper bound on the kernel from condition 1 yields the following inequality:

$$\begin{aligned}\int_{\mathcal{X}} \rho(x,y)^{\beta\eta}\frac{1}{t^{n/\beta}}\Phi\biggl(\frac{\rho(x,y)}{t^{1/\beta}}\biggr)dy &= \frac{1}{t^{n/\beta}}\biggl\{\int_{B(x,t^{1/\beta})} + \sum_{k=0}^{\infty}\int_{V_k}\biggr\}\rho(x,y)^{\beta\eta}\Phi\biggl(\frac{\rho(x,y)}{t^{1/\beta}}\biggr)dy \\ &\lesssim t^{\eta}t^{-n/\beta}\biggl\{\Phi(0)\mu(B(x,t^{1/\beta})) \\ &\qquad + \sum_{k=0}^{\infty} 2^{k\eta\beta}\Phi(2^k)\mu(B(x,2^{k+1}t^{1/\beta}))\biggr\} \\ &\lesssim t^{\eta}t^{-n/\beta}\biggl\{\Phi(0)t^{n/\beta} + \sum_{k=0}^{\infty}\Phi(2^k)2^{k(n+\eta\beta)}t^{n/\beta}\biggr\} \\ &\lesssim t^{\eta}\biggl\{\Phi(0) + \int_1^{\infty}\tau^{n+\eta\beta}\Phi(\tau)\frac{d\tau}{\tau}\biggr\} \lesssim t^{\eta}.\end{aligned}$$

We used that $\eta < 1$ and condition 1 to conclude that the last integral is finite. This is the desired result. □

Proposition 3. *For every $0 < \alpha < \min\{1, \Theta/\beta\}$, there is a constant $C > 0$ such that $D_\alpha(x,y) \le C\rho(x,y)^{\alpha\beta}$.*

Proof. First, condition 2 and Lemma 2 above with $\eta = 0$ imply that whenever $\rho(x,y) \le t^{1/\beta}$,

$$\|a_t(x,\cdot) - a_t(y,\cdot)\|_1 \le \biggl(\frac{\rho(x,y)}{t^{1/\beta}}\biggr)^{\Theta}\frac{1}{t^{n/\beta}}\int_{\mathcal{X}}\Phi\biggl(\frac{\rho(x,y)}{t^{1/\beta}}\biggr)dy \lesssim \biggl(\frac{\rho(x,y)}{t^{1/\beta}}\biggr)^{\Theta}.$$

Consequently, if we define K so that $2^{-K} \le \rho(x,y)^{\beta} < 2^{-K+1}$, then

$$\begin{aligned}D_\alpha(x,y) &\lesssim \rho(x,y)^{\Theta}\sum_{k=0}^{K}2^{-k\alpha}2^{k\Theta/\beta} + \sum_{k=K+1}^{\infty}2^{-k\alpha} \lesssim \rho(x,y)^{\Theta}2^{K(\Theta/\beta-\alpha)} + 2^{-K\alpha} \\ &\lesssim \rho(x,y)^{\alpha\beta}.\end{aligned}$$

We used that $\alpha < \Theta/\beta$ for the upper bound on the first sum. □

With this upper bound on $D_\alpha(x,y)$, it is now straightforward to show that our geometric condition (G) holds for a range of α.

Proposition 4. *Under conditions 1 and 2, condition (G) holds for all $0 < \alpha < \min\{1, \Theta/\beta\}$.*

Proof. From Proposition 3, we have the upper bound $D_\alpha(x,y) \lesssim \rho(x,y)^{\alpha\beta}$. Consequently, taking $\eta = \alpha$ in Lemma 2 yields

$$\int_{\mathcal{X}}|a_t(x,y)|D_\alpha(x,y)dy \lesssim \int_{\mathcal{X}}\rho(x,y)^{\alpha\beta}\frac{1}{t^{n/\beta}}\Phi\biggl(\frac{\rho(x,y)}{t^{1/\beta}}\biggr)dy \lesssim t^{\alpha}$$

which is the desired result. □

3.2 The distance $D_\alpha(x,y)$ for kernels with a matching lower bound

Having established conditions (G) and (R) from the upper bound $D_\alpha(x,y) \lesssim \rho(x,y)^{\alpha\beta}$ for all $\alpha < \min\{1, \Theta/\beta\}$ under the continuity and decay conditions 1 and 2 of the previous section, we now formulate general conditions under which we can prove a corresponding lower bound, $D_\alpha(x,y) \gtrsim \rho(x,y)^{\alpha\beta}$. We will then study several examples where both conditions are satisfied. Note that the lower bound is not necessary for the general results of our paper to hold; in particular, our primary concern is to establish condition (G) (and hence condition (R) as well) for a large class of examples. We only prove the lower bounds to show that the distance $D_\alpha(x,y)$ is equivalent to the "natural" geometry of the space under consideration in a plethora of cases.

Again, we suppose in this section that $\mathcal{X}$ comes equipped with a metric $\rho(x,y)$. We assume, however, a stronger relation between the measure and the metric μ, namely the two-sided estimate $\mu(B(x,r)) \simeq r^n$.

We suppose that in addition to the conditions 1 and 2 of the previous section, we also have the following condition:

3. A lower bound on the kernel: there is a monotonic decreasing function Ψ on $[0,\infty)$ and $R > 0$ such that for all $t \in (0,1]$ and all $\rho(x,y) < R$
$$|a_t(x,y)| \geq \frac{1}{t^{n/\beta}} \Psi\left(\frac{\rho(x,y)}{t^{1/\beta}}\right).$$

We will show

Proposition 5. *Under the conditions 1,2 and 3, $D_\alpha(x,y) \gtrsim \min\{1, \rho(x,y)^{\alpha\beta}\}$.*

We will deduce the result from the following lemmas.

Lemma 3. *There is a constant $A > 1$ and a constant $\epsilon > 0$ such that whenever $At^{1/\beta} \leq \rho(x,y) < R$, we have*
$$\|a_t(x,\cdot) - a_t(y,\cdot)\|_1 > \epsilon.$$

Proof. Temporarily fix any $A > 1$ and suppose $At^{1/\beta} \leq \rho(x,y) < R$. Then for any $u \in B(x, t^{1/\beta})$, the triangle inequality implies
$$\rho(y,u) \geq \rho(x,y) - \rho(x,u) \geq (A-1)t^{1/\beta}.$$

From the monotonicity of Φ it follows that $\Phi(\rho(y,u)/t^{1/\beta}) \leq \Phi(A-1)$. Consequently, using the upper and lower bounds on $|a_t(x,y)|$ and the fact that $\mu(B(x,r)) \simeq r^n$, we have
$$\begin{aligned}
\|a_t(x,\cdot) - a_t(y,\cdot)\|_1 &\geq \int_{B(x,t^{1/\beta})} |a_t(x,u)|du - \int_{B(x,t^{1/\beta})} |a_t(y,u)|du \\
&\geq \frac{1}{t^{n/\beta}} \int_{B(x,t^{1/\beta})} \Psi\left(\frac{\rho(x,u)}{t^{1/\beta}}\right)du - \frac{1}{t^{n/\beta}} \int_{B(x,t^{1/\beta})} \Phi\left(\frac{\rho(y,u)}{t^{1/\beta}}\right) \\
&\geq \frac{1}{t^{n/\beta}} \int_{B(x,t^{1/\beta})} \Psi\left(\frac{\rho(x,u)}{t^{1/\beta}}\right)du - \frac{1}{t^{n/\beta}} \int_{B(x,t^{1/\beta})} \Phi(A-1) \\
&\gtrsim \Psi(1) - \Phi(A-1).
\end{aligned}$$

Since Φ is decreasing, by choosing A large enough we can guarantee $\epsilon \equiv \Psi(1) - \Phi(A-1) > 0$, yielding the desired result. □

Corollary 1. *For all $\rho(x,y) < R$,*

$$D_\alpha(x,y) \gtrsim \rho(x,y)^{\alpha\beta}$$

Proof. By the previous lemma, $\|a_t(x,\cdot) - a_t(y,\cdot)\|_1 > \epsilon > 0$ whenever $At^{1/\beta} \le \rho(x,y)$. Let $L = \lfloor \log_2(\rho(x,y)^\beta / A^\beta) \rfloor$. Then

$$D_\alpha(x,y) \ge \sum_{k=L}^{\infty} 2^{-k\alpha} \|p_k(x,\cdot) - p_k(y,\cdot)\|_1 \ge \epsilon \sum_{k=L}^{\infty} 2^{-k\alpha} \simeq 2^{-L\alpha} \simeq \rho(x,y)^{\alpha\beta}.$$

□

Lemma 4. *There are constants $C > 0$ and $\delta > 0$ such that whenever $\rho(x,y) \ge R$ and $t^{1/\beta} < \delta R$,*

$$\|a_t(x,\cdot) - a_t(y,\cdot)\|_1 \ge C.$$

Proof. Since $\rho(x,y) \ge R$, the balls $B(x, R/2)$ and $B(y, R/2)$ are disjoint. Consequently

$$\begin{aligned}\|a_t(x,\cdot) - a_t(y,\cdot)\|_1 &\ge \int_{B(x,R/2)} |a_t(x,u)|du - \int_{B(x,R/2)} |a_t(y,u)|du \\ &\ge \int_{B(x,R/2)} |a_t(x,u)|du - \int_{B(y,R/2)^c} |a_t(y,u)|du \\ &\ge 1 - \int_{B(x,R/2)^c} |a_t(x,u)|du - \int_{B(y,R/2)^c} |a_t(y,u)|du.\end{aligned}$$

The result follows from the next lemma. □

Lemma 5. *Fix any $r > 0$ and $\epsilon > 0$. Then there exists some $\delta > 0$ sufficiently small so that whenever $0 < t^{1/\beta} < \delta r$,*

$$\int_{B(x,r)^c} |a_t(x,u)|du \le \epsilon.$$

Proof. Temporarily fix $\delta > 0$ and suppose $0 < t^{1/\beta} < \delta r$. Then if we let $V_k = B(x, 2^{k+1}r) \setminus B(x, 2^k r)$, we have

$$\begin{aligned}\int_{B(x,r)^c} |a_t(x,u)|du &\le \frac{1}{t^{n/\beta}} \int_{B(x,r)^c} \Phi\left(\frac{\rho(x,y)}{t^{1/\beta}}\right) du = \frac{1}{t^{n/\beta}} \sum_{k\ge1} \int_{V_k} \Phi\left(\frac{\rho(x,y)}{t^{1/\beta}}\right) du \\ &\lesssim \frac{1}{t^{n/\beta}} \sum_{k\ge1} \Phi\left(\frac{2^k r}{t^{1/\beta}}\right)(2^{k+1}r)^n \lesssim \frac{r^n}{t^{n/\beta}} \int_1^\infty \tau^n \Phi\left(\frac{\tau r}{t^{1/\beta}}\right) \frac{d\tau}{\tau} \\ &= \frac{r^n}{t^{n/\beta}} \int_{rt^{-1/\beta}}^\infty t^{n/\beta} r^{-n} s^n \Phi(s) \frac{ds}{s} \le \int_{\delta^{-1}}^\infty s^n \Phi(s) \frac{ds}{s}.\end{aligned}$$

By taking δ small enough, the integral can be made as small as desired, completing the proof. □

Corollary 2. *There is a constant $B > 0$ such that whenever $\rho(x,y) \geq R$, $D_\alpha(x,y) \geq B$.*

Proof. Take C and δ from Lemma 4. Let $K = \lfloor \log_2(1/(\delta^\beta R^\beta)) \rfloor$. Then $2^{-K} \leq \delta^\beta R^\beta$, and so

$$D_\alpha(x,y) \geq \sum_{k=K}^{\infty} 2^{-k\alpha} C \simeq C\delta^{\alpha\beta} R^{\alpha\beta} > 0$$

which completes the proof. □

Corollary 1 and Corollary 2 easily imply Proposition 5. Furthermore, Proposition 5 and Proposition 3 yield the following theorem:

Theorem 1. *If all the conditions 1, 2 and 3 on $a_t(x,y)$ apply, then for $0 < \alpha < \min\{1, \Theta/\beta\}$ the distance $D_\alpha(x,y)$ is equivalent to the thresholded distance $\min\{1, \rho(x,y)^{\alpha\beta}\}$.*

3.3 Heat kernel on a Riemannian manifold

We illustrate the results above on some selected examples. First, we consider the case where $\mathcal{X}$ is a closed (compact, without boundary) Riemannian manifold of dimension n, and $a_t(x,y)$ is its heat kernel. Since $\mathcal{X}$ is compact, it is true that $\mu(B(x,r)) \simeq r^n$. Furthermore, the following two lemmas can be easily derived from the parametrix construction of the heat kernel given in Chapter VI, Section 4 of [4]. Here, $\rho(x,y)$ is geodesic distance on the manifold.

Lemma 6. *There are positive constants A, B such that*

$$a_t(x,y) \leq \frac{A}{t^{n/2}} e^{-B\rho(x,y)^2/t}$$

for all t sufficiently small.

Lemma 7. *There are positive constants C, D*

$$\frac{C}{t^{n/2}} e^{-D\rho(x,y)^2/t} \leq a_t(x,y)$$

whenever $t \in (0,1]$ and $\rho(x,y)$ are sufficiently small.

In other words, we have the upper and lower bounds on the kernel from conditions 1 and 3, with $\Phi(\tau) \simeq \Psi(\tau) \simeq e^{-\tau^2}$ and $\beta = 2$. It remains to show condition 2. We will deduce the continuity estimate from the following gradient bound:

Lemma 8. *There are constants $E, F > 0$ such that for all $t \in (0,1]$ and for all x and y in $\mathcal{X}$,*

$$\|\nabla_x a_t(x,y)\| \leq \frac{E}{\sqrt{t}} \frac{e^{-F\rho(x,y)^2/t}}{t^{n/2}}$$

where ∇_x denotes the gradient with respect to the first variable.

Proof. Using the asymptotic expansion of $a_t(x,y)$ in Chapter VI, Section 4 of [4], it is easy to show a Gaussian upper bound on the time derivative of $a_t(x,y)$, namely

$$\left|\frac{\partial a_t}{\partial t}(x,y)\right| \leq \frac{b}{t}\frac{e^{-c\rho(x,y)^2/t}}{t^{n/2}}$$

for some positive constants b, c. Since the curvature of $\mathcal{X}$ is bounded (because $\mathcal{X}$ is compact) we can apply Theorem 1.4 from [14], which states that there are constants A_1, A_2, A_3 such that

$$\|\nabla_x a_t(x,y)\|^2 \leq \left(A_1 + \frac{A_2}{t}\right) a_t(x,y)^2 + A_3 a_t(x,y)\frac{\partial a_t}{\partial t}(x,y).$$

For $t \in (0,1]$, it follows from the Gaussian estimates on $a_t(x,y)$ and $\partial_t a_t(x,y)$ that

$$\|\nabla_x a_t(x,y)\|^2 \leq \frac{\tilde{A}}{t}\frac{e^{-2B\rho(x,y)^2/t}}{t^n} + \frac{b}{t}\frac{e^{-c\rho(x,y)^2/t}}{t^{n/2}}\frac{A}{t^{n/2}}e^{-B\rho(x,y)^2/t} \lesssim \frac{1}{t}\frac{e^{-\tilde{B}\rho(x,y)^2/t}}{t^n}$$

for sufficiently small $\tilde{B} > 0$, from which the result follows. □

Lemma 9. *If x, y are sufficiently close, then for any smooth function $h : \mathcal{X} \to \mathbb{R}$ and any two points x and y in $\mathcal{X}$, there is a point $\tilde{x}$ lying on the minimal geodesic from x to y such that*

$$|h(x) - h(y)| \leq \|\nabla h(\tilde{x})\|\,\rho(x,y).$$

Proof. Suppose $r \equiv \rho(x,y)$ is less than the injectivity radius of the manifold M (which is positive, since M is compact). Let $\gamma(t)$ be the unit speed geodesic connecting x to y. Then $\gamma(r) = y$, and $\gamma(0) = x$. For details, see, for instance, Chapter 13, Section 2 of [8].

Consider the function $\hat{h}(t) = h(\gamma(t))$. Observe that $\hat{h}(0) = h(x)$ and $\hat{h}(r) = h(y)$. By the mean value theorem, there is some point t_1 between 0 and r such that

$$\frac{h(y) - h(x)}{\rho(x,y)} = \frac{\hat{h}(r) - \hat{h}(0)}{r} = \hat{h}'(t_1) = \left.\frac{d}{dt}h(\gamma(t))\right|_{t=t_1} = \langle \nabla h(\gamma(t_1)), \gamma'(t_1)\rangle$$

Consequently, since γ has unit speed, the Cauchy-Schwarz inequality gives

$$|h(y) - h(x)| = |\langle \nabla h(\gamma(t_1)), \gamma'(t_1)\rangle|\rho(x,y) \leq \|\nabla h(\gamma(t_1))\|\,\rho(x,y).$$

Consequently, if we let $\tilde{x} = \gamma(t_1)$, then $\tilde{x}$ lies on the minimal geodesic connecting x and y, and

$$|h(x) - h(y)| \leq \|\nabla h(\tilde{x})\|\,\rho(x,y).$$

□

Corollary 3. *There are positive constants G, H such that whenever $\rho(x,y) \leq t^{1/2}$,*

$$|a_t(x,u) - a_t(y,u)| \leq G\frac{\rho(x,y)}{\sqrt{t}}\frac{e^{-H\rho(x,y)^2/t}}{t^{n/2}}.$$

Proof. From the mean value theorem (Lemma 9) and the gradient estimate from Lemma 8, we have the bound

$$|a_t(x,u) - a_t(y,u)| \leq \rho(x,y)\frac{E}{\sqrt{t}}\frac{e^{-F\rho(u,\tilde{x})^2/t}}{t^{n/2}}.$$

where $\tilde{x}$ is some point on the minimal geodesic connecting x and y. Since $\rho(x,y) \leq t^{1/2}$, it is also true that $\rho(x,\tilde{x}) \leq t^{1/2}$. Consequently, we have

$$\rho(u,x)^2 \leq 2\rho(u,\tilde{x})^2 + 2\rho(\tilde{x},x)^2 \leq 2\rho(u,\tilde{x})^2 + 2t$$

and so

$$\begin{aligned}|a_t(x,u) - a_t(y,u)| &\leq \rho(x,y)\frac{E}{\sqrt{t}}\frac{e^{-F\rho(u,\tilde{x})^2/t}}{t^{n/2}} \leq \rho(x,y)\frac{E}{\sqrt{t}}\frac{e^{-F(\rho(u,x)^2-2t)/2t}}{t^{n/2}}\\ &\leq \rho(x,y)\frac{Ee}{\sqrt{t}}\frac{e^{-(F/2)\rho(u,x)^2/t}}{t^{n/2}}\end{aligned}$$

which is the desired result. □

We can therefore apply the results of the previous sections to the Gaussian upper bound from Lemma 6 and the continuity estimate from Corollary 3 to conclude that condition (G) is satisfied for all $\alpha < 1/2$, and that $D_\alpha(x,y) \lesssim \rho(x,y)^{2\alpha}$. Furthermore, the Gaussian lower bound from Lemma 7 and the fact that the geodesic distance is bounded (since $\mathcal{X}$ is compact) shows that $D_\alpha(x,y) \gtrsim \rho(x,y)^{2\alpha}$ as well. Consequently, $D_\alpha(x,y)$ is equivalent to the distance $\rho(x,y)^{2\alpha}$ when $\alpha < 1/2$.

Of course we can derive a similar result, namely that condition (G) holds for $\alpha < 1/2$ and that $D_\alpha(x,y) \simeq \min\{1, \rho(x,y)^{2\alpha}\}$, for non-closed manifolds whose heat kernels satisfy the same Gaussian bounds. We note again that we only need condition (G) to hold for our theory to hold, and hence only need the Gaussian upper bound, the continuity estimate, and the upper bound $\mu(B(x,r)) \lesssim r^n$. For example, as discussed in [11], the Gaussian upper bounds and continuity estimates hold for the heat kernel on any geodesically complete Riemannian manifold with non-negative curvature. The lower bounds are only used to prove the lower bound $D_\alpha(x,y) \gtrsim \min\{1, \rho(x,y)^{2\alpha}\}$.

We note too that this section may be of particular interest to those in the machine-learning community, as approximations to the heat kernel on data sampled from submanifolds of $\mathbb{R}^n$ are a widely-used model for many data sets, e.g. a collection of high-dimensional images defined by a relatively small number of parameters. See, for instance, [16, 17].

3.4 Subordinated heat kernels with shifts on $\mathbb{R}^n$

Next we consider the case in which $a_t(x,y) = K_t(x-y)$, where $K_t(u)$ is a radial kernel, i.e. $K_t(x) = K_t(y)$ if $|x| = |y|$, satisfying the following scaling property:

$$K_t(x) = t^{-n/\beta} K_1(t^{-1/\beta}x)$$

where $0 < \beta \leq 2$. For details on the construction of such kernels in one dimension, the reader can refer to the book [26]; also see the paper [10] for a useful summary. These kernels are known as *subordinated heat kernels* on $\mathbb{R}^n$, since they can be expressed as an average of the heat kernel at different scales. Concretely, when $0 < \beta < 2$ (that is, $\beta \neq 2$), $K_t(x)$ is of the form

$$K_t(x) = \int_0^\infty \eta_t(s) g_s(x) ds$$

where g_s is the Gaussian kernel at time s, and for each t the function $\eta_t(s)$ is a probability density on $(0,\infty)$, known as the *subordinator.* In fact, η_t satisfies the identity

$$\exp(-t\lambda^{\beta/2}) = \int_0^\infty \eta_t(s) e^{-s\lambda} ds$$

for all $\lambda \geq 0$, from which it easily follows that the Fourier transform of K_t is

$$\hat{K}_t(\xi) = \exp(-t|\xi|^\beta).$$

It is shown in [11] that any subordinated heat kernel satisfies conditions 1, 2 and 3. More precisely,

$$a_t(x,y) \simeq \frac{1}{t^{n/\beta}} \left(1 + \frac{|x-y|}{t^{1/\beta}}\right)^{-(n+\beta)}$$

and

$$|a_t(x,u) - a_t(y,u)| \lesssim \frac{|x-y|}{t^{1/m}} \frac{1}{t^{n/m}} \left(1 + \frac{|x-y|}{t^{1/m}}\right)^{-(n+m)}.$$

It follows immediately that the distance $D_\alpha(x,y)$ with respect to the kernel $a_t(x,y)$ is equivalent to $\min\{1, |x-y|^{\alpha\beta}\}$ whenever $\alpha < 1/\beta$. Note that our use of the parameter β in the definition of the subordinated heat kernel coincides with its use in the conditions 1, 2 and 3.

This leads us immediately to a family of examples of *non-symmmetric* semigroup for which condition (G) holds, namely the subordinated heat kernels with shifts. Take $\beta \in [1,2]$ and define $K_t(u)$ as above. Then for a fixed parameter $\theta \in \mathbb{R}$, define

$$a_t(x,y) = t^{-n\beta} K_1(t^{-1/\beta}(x - \theta t - y)).$$

It is easy to check from the semigroup property for the non-shift case $\theta = 0$ that $a_t(x,y)$ is also a semigroup. Furthermore, we still have $D_\alpha(x,y) \simeq \min\{1, |x-y|^{\alpha\beta}\}$. Therefore, we can verify condition (G) directly by writing

$$\begin{aligned}\int_{\mathbb{R}^n} a_t(x,y)D_\alpha(x,y)dy &\lesssim \int_{\mathbb{R}^n} t^{-n/\beta}K_1(t^{-1/\beta}(x-\theta t-y))\min\{1,|x-y|^{\alpha\beta}\}dy\\ &= \int_{\mathbb{R}^n} t^{-n/\beta}K_1(t^{-1/\beta}(x-y))\min\{1,|x-y+\theta t|^{\alpha\beta}\}dy\\ &\leq \int_{\mathbb{R}^n} t^{-n/\beta}K_1(t^{-1/\beta}(x-y))\min\{1,|x-y|^{\alpha\beta}\}dy\\ &\qquad\qquad + \int_{\mathbb{R}^n} t^{-n/\beta}K_1(t^{-1/\beta}(x-y))|\theta t|^{\alpha\beta}dy\\ &\lesssim t^\alpha + t^{\alpha\beta}.\end{aligned}$$

The last line follows from condition (G) in the case $\theta = 0$. As long as $\beta \geq 1$, condition (G) is satisfied. Note that this range of β includes both the heat kernel ($\beta = 2$) and the Poisson kernel ($\beta = 1$).

3.5 Products of kernels and mixed homogeneity kernels

Suppose that $a_t(x_1, x_2)$ and $b_t(y_1, y_2)$ are two semigroups on spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively, for which the conditions (S), (C), (I) and (R) (and thus (G)) hold. We define their product by $c_t((x_1, y_1), (x_2, y_2)) = a_t(x_1, x_2) \cdot b_t(y_1, y_2)$. It is easy to check that the kernel c_t defines a semigroup on $\mathcal{X} \times \mathcal{Y}$, and that the three conditions (S), (C), and (I) all hold. We will check that (G) holds as well. Since we have three semigroups, it will be convenient to distinguish between the distances each one induces. Fixing the distance parameter α, we will write $D^a_\alpha(x_1, x_2)$ for the distance induced by a_t, and similarly for b_t and c_t. We then have:

Proposition 6. *The distance $D^c_\alpha((x_1, y_1), (x_2, y_2))$ on $\mathcal{X}\times\mathcal{Y}$ is equivalent to $D^a_\alpha(x_1, x_2)+D^b_\alpha(y_1, y_2)$.*

Proof. This follows immediately from the following lemma. □

Lemma 10. *For every $z_1 = (x_1, y_1), z_2 = (x_2, y_2)$ in $\mathcal{X} \times \mathcal{Y}$, we have*

$$\|c_t(z_1,\cdot) - c_t(z_2,\cdot)\|_1 \simeq \|a_t(x_1,\cdot) - a_t(x_2,\cdot)\|_1 + \|b_t(y_1,\cdot) - b_t(y_2,\cdot)\|_1 .$$

Proof. First, we prove that

$$\|c_t(z_1,\cdot) - c_t(z_2,\cdot)\|_1 \lesssim \|a_t(x_1,\cdot) - a_t(x_2,\cdot)\|_1 + \|b_t(y_1,\cdot) - b_t(y_2,\cdot)\|_1 .$$

To see this, observe that if $x_1 = x_2$ then

$$\begin{aligned}\|c_t((x_1,y_1),\cdot) - c_t((x_1,y_2),\cdot)\|_1 &= \int_{\mathcal{X}}\int_{\mathcal{Y}} |a_t(x_1,x)||b_t(y_1,y) - b_t(y_2,y)|dydx\\ &\lesssim \|b_t(y_1,\cdot) - b_t(y_2,\cdot)\|_1\end{aligned}$$

where have used condition (I) on the kernels a_t. Similarly,

$$\|c_t((x_1,y_2),\cdot) - c_t((x_2,y_2),\cdot)\|_1 \lesssim \|a_t(x_1,\cdot) - a_t(x_2,\cdot)\|_1 .$$

We therefore have

$$\begin{aligned}\|c_t((x_1,y_1),\cdot) - c_t((x_2,y_2),\cdot)\|_1 &\le \|c_t((x_1,y_1),\cdot) - c_t((x_1,y_2),\cdot)\|_1 \\ &\qquad + \|c_t((x_1,y_2),\cdot) - c_t((x_2,y_2),\cdot)\|_1 \\ &\lesssim \|a_t(x_1,\cdot) - a_t(x_2,\cdot)\|_1 + \|b_t(y_1,\cdot) - b_t(y_2,\cdot)\|_1 ,\end{aligned}$$

as desired.

For the other direction, observe that

$$\begin{aligned}\|c_t(z_1,\cdot) - c_t(z_2,\cdot)\|_1 &= \int_{\mathcal{X}}\int_{\mathcal{Y}} |a_t(x_1,x)b_t(y_1,y) - a_t(x_2,x)b_t(y_2,y)|dydx \\ &\ge \int_{\mathcal{X}}\left|\int_{\mathcal{Y}}[a_t(x_1,x)b_t(y_1,y) - a_t(x_2,x)b_t(y_2,y)]dy\right|dx \\ &= \int_{\mathcal{X}}\left|a_t(x_1,x)\int_{\mathcal{Y}} b_t(y_1,y)dy - a_t(x_2,x)\int_{\mathcal{Y}} b_t(y_2,y)dy\right|dx \\ &= \int_{\mathcal{X}} |a_t(x_1,x) - a_t(x_2,x)|dx = \|a_t(x_1,\cdot) - a_t(x_2,\cdot)\|_1 .\end{aligned}$$

Similarly

$$\|c_t(z_1,\cdot) - c_t(z_2,\cdot)\|_1 \ge \|b_t(y_1,\cdot) - b_t(y_2,\cdot)\|_1$$

from which it follows

$$\|c_t(z_1,\cdot) - c_t(z_2,\cdot)\|_1 \ge \frac{1}{2}(\|a_t(x_1,\cdot) - a_t(x_2,\cdot)\|_1 + \|b_t(y_1,\cdot) - b_t(y_2,\cdot)\|_1)$$

completing the proof. □

From Proposition 6 we can easily deduce that condition (G) holds for c_t if it holds for a_t and b_t.

Proposition 7. *If condition (G) holds for a_t and b_t, then it holds for their product c_t as well.*

Proof. We have, using condition (I) for both a_t and b_t,

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{Y}} |c_t(z_1,z_2)|D^c_\alpha(z_1,z_2)dz_2 \\ &\qquad\lesssim \int_{\mathcal{X}}\int_{\mathcal{Y}} |a_t(x_1,x_2)\cdot b_t(y_1,y_2)|(D^a_\alpha(x_1,x_2) + D^b_\alpha(y_1,y_2))dx_2dy_2 \\ &\qquad\lesssim \int_{\mathcal{X}} |a_t(x_1,x_2)|D^a_\alpha(x_1,x_2)dx_2 + \int_{\mathcal{Y}} |b_t(y_1,y_2)|D^b_\alpha(y_1,y_2)dy_2 \\ &\qquad\lesssim t^\alpha\end{aligned}$$

which is the desired result. □

Of course, these results hold for the product of any number of kernels, not just two, and the proofs are similar. A natural example is the product of subordinated heat kernels on $\mathbb{R}^n$. Suppose that $n = n_1 + \cdots + n_l$ and that on each space $\mathbb{R}^{n_i}$ we have a subordinated heat kernel $a_t^{(i)}(x_i, y_y)$ with scaling β_i. Then as long as $\alpha < \min\{1/\beta_1, \ldots, 1/\beta_l\}$, for $x = (x_1, \ldots, x_l), y = (y_1, \ldots, y_l)$, $x_i, y_i \in \mathbb{R}^{n_i}$, the kernel

$$a_t(x, y) = \prod_{i=1}^{l} a_t^{(i)}(x_i, y_i)$$

generates the distance

$$D_\alpha(x, y) \simeq \min\{1, |x - y|_{\mathrm{MH}}\}$$

where

$$|x - y|_{\mathrm{MH}} = \sum_{i=1}^{l} |x_i - y_i|^{\alpha\beta_i}$$

is a mixed-homogeneity distance on $\mathbb{R}^n$.

4 Lipschitz functions

We now turn to characterizing functions that are Lipschitz with respect to the distance $D_\alpha(x, y)$, for a fixed $\alpha \in (0, 1)$. We assume that α is chosen so that condition (G) holds; in particular, by Proposition 1 if the kernel satisfies condition (R) for some α', we take any $0 < \alpha < \alpha'$.

For a function f on $\mathcal{X}$ define the seminorm

$$V(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{D_\alpha(x, y)}$$

We then define the Lipschitz norm of a function f on $\mathcal{X}$ to be

$$\|f\|_{\Lambda_\alpha} = \sup_x |f(x)| + V(f)$$

and Λ_α is the space of functions f for which this norm is finite.

Our goal in this section is to define two norms equivalent to this one on Λ_α. We define the difference operators

$$\Delta_k = P_{k+1} - P_k, \quad \delta_k = I - P_k.$$

We also define the seminorms

$$V^{(1)}(f) = \sup_{k \geq 0} \sup_x 2^{k\alpha} |\Delta_k f(x)|$$

and

$$V^{(2)}(f) = \sup_{k\geq 0}\sup_{x} 2^{k\alpha}|\delta_k f(x)|.$$

The two norms can now be defined as

$$\|f\|^{(1)}_{\Lambda_\alpha} = \sup_x |f(x)| + V^{(1)}(f)$$

and

$$\|f\|^{(2)}_{\Lambda_\alpha} = \sup_x |f(x)| + V^{(2)}(f).$$

We immediately see the use of condition (R) and its equivalent condition (G) in the following result:

Proposition 8. $V^{(2)}(f) \lesssim V(f)$.

Proof. Take any $k \geq 0$. Since $p_k(x,\cdot)$ has integral 1 for every x, we have

$$\begin{aligned}|f(x) - P_k f(x)| &= \left| f(x) - \int_{\mathcal{X}} p_k(x,y)f(y)dy \right| = \left| \int_{\mathcal{X}} p_k(x,y)(f(x) - f(y))dy \right| \\ &\leq V(f)\int_{\mathcal{X}} |p_k(x,y)|D_\alpha(x,y)dy \lesssim V(f)2^{-k\alpha}\end{aligned}$$

from which the desired inequality follows trivially. □

Corollary 4. $\|f\|^{(2)}_{\Lambda_\alpha} \lesssim \|f\|_{\Lambda_\alpha}$.

Next, we make the following simple observation about uniform convergence:

Lemma 11. *If* $\|f\|^{(2)}_{\Lambda_\alpha} < \infty$*, then* $P_k f$ *converges to* f *uniformly as* $k \to \infty$*.*

Proof. This is clear from the definition of $||f||^{(2)}_{\Lambda_\alpha}$ (more specifically, the definition of $V^{(2)}(f)$). □

Since $\|f\|^{(2)}_{\Lambda_\alpha} \lesssim \|f\|_{\Lambda_\alpha}$, it follows that:

Lemma 12. *For all* $f \in \Lambda_\alpha$*,* $P_k f$ *converges to* f *uniformly as* $k \to \infty$*.*

We now prove:

Proposition 9. *The seminorms* $V^{(1)}(f)$ *and* $V^{(2)}(f)$ *are equivalent for* $f \in \Lambda_\alpha$*.*

Proof. First, write f as a telescopic series:

$$f - P_0 f = \sum_{l=0}^{\infty}[P_{l+1}f - P_l f]$$

where the series converges uniformly by Lemma 12.

Similarly we can write $P_k f$ as a telescopic series

$$P_k f - P_0 f = \sum_{l=0}^{k} [P_{l+1} f - P_l f]$$

and subtracting the two series gives:

$$\begin{aligned}
|f - P_k f| &= \left| \sum_{l=0}^{\infty} (P_{l+1} f - P_l f) - \sum_{l=0}^{k} (P_{l+1} f - P_l f) \right| \\
&= \left| \sum_{l=k+1}^{\infty} (P_{l+1} f - P_l f) \right| \le \sum_{l=k+1}^{\infty} |P_{l+1} f - P_l f| \\
&\le V^{(1)}(f) \sum_{l=k+1}^{\infty} 2^{-l\alpha} = V^{(1)}(f) \frac{2^{-\alpha}}{1 - 2^{-\alpha}} 2^{-k\alpha}
\end{aligned}$$

and consequently

$$2^{k\alpha} \sup_x |\delta_k f(x)| \le \frac{2^{-\alpha}}{1 - 2^{-\alpha}} V^{(1)}(f).$$

Taking the supremum over all $k \ge 0$ shows $V^{(2)}(f) \lesssim V^{(1)}(f)$.

For the other direction, we simply observe

$$|P_k f - P_{k+1} f| \le |(P_k - I) f| + |(P_{k+1} - I) f| \le 2 V^{(2)}(f) 2^{-k\alpha}$$

implying

$$2^{k\alpha} \sup_x |\Delta_k f(x)| \le 2 V^{(2)}(f).$$

Taking the supremum over all $k \ge 0$ gives the result. □

Corollary 5. *The norms $\|f\|_{\Lambda_\alpha}^{(1)}$ and $\|f\|_{\Lambda_\alpha}^{(2)}$ are equivalent on Λ_α.*

Now we turn to proving the main result of this section, namely that $\|f\|_{\Lambda_\alpha}^{(1)}$ and $\|f\|_{\Lambda_\alpha}^{(2)}$ are equivalent to $\|f\|_{\Lambda_\alpha}$. The following simple observation will be useful:

Lemma 13. $(P_{k+1} + P_k)\Delta_k = \Delta_{k-1}$.

Proof. This is a simple algebraic computation:

$$\begin{aligned}
(P_{k+1} + P_k)\Delta_k &= (P_{k+1} + P_k)(P_{k+1} - P_k) = P_{k+1}P_{k+1} - P_{k+1}P_k + P_k P_{k+1} - P_k P_k \\
&= A_{2^{-(k+1)}} A_{2^{-(k+1)}} - A_{2^{-k}} A_{2^{-k}} = A_{2^{-(k+1)} + 2^{-(k+1)}} - A_{2^{-k} + 2^{-k}} \\
&= A_{2^{-k}} - A_{2^{-(k-1)}} = P_k - P_{k-1} = \Delta_{k-1}.
\end{aligned}$$

□

Lemma 14. *Suppose f is bounded. Then $V(P_k f) \le 2^{k\alpha} \sup_x |f(x)|$.*

Proof.

$$\begin{aligned}|P_kf(x)-P_kf(y)| &= \left|\int_{\mathcal{X}} p_k(x,u)f(u)du - \int_{\mathcal{X}} p_k(y,u)f(u)du\right| \\ &= \left|\int_{\mathcal{X}} [p_k(x,u)-p_k(y,u)]f(u)du\right| \\ &\le \sup_{x'}|f(x')|\,\|p_k(x,\cdot)-p_k(y,\cdot)\|_1 \\ &\le \sup_{x'}|f(x')|2^{k\alpha}D_\alpha(x,y).\end{aligned}$$

□

Proposition 10. *For $f\in\Lambda_\alpha$, $\|f\|_{\Lambda_\alpha}\lesssim\|f\|^{(1)}_{\Lambda_\alpha}$.*

Proof. Expand f in a telescopic series:

$$f-P_0f=\sum_{k=0}^{\infty}[P_{k+1}f-P_kf]=\sum_{k=0}^{\infty}\Delta_kf(x)=\sum_{k=0}^{\infty}[(P_{k+1}+P_{k+2})\Delta_{k+1}f]$$

where we have used Lemma 13. The series converges uniformly by Lemma 12.

For all $x,y\in\mathcal{X}$,

$$\begin{aligned}|P_k\Delta_kf(x)-P_k\Delta_kf(y)| &= \left|\int_{\mathcal{X}} p_k(x,u)(\Delta_kf)(u)du-\int_{\mathcal{X}} p_k(y,u)(\Delta_kf)(u)du\right| \\ &= \left|\int_{\mathcal{X}}(p_k(x,u)-p_k(y,u))(\Delta_kf)(u)du\right| \qquad (2) \\ &\le V^{(1)}(f)2^{-k\alpha}D_k(x,y).\end{aligned}$$

Similarly,

$$|P_{k+1}\Delta_kf(x)-P_{k+1}\Delta_kf(y)|\le V^{(1)}(f)2^{-k\alpha}D_{k+1}(x,y). \qquad (3)$$

For every $x,y\in\mathcal{X}$

$$\begin{aligned}f(x)-f(y)=\sum_{k=0}^{\infty}[(P_{k+1}+P_{k+2})\Delta_{k+1}f](x)-\sum_{k=0}^{\infty}[(P_{k+1}+P_{k+2})\Delta_{k+1}f](y) \\ +P_0f(x)-P_0f(y).\end{aligned}$$

From the inequalities (2) and (3) we get

$$\begin{aligned}
&\left|\sum_{k=0}^{\infty}[(P_{k+1}+P_{k+2})\Delta_{k+1}f](x)-\sum_{k=0}^{\infty}[(P_{k+1}+P_{k+2})\Delta_{k+1}f](y)\right|\\
\leq&\left|\sum_{k=0}^{\infty}(P_{k+1}\Delta_{k+1}f(x)-P_{k+1}\Delta_{k+1}f(y))\right|+\left|\sum_{k=0}^{\infty}(P_{k+2}\Delta_{k+1}f(x)-P_{k+2}\Delta_{k+1}f(y))\right|\\
\leq&\sum_{k=0}^{\infty}V^{(1)}(f)2^{-(k+1)\alpha}D_{k+1}(x,y)+\sum_{k=0}^{\infty}V^{(1)}(f)2^{-(k+1)\alpha}D_{k+2}(x,y)\\
\leq&V^{(1)}(f)(1+2^{\alpha})D_{\alpha}(x,y).
\end{aligned}$$

By Lemma 14, we also know $|P_0f(x)-P_0f(y)|\leq\sup_{x'}|f(x')|D_{\alpha}(x,y)$. Consequently, for every $x,y\in\mathcal{X}$

$$|f(x)-f(y)|\leq(V^{(1)}(f)(1+2^{\alpha})+\sup_{x'}|f(x')|)D_{\alpha}(x,y)$$

and so

$$\sup_{x\neq y}\frac{|f(x)-f(y)|}{D_{\alpha}(x,y)}\leq V^{(1)}(f)(1+2^{\alpha})+\sup_{x'}|f(x')|.$$

Therefore

$$\|f\|_{\Lambda_{\alpha}}=\sup_{x}|f(x)|+\sup_{x\neq y}\frac{|f(x)-f(y)|}{D_{\alpha}(x,y)}\leq V^{(1)}(f)(1+2^{\alpha})+2\sup_{x}|f(x)|\leq 3\|f\|_{\Lambda_{\alpha}}^{(1)}.$$

□

Putting together Corollary 4, Corollary 5 and Proposition 10, we have shown:

Theorem 2. *The norms $\|f\|_{\Lambda_{\alpha}}$, $\|f\|_{\Lambda_{\alpha}}^{(1)}$ and $\|f\|_{\Lambda_{\alpha}}^{(2)}$ are equivalent on Λ_{α}.*

5 Norm dual to Lipschitz

We now turn to the space of L^1 measures defined on $\mathcal{X}$. Since all Lipschitz functions are in L^{∞}, any such distribution can be integrated against any Lipschitz function. We will denote the action of a distribution T on a Lipschitz function f by $\langle f,T\rangle=\int_{\mathcal{X}}f(x)dT(x)$. The dual norm to the Lipschitz space Λ_{α} is defined as:

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

The space Λ_{α}^{*} is the space of all L^1 distributions T equipped with the norm $\|T\|_{\Lambda_{\alpha}^{*}}$. In Section 6, we will give a well-known interpretation of the dual norm of the difference of two probability measures.

Our goal in this section is to define two other norms on Λ_α^* and prove their equivalence to $\|T\|_{\Lambda_\alpha^*}$. First, we define

$$W^{(1)}(T) = \sum_{k\geq 0} 2^{-k\alpha} \left\| \Delta_k^* T \right\|_1$$

and

$$W^{(2)}(T) = \sum_{k\geq 0} 2^{-k\alpha} \left\| d_k^* T \right\|_1$$

where

$$d_k = P_k - P_0.$$

Now we define the equivalent norms. The first is defined by

$$\|T\|_{\Lambda_\alpha^*}^{(1)} = \|P_0^* T\|_1 + W^{(1)}(T)$$

and the second is defined by

$$\|T\|_{\Lambda_\alpha^*}^{(2)} = \|P_0^* T\|_1 + W^{(2)}(T).$$

We show that all three norms $\|T\|_{\Lambda_\alpha^*}$, $\|T\|_{\Lambda_\alpha^*}^{(1)}$ and $\|T\|_{\Lambda_\alpha^*}^{(2)}$ are equivalent on Λ_α^*.

Proposition 11. *The seminorms $W^{(1)}(T)$ and $W^{(2)}(T)$ are equivalent on Λ_α^*.*

Proof. We first show $W^{(1)}(T) \leq 2W^{(2)}(T)$.

$$\begin{aligned} W^{(1)}(T) &= \sum_{k=0}^{\infty} 2^{-k\alpha} \left\| \Delta_k^* T \right\|_1 = \sum_{k=0}^{\infty} 2^{-k\alpha} \left\| (P_k^* - P_{k+1}^*) T \right\|_1 \\ &\leq \sum_{k=0}^{\infty} 2^{-k\alpha} \left\| (P_k^* - P_0^*) T \right\|_1 + \sum_{k=0}^{\infty} 2^{-k\alpha} \left\| (P_{k+1}^* - P_0^*) T \right\|_1 \\ &\leq 2W^{(2)}(T). \end{aligned}$$

For the other direction, we write d_k^* as the telescopic sum

$$d_k^* T = P_k^* T - P_0^* T = \sum_{l=0}^{k-1} [(P_{l+1}^* T(x) - P_l^*) T] = \sum_{l=0}^{k-1} \Delta_l^* T.$$

Then $\|d_k^* T\|_1 \leq \sum_{l=0}^{k-1} \|\Delta_l^* T\|_1$, and consequently Fubini's theorem yields

$$
\begin{aligned}
W^{(2)}(T) &= \sum_{k=0}^{\infty} 2^{-k\alpha} \left\| d_k^* T \right\|_1 \leq \sum_{k=0}^{\infty} 2^{-k\alpha} \sum_{l=0}^{k-1} \left\| \Delta_l^* T \right\|_1 \\
&= \sum_{l=0}^{\infty} \left\| \Delta_l^* T \right\|_1 \sum_{k \geq l+1} 2^{-k\alpha} \\
&= \sum_{l=0}^{\infty} \left\| \Delta_l^* T \right\|_1 \frac{2^{-(l+1)\alpha}}{1-2^{-\alpha}} = \frac{2^{-\alpha}}{1-2^{-\alpha}} W^{(1)}(T)
\end{aligned}
$$

completing the proof. □

Corollary 6. *The norms $\|T\|^{(1)}_{\Lambda^*_\alpha}$ and $\|T\|^{(2)}_{\Lambda^*_\alpha}$ are equivalent.*

Next we turn to the main result of this section, namely that $\|T\|^{(1)}_{\Lambda^*_\alpha}$ and $\|T\|^{(2)}_{\Lambda^*_\alpha}$ are equivalent to $\|T\|_{\Lambda^*_\alpha}$.

Proposition 12. $\|T\|_{\Lambda^*_\alpha} \lesssim \|T\|^{(2)}_{\Lambda^*_\alpha}$.

Proof. Suppose f is any function with $\|f\|_{\Lambda_\alpha} \leq 1$. Making use of Lemma 13 and the uniform convergence of $P_k f$ to f as $k \to \infty$ we can write

$$
\begin{aligned}
f - P_0 f &= \sum_{j=0}^{\infty} \Delta_j f = \sum_{j=1}^{\infty} P_j \Delta_j f + \sum_{j=1}^{\infty} P_{j+1} \Delta_j f \\
&= \sum_{j=1}^{\infty} (P_j - P_0) \Delta_j f + \sum_{j=1}^{\infty} (P_{j+1} - P_0) \Delta_j f + 2 P_0 (I - P_1) f.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\langle f, T \rangle &= \sum_{j=1}^{\infty} \langle T, (P_j - P_0) \Delta_j f \rangle + \sum_{j=1}^{\infty} \langle T, (P_{j+1} - P_0) \Delta_j f \rangle \\
&\qquad + \langle T, (3P_0 - 2P_0 P_1) f \rangle \\
&= \sum_{j=1}^{\infty} \langle (P_j^* - P_0^*) T, \Delta_j \rangle + \sum_{j=1}^{\infty} \langle (P_{j+1}^* - P_0^*) T, \Delta_j f \rangle \\
&\qquad + \langle P_0^* T, (3I - 2P_1) f \rangle.
\end{aligned}
$$

Consequently, we have

$$
\begin{aligned}
|\langle f,T\rangle| \le \sum_{j=1}^{\infty} \left\|(P_j^* - P_0^*)T\right\|_1 \sup_x |\Delta_j f(x)| + \sum_{j=1}^{\infty} \left\|(P_{j+1}^* - P_0^*)T\right\|_1 \sup_x |\Delta_j f(x)| \\
+ \|P_0^* T\|_1 \sup_x |(3I - 2P_1) f(x)| \\
\lesssim \sum_{j=1}^{\infty} 2^{-j\alpha} \left\|d_j^* T\right\|_1 + \sum_{j=1}^{\infty} 2^{-j\alpha} \left\|d_{j+1}^* T\right\|_1 + \|P_0^* T\|_1 \sup_x |f(x)|
\end{aligned}
$$

where in the last inequality we have used the equivalence of $\|f\|_{\Lambda_\alpha}$ and $\|f\|_{\Lambda_\alpha}^{(1)}$ and the fact that $\sup_x |P_k f(x)| \lesssim \sup_x |f(x)|$ (a trivial consequence of condition (I) on the kernel). Since $\sup_x |f(x)| \le \|f\|_{\Lambda_\alpha} \le 1$, it follows immediately that

$$
|\langle f,T\rangle| \lesssim \sum_{j=1}^{\infty} 2^{-j\alpha} \left\|d_j^* T\right\|_1 + \|P_0 T\|_1 = \|T\|_{\Lambda_\alpha^*}^{(2)}.
$$

Now take the supremum over all f with $\|f\|_{\Lambda_\alpha} \le 1$ to reach the desired conclusion. □

Proposition 13. $\|T\|_{\Lambda_\alpha^*}^{(2)} \lesssim \|T\|_{\Lambda_\alpha^*}$.

Proof. Define the function f by

$$
\begin{aligned}
f(x) &= \sum_{k=1}^{\infty} 2^{-k\alpha} (P_k - P_0) \operatorname{sgn}[(P_k^* - P_0^*)T](x) + P_0[\operatorname{sgn}(P_0^* T)](x) \\
&= \sum_{k=1}^{\infty} 2^{-k\alpha} P_k \operatorname{sgn}[(P_k^* - P_0^*)T](x) + P_0 F(x)
\end{aligned}
$$

where

$$
F(x) = \operatorname{sgn}(P_0^* T)(x) - \sum_{k=1}^{\infty} 2^{-k\alpha} \operatorname{sgn}[(P_k^* - P_0^*)T](x).
$$

Since

$$
\sup_x |F(x)| \le 1 + \sum_{k=1}^{\infty} 2^{-k\alpha} \le 1 + \frac{2^{-\alpha}}{1 + 2^{-\alpha}}
$$

therefore, by Lemma 14

$$
|P_0 F(x) - P_0 F(y)| \le \left(1 + \frac{2^{-\alpha}}{1 + 2^{-\alpha}}\right) D_\alpha(x,y)
$$

for all $x, y \in \mathcal{X}$. Furthermore, letting $h_k = \operatorname{sgn}[(P_k^* - P_0^*)T]$, Lemma 14 also implies that

$|P_k h_k(x) - P_k h_k(y)| \leq D_k(x,y)$, and consequently

$$\begin{aligned}
&\left| \sum_{k=1}^{\infty} 2^{-k\alpha} P_k \operatorname{sgn}[(P_k^* - P_0^*)T](x) - \sum_{k=1}^{\infty} 2^{-k\alpha} P_k \operatorname{sgn}[(P_k^* - P_0^*)T](y) \right| \\
&\leq \sum_{k=1}^{\infty} 2^{-k\alpha} |P_k h_k(x) - P_k h_k(y)| \\
&\leq \sum_{k=1}^{\infty} 2^{-k\alpha} D_k(x,y) \leq D_\alpha(x,y).
\end{aligned}$$

We also have the estimate

$$\|f\|_\infty \lesssim \sum_{k=1}^{\infty} 2^{-k\alpha} + 1 \leq 1 + \frac{2^{-\alpha+1}}{1+2^{-\alpha}}.$$

It follows that $\|f\|_{\Lambda_\alpha} \leq C(\alpha)$, where $C(\alpha)$ is a constant depending only on α (in particular, not on T). By the definition of f, we see

$$\begin{aligned}
\langle f, T\rangle &= \sum_{k=1}^{\infty} 2^{-k\alpha} \langle (P_k - P_0) \operatorname{sgn}[(P_k^* - P_0^*)T], T\rangle + \langle P_0[\operatorname{sgn}(P_0^* T)], T\rangle \\
&= \sum_{k=1}^{\infty} 2^{-k\alpha} \langle \operatorname{sgn}[(P_k^* - P_0^*)T], (P_k^* - P_0^*)T\rangle + \langle \operatorname{sgn}(P_0^* T), P_0^* T\rangle \\
&= \sum_{k=1}^{\infty} 2^{-k\alpha} \left\|(P_k^* - P_0^*)T\right\|_1 + \left\|P_0^* T\right\|_1.
\end{aligned}$$

So

$$\|T\|_{\Lambda_\alpha^*} \geq C(\alpha)^{-1} \langle f, T\rangle \simeq \sum_{k=1}^{\infty} 2^{-k\alpha} \left\|(P_k^* - P_0^*)T\right\|_1 + \left\|P_0^* f\right\|_1$$

which is the desired result.

□

Putting together Corollary 6, Proposition 12 and Proposition 13 we have shown:

Theorem 3. *The norms $\|T\|_{\Lambda_\alpha^*}$, $\|T\|_{\Lambda_\alpha^*}^{(1)}$ and $\|T\|_{\Lambda_\alpha^*}^{(2)}$ are equivalent on Λ_α^*.*

6 Application to Earth Mover's Distance

The dual norm $\|T\|_{\Lambda_\alpha^*}$ has a natural interpretation when the distribution T is the difference of two probability measures μ and ν. We will explain this in a more general setting. Suppose Ω is any metric/measure space with metric ρ. A measure π on $\Omega \times \Omega$ satisfies the equality-of-marginals condition with respect to μ and ν if

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

for all measurable sets $E \subset \Omega$. The Kantorovich-Rubinstein Theorem states that if Ω is separable, and if the expected distance under μ and ν from any point is finite, we have the dual relationship

$$\sup_{g:|g(x)-g(y)|\leq\rho(x,y)} \left\{ \int_\Omega g d\mu - \int_\Omega g d\nu \right\} = \inf_{\pi\text{: (EM) holds}} \int_{\Omega\times\Omega} \rho(x,y) d\pi(x,y). \tag{KR}$$

For a proof, see [9].

The quantity on the right of (KR) is known as the Earth Mover's Distance between μ and ν, denoted $\text{EMD}(\mu,\nu)$. It has the following interpretation. We view each measure π satisfying the equality-of-marginals condition (EM) with respect to μ and ν as a transport between the measures ν and μ; that is, for any two measurable sets $A, B \subset \Omega$, $\pi(A, B)$ is interpreted as the amount of mass moved from set A to set B. The equality-of-marginals condition (EM) guarantees that the transport rearranges the mass distribution described by ν to end up with the distribution described by μ. If $\rho(x, y)$ is the cost-per-mass of moving mass from location x to location y, then $\text{EMD}(\mu,\nu)$ 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 quantity on the left of (KR) is equal to the norm of $T = \mu - \nu$ in the space dual to Lipschitz functions, except we do not require that the functions T is integrated against lie in L^∞. However, when the diameter of the space is finite, as for the distances D_α we have defined, and $\int d\mu = \int d\nu$, then the two definitions are easily seen to be equal, and the norm $\|\mu - \nu\|_{\Lambda_\alpha^*}$ is equal to the left side of (KR).

Due to the way it exploits the geometry of the metric space on which probability distributions are defined, EMD has many desirable properties that make it a natural choice of metric for many problems in machine learning [15, 18, 20, 21, 24]. We now describe one such property, which helps explain its robustness.

Suppose p_1 is a probability distribution on a space Ω with metric $\rho(x, y)$ and measure μ such that the Kantorovich-Rubinstein Theorem holds; for instance, Ω might be separable. Let $h : \Omega \to \Omega$ be a 1-1, absolutely continuous (with respect to μ) transformation satisfying

$$\rho(x, h(x)) \leq \epsilon \tag{4}$$

for all $x \in \Omega$. Let ν be the measure induced by the change-of-variable h, that is, $\nu(S) = \mu(h(S))$ for measurable subsets $S \subset \Omega$; and let $\frac{d\nu}{d\mu}$ denote the Radon-Nikodym derivative of ν with respect to μ. Then we define the distribution

$$p_2(x) = p_1(h(x))\frac{d\nu}{d\mu}(x)$$

obtained from p_1 by the change-of-variable h. We think of p_2 as a perturbation of p_1. In L^1, for example, the distance between p_1 and p_2 could be quite large; however, we now show that $\mathrm{EMD}(p_1, p_2)$ is no greater than the size of the perturbation itself.

Theorem 4. *Under the assumptions described above,* $\mathrm{EMD}(p_1, p_2) \leq \epsilon$.

Proof. We use that

$$\mathrm{EMD}(p_1, p_2) = \sup \left\{ \int_\Omega f(x)(p_1(x) - p_2(x)) d\mu(x) : |f(x) - f(y)| \leq \rho(x, y) \right\}.$$

Take any f with $|f(x) - f(y)| \leq \rho(x, y)$ for all x and y, and observe that

$$\begin{aligned} \int_\Omega f(x) p_2(x) d\mu(x) &= \int_\Omega f(x) p_1(h(x)) \frac{d\nu}{d\mu}(x) d\mu(x) \\ &= \int_\Omega f(x) p_1(h(x)) d\nu(x) = \int_\Omega f(x) p_1(h(x)) d\mu(h(x)) \\ &= \int_\Omega f(h^{-1}(y)) p_1(y) d\mu(y) \end{aligned}$$

and consequently

$$\int_\Omega f(x)(p_1(x) - p_2(x)) d\mu(x) = \int_\Omega p_1(x)(f(x) - f(h^{-1}(x))) d\mu(x).$$

Now by assumption (4) on h and the fact that h is 1-1, we have $\rho(x, h^{-1}(x)) \leq \epsilon$; hence, since f has Lipschitz constant 1, we have

$$|f(x) - f(h^{-1}(x))| \leq \rho(x, h^{-1}(x)) \leq \epsilon$$

and consequently

$$\int_\Omega f(x)(p_1(x) - p_2(x)) d\mu(x) = \int_\Omega p_1(x)(f(x) - f(h^{-1}(x))) d\mu(x) \leq \epsilon \int_\Omega p_1(x) d\mu(x) = \epsilon$$

since p_1 is a probability distribution; then taking the supremum over all Lipschitz f gives

$$\mathrm{EMD}(p_1, p_2) \leq \epsilon$$

as desired. □

In order to apply this theory to the setting of this paper, we need to check that the Kantorovich-Rubinstein Theorem applies when we equip our space $\mathcal{X}$ with the metric $D_\alpha(x, y)$. As noted, a sufficient condition is to check that the resulting metric space is separable. We can prove separability under the additional assumption that $\mathcal{X}$ is sigma-finite.

Lemma 15. *Under the metric $D_\alpha(x,y)$, balls in $\mathcal{X}$ of positive radius have positive measure.*

Proof. We deduce this from condition (G) as follows. Suppose that there were some ball $B(x,r)$, $r>0$, with measure zero. Then

$$1 = \int_{\mathcal{X}} p_k(x,y)dy \le \int_{\mathcal{X}} |p_k(x,y)|dy = \int_{B(x,r)^c} |p_k(x,y)|dy$$

and consequently

$$r \le \int_{B(x,r)^c} |p_k(x,y)| D_\alpha(x,y)dy \le C2^{-k\alpha}.$$

Since $r>0$, taking $k\to\infty$ yields a contradiction. □

Proposition 14. *Suppose that the measure space $\mathcal{X}$ is sigma-finite. Then the metric $D_\alpha(x,y)$ turns $\mathcal{X}$ into a separable metric space; in particular, the Kantorovich-Rubinstein Theorem holds on $\mathcal{X}$.*

Proof. By sigma-finiteness, we can write $\mathcal{X}$ as a countable union of finite measure sets. Without loss of generality, we can therefore assume that $\mathcal{X}$ itself has finite measure. Use Zorn's Lemma to find a maximal collection of points $\{x_i\}_{i\in\mathcal{I}}$ so that $D_\alpha(x_i,x_j)\ge 1/n$, where $\mathcal{I}$ is some index set. By maximality, every point in $\mathcal{X}$ is within $1/n$ of one of the points x_i; so we will be done if we can show that $\mathcal{I}$ is necessarily countable. To see this, observe that the balls $B(x_i,1/2n)$ are pairwise disjoint and have positive measure. Since $\mathcal{X}$ has finite measure, there can only be finitely many balls whose measure lies in the interval $(2^{-k-1},2^{-k}]$, for each $k\in\mathbb{Z}$. Since the measure of each ball must lie in one such interval, and there are countably many intervals, there are only countably many balls, and the proof is complete. □

In our setting of the space $\mathcal{X}$ with the semigroup $a_t(x,y)$, the formulas for the norm $\|T\|_{\Lambda_\alpha^*}$ from Section 5 provides an approximation to Earth Mover's Distance. From Theorem 3, and the Kantorovich-Rubinstein Theorem, the Earth Mover's Distance between two probability measures μ and ν is equivalent to the expressions

$$\|\mu-\nu\|_{\Lambda_\alpha^*}^{(1)} = \|P_0^*(\mu-\nu)\|_1 + \sum_{k\ge 0} 2^{-k\alpha}\,\|\Delta_k^*(\mu-\nu)\|_1$$

and

$$\|\mu-\nu\|_{\Lambda_\alpha^*}^{(2)} = \|P_0^*(\mu-\nu)\|_1 + \sum_{k\ge 0} 2^{-k\alpha}\,\|d_k^*(\mu-\nu)\|_1\,.$$

In machine learning applications, these formulas can often be computed fast, and thus provide a fast approximation to Earth Mover's Distance. We only give a sketch of how this works, waving our hands regarding the issues that arise when using discrete data. We take $\mathcal{X}$ to be a collection of n data points, and the operators P_k to be dyadic powers of a Markov matrix M on the data, as in the theory of diffusion maps [5].

We assume that the one-step Markov matrix on the data is at scale $t = 1/n$. Therefore, we need only take $N = \lfloor \log_2(n) \rfloor$ dyadic powers of M before reaching scale 1. So to approximate Earth Mover's Distance (with respect to the ground distance $D_\alpha(x, y)$) between two probability vectors μ and ν on the data, we propose the heuristic formula

$$\|(M^n)^*(\mu - \nu)\|_1 + \sum_{k=0}^{N} 2^{(k-N)\alpha} \|(M^{2^{k+1}} - M^{2^k})^*(\mu - \nu)\|_1$$

or the formula

$$\|(M^n)^*(\mu - \nu)\|_1 + \sum_{k=0}^{N} 2^{(k-N)\alpha} \|(M^{2^k} - M^n)^*(\mu - \nu)\|_1.$$

If all dyadic powers of the matrix M can be applied rapidly, say in time $\mathcal{O}(n \log^k n)$, then these formulas can be evaluated at the same cost. This is not an unreasonable supposition; for instance, see the papers [6] and [7]. Note that a simplifying consideration in the case of diffusion maps is that the Markov matrices M considered there are similar to a symmetric positive definite matrix (with maximum eigenvalue equal to 1), and so any algorithm that permits fast application of all powers of such matrices will enable a fast approximation of EMD in our setting.

In more specialized cases similar formulas have been shown to approximate EMD as well. The work that most closely resembles this one is wavelet EMD [19]. Here, wavelets are used in place of the operators Δ_k and d_k. The applicability of this method limited to $\mathbb{R}^n$, where the ground distance is a snowflake of the Euclidean metric.

The reader can also refer to the papers by Charikar [3] and Indyk and Thaper [13]. Though the particulars are quite different than those in the present work, the general spirit is the same; EMD can be approximated by a weighted sum of L^1 norms of difference operators at different scales, whatever the notion of "scale" might mean for the geometry under consideration.

7 Mixed Lipschitz functions on product spaces

We now consider the setting where we have a product of spaces, each equipped with its own semigroup satisfying (S), (C), (I) and (R) so that the theory developed so far can be applied. For simplicity, we will consider only two spaces, which we will denote $\mathcal{X}$ and $\mathcal{Y}$, each with a semigroup A_s and B_t with kernels $a_t(x, x')$ and $b(y, y')$, respectively. All the results and their proofs can be extended to arbitrarily many semigroups. We define the dyadic discretizations for times between 0 and 1

$$P_k = A_{2^{-k}},\ p_k(x, x') = a_{2^{-k}}(x, x'),\ k \geq 0$$

and

$$Q_l = B_{2^{-l}},\ q_l(y, y') = b_{2^{-k}}(y, y'),\ l \geq 0$$

and the distances

$$D_{\mathcal{X},k}(x,x') = \left\|p_k(x,\cdot) - p_k(x',\cdot)\right\|_1$$

and

$$D_{\mathcal{Y},l}(y,y') = \left\|q_l(y,\cdot) - q_l(y',\cdot)\right\|_1 .$$

For $0 < \alpha, \beta < 1$, such that the geometric condition (G) holds for P_k with respect to α and (G) holds for Q_l with respect to β, we define metrics on $\mathcal{X}$ and $\mathcal{Y}$ by

$$D_{\mathcal{X},\alpha}(x,x') = \sum_{k\geq 0} 2^{-k\alpha} D_{\mathcal{X},k}(x,x')$$

and

$$D_{\mathcal{Y},\beta}(y,y') = \sum_{l\geq 0} 2^{-l\beta} D_{\mathcal{Y},k}(y,y').$$

For brevity, we will let $D_{\mathcal{X}} = D_{\mathcal{X},\alpha}$ and $D_{\mathcal{Y}} = D_{\mathcal{Y},\beta}$.

We will define a regularity norm and its dual on the product space $\mathcal{X} \times \mathcal{Y}$. We first define the following quantities:

$$V_{\mathcal{X}}(f) = \sup_{y, x\neq x'} \frac{f(x,y) - f(x',y)}{D_{\mathcal{X}}(x,x')},$$

$$V_{\mathcal{Y}}(f) = \sup_{x, y\neq y'} \frac{f(x,y) - f(x,y')}{D_{\mathcal{Y}}(y,y')},$$

and

$$M(f) = \sup_{x\neq x', y\neq y'} \frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{D_{\mathcal{X}}(x,x') D_{\mathcal{Y}}(y,y')}.$$

We then define the norm

$$\|f\|_{\Lambda_{\alpha,\beta}} \equiv M(f) + V_{\mathcal{X}}(f) + V_{\mathcal{Y}}(f) + \sup_x |f(x)|$$

and denote by $\Lambda_{\alpha,\beta}$ the space of all functions f where $\|f\|_{\Lambda_{\alpha,\beta}} < \infty$.

Lemma 16. *Taking*

$$\tilde{V_{\mathcal{X}}}(f) = \sup_{y, x\neq x'} \frac{(Q_0 f)(x,y) - (Q_0 f)(x',y)}{D_{\mathcal{X}}(x,x')},$$

$$\tilde{V_{\mathcal{Y}}}(f) = \sup_{x, y\neq y'} \frac{(P_0 f)(x,y) - (P_0 f)(x,y')}{D_{\mathcal{Y}}(y,y')}$$

in place of, respectively, the seminorms $V_{\mathcal{X}}$ and $V_{\mathcal{Y}}$ in the definition of $\|f\|_{\Lambda_{\alpha,\beta}}$ yields an equivalent norm.

Proof. From condition (I), it is immediate that $\tilde{V}_{\mathcal{X}}(f) \lesssim V_{\mathcal{X}}(f)$ and $\tilde{V}_{\mathcal{Y}}(f) \leq V_{\mathcal{Y}}(f)$. For the other inequality, we can control $V_{\mathcal{X}}(f)$ by $\tilde{V}_{\mathcal{X}}(f)$ and $M(f)$, and control $V_{\mathcal{Y}}(f)$ by $\tilde{V}_{\mathcal{Y}}(f)$ and $M(f)$. To see this, observe that

$$\begin{aligned}
&|(Q_0f)(x,y) - (Q_0f)(x',y) - f(x,y) + f(x',y)| \\
=&\left| \int_{\mathcal{X}} q_0(y,y')[f(x,y') - f(x',y') - f(x,y) + f(x',y)]dy' \right| \\
\leq& CD_{\mathcal{X}}(x,x') \operatorname{diam}(\mathcal{Y}) M(f) \\
\lesssim& M(f) D_{\mathcal{X}}(x,x')
\end{aligned}$$

where we have used condition (I) in the second-to-last inequality. Consequently, $V_{\mathcal{X}}(f) \lesssim \tilde{V}_{\mathcal{X}}(f) + M(f)$; similarly, $V_{\mathcal{Y}}(f) \lesssim \tilde{V}_{\mathcal{Y}}(f) + M(f)$. It follows that replacing $V_{\mathcal{X}}(f)$ and $V_{\mathcal{Y}}(f)$ by, respectively, $\tilde{V}_{\mathcal{X}}(f)$ and $\tilde{V}_{\mathcal{Y}}(f)$ in the definition of $\|f\|_{\Lambda_{\alpha,\beta}}$ yields an equivalent norm. □

Of course, other minor variations in the definition of $\|f\|_{\Lambda_{\alpha,\beta}}$ yielding equivalent norms are also possible. However, as in the case of a single space our primary goal is to give simpler characterizations of the norm $\|f\|_{\Lambda_{\alpha,\beta}}$ involving the changes in the function's averages across scales. In Section 8, we will use these to give simple characterizations of the norm on the space dual to $\Lambda_{\alpha,\beta}$.

We define the difference operators

$$\Delta_{P,k} = P_{k+1} - P_k, \quad \Delta_{Q,l} = Q_{l+1} - Q_l.$$

as well as

$$\delta_{P,k} = I - P_k, \quad \delta_{Q,l} = I - Q_l.$$

We then define

$$V_{\mathcal{X}}^{(1)}(f) = \sup_{k\geq 0} \sup_{x,y} 2^{k\alpha} |\Delta_{P,k} f(x,y)|, \quad V_{\mathcal{Y}}^{(1)}(f) = \sup_{l\geq 0} \sup_{x,y} 2^{l\beta} |\Delta_{Q,l} f(x,y)|,$$

and

$$M^{(1)}(f) = \sup_{k\geq 0, l\geq 0} \sup_{x,y} 2^{k\alpha + l\beta} |\Delta_{P,k} \Delta_{Q,l} f(x,y)|.$$

Similarly, define

$$V_{\mathcal{X}}^{(2)}(f) = \sup_{k\geq 0} \sup_{x,y} 2^{k\alpha} |\delta_{P,k} f(x,y)|, \quad V_{\mathcal{Y}}^{(2)}(f) = \sup_{l\geq 0} \sup_{x,y} 2^{l\beta} |\delta_{Q,l} f(x,y)|,$$

and

$$M^{(2)}(f) = \sup_{k\geq 0, l\geq 0} \sup_{x,y} 2^{k\alpha + l\beta} \; |delta_{P,k} \delta_{Q,l} f(x,y)|.$$

We can now define the equivalent regularity norms by

$$\|f\|^{(1)}_{\Lambda_{\alpha,\beta}} = M^{(1)}(f) + V^{(1)}_{\mathcal{X}}(f) + V^{(1)}_{\mathcal{Y}}(f) + \sup_{x,y} |f(x,y)|$$

and

$$\|f\|^{(2)}_{\Lambda_{\alpha,\beta}} = M^{(2)}(f) + V^{(2)}_{\mathcal{X}}(f) + V^{(2)}_{\mathcal{Y}}(f) + \sup_{x,y} |f(x,y)|.$$

We first show that $\|f\|_{\Lambda_{\alpha,\beta}}$ controls $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}}$. It will follow that on $\Lambda_{\alpha,\beta}$ we have uniform convergence of the semigroups and their products to the identity.

Proposition 15. *For any function f, $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}} \lesssim \|f\|_{\Lambda_{\alpha,\beta}}$.*

Proof. Showing that $V^{(2)}_{\mathcal{X}}(f)$ and $V^{(2)}_{\mathcal{Y}}(f)$ are controlled by, respectively, $V_{\mathcal{X}}(f)$ and $V_{\mathcal{Y}}(f)$ is an immediate consequence of the one-dimensional result, Lemma 8. To show $M^{(2)}(f) \lesssim M(f)$, observe that Lemma 8 also gives

$$\begin{aligned}|2^{k\alpha}\delta_{P,k}2^{l\beta}\delta_{Q,l}f(x,y)| &\lesssim \sup_{x\neq x'} \frac{2^{l\beta}\delta_{Q,l}f(x,y) - 2^{l\beta}\delta_{Q,l}f(x',y)}{D_{\mathcal{X}}(x,x')} \\ &= \sup_{x\neq x'} \frac{2^{l\beta}\delta_{Q,l}[f(x,\cdot) - f(x',\cdot)](y)}{D_{\mathcal{X}}(x,x')}.\end{aligned}$$

Now apply Lemma 8 again to the function $y \mapsto f(x,y) - f(x',y)$ to obtain the bound

$$|2^{l\beta}\delta_{Q,l}[f(x,\cdot) - f(x',\cdot)](y)| \lesssim \sup_{y\neq y'} \frac{f(x,y) - f(x',y) - f(x,y') + f(x',y')}{D_{\mathcal{Y}}(y,y')}.$$

The result follows. □

It is easy to see that if $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}} < \infty$, then

$$\lim_{k\to\infty, l\to\infty} P_k Q_l f = f$$

uniformly, where the limits can be taken in either order or simultaneously. Since $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}} \lesssim \|f\|_{\Lambda_{\alpha,\beta}}$, the same convergence applies for any $f \in \Lambda_{\alpha,\beta}$.

We will next show that $\|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$ and $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}}$ are equivalent, and then that $\|f\|_{\Lambda_{\alpha,\beta}} \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$. To that end:

Lemma 17. *The seminorms $V^{(1)}_{\mathcal{X}}(f)$ and $V^{(2)}_{\mathcal{X}}(f)$ are equivalent, as are the seminorms $V^{(1)}_{\mathcal{Y}}(f)$ and $V^{(2)}_{\mathcal{Y}}(f)$.*

Proof. This follows immediately from Proposition 9 for a single semigroup. □

Lemma 18. *The seminorms $M^{(1)}(f)$ and $M^{(2)}(f)$ are equivalent.*

Proof. From Proposition 9, we have

$$\begin{aligned}|2^{k\alpha}\Delta_{P,k}2^{l\beta}\Delta_{Q,l}f(x,y)| &\lesssim \sup_{x'}\sup_{k'\geq 0} 2^{k'\alpha}|\delta_{P,k'}2^{l\beta}\Delta_{Q,l}f(x',y)| \\ &= \sup_{x'}\sup_{k'\geq 0} 2^{l\beta}|\Delta_{Q,l}2^{k'\alpha}\delta_{P,k'}f(x',y)| \\ &\lesssim \sup_{x'}\sup_{k'\geq 0}\sup_{y'}\sup_{l'} 2^{l'\beta}|\delta_{Q,l}2^{k'\alpha}\delta_{P,k'}f(x',y)|\end{aligned}$$

which proves $M^{(2)}(f) \lesssim M^{(1)}(f)$. The other direction is proved similarly. □

Combining Lemmas 17 and 18, we get:

Proposition 16. *The norms $\|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$ and $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}}$ are equivalent.*

To finish proving that all three norms are equivalent, we will show that $\|f\|_{\Lambda_{\alpha,\beta}} \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$.

Proposition 17. *For all $f \in \Lambda_{\alpha,\beta}$, $\|f\|_{\Lambda_{\alpha,\beta}} \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$.*

Proof. First, it is trivial to deduce $V_{\mathcal{X}}(f) \lesssim V^{(1)}_{\mathcal{X}}(f) + \sup_{x,y}|f(x,y)|$ and $V_{\mathcal{Y}}(f) \lesssim V^{(1)}_{\mathcal{Y}}(f) + \sup_{x,y}|f(x,y)|$ from Proposition 10. Therefore, it remains to show $M(f) \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$.

Fix any $y, y' \in \mathcal{Y}$ and define

$$g(x) = \frac{f(x,y) - f(x,y')}{D_{\mathcal{Y}}(y,y')}.$$

From Proposition 10 again, we have that for all $x \neq x'$,

$$\frac{f(x,y) - f(x,y') - f(x',y) + f(x',y')}{D_{\mathcal{X}}(x,x')D_{\mathcal{Y}}(y,y')} \lesssim \sup_{k\geq 0}\sup_{x''} 2^{k\alpha}|\Delta_{P,k}g(x'')| + \sup_{x''}|g(x'')|.$$

The supremum of g is bounded by

$$\sup_{x''}|g(x'')| \leq V_{\mathcal{Y}}(f) \lesssim V^{(1)}_{\mathcal{Y}}(f).$$

Furthermore, we have

$$\begin{aligned}2^{k\alpha}|\Delta_{P,k}g(x'')| &= 2^{k\alpha}\frac{|\Delta_{P,k}f(x'',y) - \Delta_{P,k}f(x'',y')|}{D_{\mathcal{Y}}(y,y')} \\ &\lesssim 2^{k\alpha}\sup_{l\geq 0}\sup_{y''} 2^{l\beta}|\Delta_{Q,l}\Delta_{P,k}f(x'',y'')| \\ &\leq M^{(1)}(f).\end{aligned}$$

It follows that $M(f) \lesssim M^{(1)}(f) + V^{(1)}_{\mathcal{Y}}(f) \leq \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$, which completes the proof. □

Combining Proposition 15, Proposition 16, and Proposition 17, we have shown

Theorem 5. *The norms $\|f\|_{\Lambda_{\alpha,\beta}}$, $\|f\|^{(1)}_{\Lambda_{\alpha,\beta}}$ and $\|f\|^{(2)}_{\Lambda_{\alpha,\beta}}$ are equivalent on $\Lambda_{\alpha,\beta}$.*

7.1 Approximating mixed Lipschitz functions

The reader will recall Lemma 8, which states that for a Lipschitz function f on a single space $\mathcal{X}$,

$$\sup_x |f(x) - P_L f(x)| \leq CV(f)2^{-L\alpha}$$

for some constant $C > 0$ depending only on the semigroup. Another way of saying this is that given any $\epsilon > 0$, if we take $L \geq \log_2(1/\epsilon)$ then $\sup_x |f(x) - P_L f(x)| \lesssim \epsilon^\alpha$. In other words, Lipschitz functions f are well-approximated by their averages under the semigroup. We will derive a similar result for mixed Lipschitz functions. For any integer L, define the operator $\mathcal{P}_L$ by

$$\mathcal{P}_L f = \sum_{k,l:k+l\leq L} \Delta_{P,k}\Delta_{Q,l} f + \delta_{Q,L} P_0 f + \delta_{P,L} Q_0 f + P_0 Q_0 f.$$

We then have the following result:

Proposition 18. *Fix any ϵ and let $L \geq \log_2(1/\epsilon)$. Then*

$$\sup_{x,y} |\mathcal{P}_L f(x,y) - f(x,y)| \leq C \|f\|_{\Lambda_{\alpha,\beta}} \begin{cases} \epsilon^\alpha \log_2(1/\epsilon), & \text{if } \alpha = \beta \\ \epsilon^{\min(\alpha,\beta)}, & \text{if } \alpha \neq \beta \end{cases}$$

where $C > 0$ is some constant depending only on the semigroup in both cases.

Proof. We can write f as

$$\begin{aligned} f &= \sum_{k,l\geq 0} \Delta_{P,k}\Delta_{Q,l} f + \sum_{l\geq 0} \Delta_{Q,l} P_0 f + \sum_{k\geq 0} \Delta_{P,k} Q_0 f + P_0 Q_0 f \\ &= \sum_{k,l\geq 0} \Delta_{P,k}\Delta_{Q,l} f + \sum_{l\geq L} \Delta_{Q,l} P_0 f + \delta_{Q,L} P_0 f + \sum_{k\geq L} \Delta_{P,k} Q_0 f + \delta_{P,L} Q_0 f + P_0 Q_0 f. \end{aligned}$$

Therefore,

$$f - \mathcal{P}_L f = \sum_{k,l:k+l>L} \Delta_{P,k}\Delta_{Q,l} f + \sum_{l\geq L} \Delta_{Q,l} P_0 f + \sum_{k\geq L} \Delta_{P,k} Q_0 f.$$

We have

$$\left| \sum_{l\geq L} \Delta_{Q,l} P_0 f \right| \leq \sum_{l\geq L} V_{\mathcal{Y}}^{(1)}(f) 2^{-l\beta} \leq \frac{1}{1-2^{-\beta}} V_{\mathcal{Y}}^{(1)}(f) 2^{-L\beta} \lesssim V_{\mathcal{Y}}^{(1)}(f)\epsilon^\beta$$

and similarly

$$\left| \sum_{l\geq L} \Delta_{Q,l} P_0 f \right| \leq \frac{1}{1-2^{-\alpha}} V_{\mathcal{X}}^{(1)}(f) 2^{-L\alpha} \lesssim V_{\mathcal{X}}^{(1)}(f)\epsilon^\alpha.$$

Finally, we control the mixed difference term:

$$\begin{aligned}\Bigg|\sum_{k,l:k+l>L}\Delta_{P,k}\Delta_{Q,l}f\Bigg| &= \sum_{k=0}^{L+1}\sum_{l=L-k}^{\infty}|\Delta_{P,k}\Delta_{Q,l}f| + \sum_{k=L+2}^{\infty}\sum_{l=0}^{\infty}|\Delta_{P,k}\Delta_{Q,l}f| \\ &\le M^{(1)}(f)\sum_{k=0}^{L+1}2^{-k\alpha}\sum_{l=L-k}^{\infty}2^{-l\beta} + M^{(1)}(f)\sum_{k=L+2}^{\infty}2^{-k\alpha}\sum_{l=0}^{\infty}2^{-l\beta} \\ &\le \frac{M^{(1)}(f)}{1-2^{-\beta}}2^{-L\beta}\sum_{k=0}^{L+1}2^{-k(\alpha-\beta)} + \frac{M^{(1)}(f)}{1-2^{-\beta}}\frac{1}{1-2^{-\alpha}}2^{-(L+2)\alpha}.\end{aligned}$$

Now, if $\alpha=\beta$, then $\sum_{k=0}^{L+1}2^{-k(\alpha-\beta)}=L+2$, and

$$\Bigg|\sum_{k,l:k+l>L}\Delta_{P,k}\Delta_{Q,l}f\Bigg| \lesssim M^{(1)}(f)L2^{-L\alpha} \le M^{(1)}(f)\epsilon^{\alpha}\log_2(1/\epsilon)$$

from which the estimate $\sup_{x,y}|\mathcal{P}_Lf(x,y)-f(x,y)| \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}\epsilon^{\alpha}\log_2(1/\epsilon)$ follows immediately.

If $\alpha<\beta$, then $\sum_{k=0}^{L+1}2^{-k(\alpha-\beta)}\simeq 2^{L(\beta-\alpha)}$, and so

$$\Bigg|\sum_{k,l:k+l>L}\Delta_{P,k}\Delta_{Q,l}f\Bigg| \lesssim M^{(1)}(f)(2^{-L\beta}2^{L(\beta-\alpha)}+2^{-L\alpha}) \lesssim M^{(1)}(f)2^{-L\alpha} \le M^{(1)}(f)\epsilon^{\alpha}$$

from which the estimate $\sup_{x,y}|\mathcal{P}_Lf(x,y)-f(x,y)| \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}\epsilon^{\alpha}$ follows.

Finally, if $\alpha>\beta$, then $\sum_{k=0}^{L+1}2^{-k(\alpha-\beta)}\simeq 1$, and so

$$\Bigg|\sum_{k,l:k+l>L}\Delta_{P,k}\Delta_{Q,l}f\Bigg| \lesssim M^{(1)}(f)(2^{-L\beta}+2^{-L\alpha}) \lesssim M^{(1)}(f)2^{-L\beta} \le M^{(1)}(f)\epsilon^{\beta}$$

from which the estimate $\sup_{x,y}|\mathcal{P}_Lf(x,y)-f(x,y)| \lesssim \|f\|^{(1)}_{\Lambda_{\alpha,\beta}}\epsilon^{\beta}$ also follows. Since $\|f\|^{(1)}_{\Lambda_{\alpha,\beta}} \lesssim \|f\|_{\Lambda_{\alpha,\beta}}$ by Theorem 5, we are done. □

8 Norm dual to mixed Lipschitz

We now consider the space $\Lambda^*_{\alpha,\beta}$ of L^1 distributions dual to the space $\Lambda_{\alpha,\beta}$ of mixed Lipschitz functions. We will derive two simpler norms that are equivalent to the canonical dual norm on $\Lambda^*_{\alpha,\beta}$, as we did for the case of a single semigroup.

We define the norm of a distribution T in $\Lambda^*_{\alpha,\beta}$ by

$$\|T\|_{\Lambda^*_{\alpha,\beta}} = \sup_{\|f\|_{\Lambda_{\alpha,\beta}}\le 1}\langle f,T\rangle$$

Before defining the equivalent norms, we introduce some notation. Define

$$d_{P,k} = P_k - P_0,\ d_{Q,l} = Q_k - Q_0$$

and

$$W^{(1)}_{\mathcal{X}}(T) = \sum_{k\geq 0} 2^{-k\alpha} \left\|\Delta^*_{P,k} Q^*_0 f\right\|_1,\ W^{(1)}_{\mathcal{Y}}(T) = \sum_{l\geq 0} 2^{-l\beta} \left\|\Delta^*_{Q,l} P^*_0 f\right\|_1$$

and

$$W^{(2)}_{\mathcal{X}}(T) = \sum_{k\geq 0} 2^{-k\alpha} \left\|d^*_{P,k} Q^*_0 T\right\|_1,\ W^{(2)}_{\mathcal{Y}}(T) = \sum_{l\geq 0} 2^{-l\beta} \left\|d^*_{Q,l} P^*_0 T\right\|_1$$

as well as

$$N^{(1)}(T) = \sum_{k\geq 0, l\geq 0} 2^{-k\alpha} 2^{-l\beta} \left\|\Delta^*_{P,k} \Delta^*_{Q,l} T\right\|_1,\ N^{(2)}(T) = \sum_{k\geq 0, l\geq 0} 2^{-k\alpha} 2^{-l\beta} \left\|d^*_{P,k} d^*_{Q,l} T\right\|_1.$$

With these definitions, we define the two norms we will show are equivalent to $\|T\|_{\Lambda^*_{\alpha,\beta}}$. The first norm is defined by

$$\|T\|^{(1)}_{\Lambda^*_{\alpha,\beta}} = N^{(1)}(T) + W^{(1)}_{\mathcal{X}}(T) + W^{(1)}_{\mathcal{Y}}(T) + \|P^*_0 Q^*_0 T\|_1$$

and the second is defined by

$$\|T\|^{(2)}_{\Lambda^*_{\alpha,\beta}} = N^{(2)}(T) + W^{(2)}_{\mathcal{X}}(T) + W^{(2)}_{\mathcal{Y}}(T) + \|P^*_0 Q^*_0 T\|_1.$$

Lemma 19. *The seminorms $W^{(1)}_{\mathcal{X}}(T)$ and $W^{(2)}_{\mathcal{X}}(T)$ are equivalent, as are the seminorms $W^{(1)}_{\mathcal{Y}}(T)$ and $W^{(2)}_{\mathcal{Y}}(T)$.*

Proof. This follows immediately from Proposition 11. □

Lemma 20. *The seminorms $N^{(1)}(T)$ and $N^{(2)}(T)$ are equivalent.*

Proof. We reduce the proof to the case of the single semigroup by applying Proposition

11 repeatedly. We have

$$\begin{aligned} N^{(1)}(T) &= \sum_{l\geq 0} 2^{-l\beta} \sum_{k\geq 0} 2^{-k\alpha} \left\|\Delta^*_{P,k}\Delta^*_{Q,l}T\right\|_1 \\ &= \int_{\mathcal{Y}} \sum_{l\geq 0} 2^{-l\beta} \sum_{k\geq 0} 2^{-k\alpha} \left\|\Delta^*_{P,k}\Delta^*_{Q,l}T(\cdot,y)\right\|_{L^1(X)} dy \\ &\simeq \int_{\mathcal{Y}} \sum_{l\geq 0} 2^{-l\beta} \sum_{k\geq 0} 2^{-k\alpha} \left\|d^*_{P,k}\Delta^*_{Q,l}T(\cdot,y)\right\|_{L^1(X)} dy \\ &= \int_{\mathcal{X}} \sum_{k\geq 0} 2^{-k\alpha} \sum_{l\geq 0} 2^{-l\beta} \left\|\Delta^*_{Q,l}d^*_{P,k}T(x,\cdot)\right\|_{L^1(Y)} dx \\ &\simeq \int_{\mathcal{X}} \sum_{k\geq 0} 2^{-k\alpha} \sum_{l\geq 0} 2^{-l\beta} \left\|d^*_{Q,l}d^*_{P,k}T(x,\cdot)\right\|_{L^1(Y)} dx \\ &= \sum_{k\geq 0} 2^{-k\alpha} \sum_{l\geq 0} 2^{-l\beta} \left\|d^*_{Q,l}d^*_{P,k}T\right\|_1 = N^{(2)}(T). \end{aligned}$$

□

Proposition 19. *The norms* $\|T\|^{(1)}_{\Lambda^*_{\alpha,\beta}}$ *and* $\|T\|^{(2)}_{\Lambda^*_{\alpha,\beta}}$ *are equivalent.*

Proof. This follows immediately from the preceding two lemmas. □

We will now prove that $\|T\|_{\Lambda^*_{\alpha,\beta}}$ and $\|T\|^{(2)}_{\Lambda^*_{\alpha,\beta}}$ are equivalent. We will work formally; all manipulations can be easily justified by the fact that P_kQ_lf converges uniformly to f as $k,l\to\infty$, whenever $f\in\Lambda_{\alpha,\beta}$. Take any function f with $\|f\|_{\Lambda_{\alpha,\beta}}\leq 1$. Write

$$f = \sum_{k\geq 0, l\geq 0} \Delta_{P,k}\Delta_{Q,l}f + \sum_{k\geq 0}\Delta_{P,k}Q_0f + \sum_{l\geq 0}\Delta_{Q,l}P_0f + P_0Q_0f. \tag{5}$$

We want to show that $|\langle f,T\rangle| \lesssim \|T\|^{(2)}_{\Lambda^*_{\alpha,\beta}}$. We will deal with the inner product of f with each of the four terms in the right side of (5) separately. First, we have

$$|\langle P_0Q_0f,T\rangle| = |\langle f, P_0^*Q_0^*T\rangle| \leq \|P_0^*Q_0^*T\|_1$$

since $\|f\|_\infty \leq 1$.

To control the inner product of T with $\sum_{k\geq 0}\Delta_{P,k}Q_0f$, first observe that

$$\begin{aligned} \Delta_{P,k}Q_0 &= \Delta_{P,k+1}(P_{k+2}+P_{k+1})Q_0 \\ &= \Delta_{P,k+1}P_{k+2}Q_0 + \Delta_{P,k+1}P_{k+1}Q_0. \end{aligned} \tag{6}$$

Now,

$$\begin{aligned}\Delta_{P,k+1}P_{k+1}Q_0 &= \Delta_{P,k+1}(P_{k+1}-P_0)Q_0 + \Delta_{P,k+1}P_0Q_0\\ &= \Delta_{P,k+1}d_{P,k+1}Q_0 + \Delta_{P,k+1}P_0Q_0,\end{aligned}$$

and so

$$\begin{aligned}|\langle \Delta_{P,k+1}P_{k+1}Q_0f,T\rangle| &\le |\langle \Delta_{P,k+1}d_{P,k+1}Q_0f,T\rangle| + |\Delta_{P,k+1}P_0Q_0f,T\rangle|\\ &= |\langle \Delta_{P,k+1}f, d^*_{P,k+1}Q^*_0T\rangle| + |\Delta_{P,k+1}f, P^*_0Q^*_0T\rangle|\\ &\le \sup_{x,y}|\Delta_{P,k+1}f(x,y)|\Big\{ \left\|d^*_{P,k+1}Q^*_0f\right\|_1 + \|P^*_0Q^*_0T\|_1 \Big\}\\ &\le 2^{-k\alpha}\left\|d^*_{P,k+1}Q^*_0T\right\|_1 + 2^{-k\alpha}\|P^*_0Q^*_0T\|_1 .\end{aligned}$$

Similarly,

$$|\langle \Delta_{P,k+1}P_{k+2}Q_0f,T\rangle| \le 2^{-k\alpha}\left\|d^*_{P,k+2}Q^*_0T\right\|_1 + 2^{-k\alpha}\|P^*_0Q^*_0T\|_1$$

Combining these inequalities yields, by equation (6),

$$|\langle \Delta_{P,k}Q_0f,T\rangle| \le 2^{-k\alpha}\Big\{ \left\|d^*_{P,k+1}Q^*_0T\right\|_1 + \left\|d^*_{P,k+2}Q^*_0T\right\|_1 + 2\|P^*_0Q^*_0T\|_1 \Big\}$$

Summing over $k \ge 0$ then yields

$$\left|\left\langle \sum_{k\ge 0}\Delta_{P,k}Q_0f,T\right\rangle\right| \lesssim W^{(2)}_{\mathcal{X}}(T) + \|P^*_0Q^*_0T\|_1 .$$

A nearly identical proof shows that

$$\left|\left\langle \sum_{l\ge 0}\Delta_{Q,l}P_0f,T\right\rangle\right| \lesssim W^{(2)}_{\mathcal{Y}}(T) + \|P^*_0Q^*_0T\|_1 .$$

The only term left to control from (5) is the inner product of T with

$$\sum_{k\ge 0,l\ge 0}\Delta_{P,k}\Delta_{Q,l}f.$$

Using the identity

$$\begin{aligned}\Delta_{P,k-1}\Delta_{Q,l-1} &= \Delta_{P,k}(P_{k+1}+P_k)\Delta_{Q,l}(Q_{l+1}+Q_k)\\ &= \Delta_{P,k}P_{k+1}\Delta_{Q,l}Q_{l+1} + \Delta_{P,k}P_k\Delta_{Q,l}Q_{l+1} \qquad (7)\\ &\qquad + \Delta_{P,k}P_{k+1}\Delta_{Q,l}Q_l + \Delta_{P,k}P_k\Delta_{Q,l}Q_l\end{aligned}$$

it follows that we must control the inner product of T with each of the four terms on the right side (applied to f). The argument is the same for each, so we will show it only for $\Delta_{P,k}P_k\Delta_{Q,l}Q_lf = \Delta_{P,k}\Delta_{Q,l}P_kQ_lf$.

We have the easily-verified identity

$$P_kQ_lf = d_{P,k}d_{Q,l}f + d_{P,k}Q_0f + d_{Q,l}P_0f + P_0Q_0f$$

and consequently

$$\begin{aligned}\Delta_{P,k}\Delta_{Q,l}P_kQ_lf = \Delta_{P,k}\Delta_{Q,l}d_{P,k}d_{Q,l}f &+ \Delta_{P,k}\Delta_{Q,l}d_{P,k}Q_0f \\ &+ \Delta_{P,k}\Delta_{Q,l}d_{Q,l}P_0f + \Delta_{P,k}\Delta_{Q,l}P_0Q_0f\end{aligned} \tag{8}$$

We will bound the inner product of T with the sum over $k \geq 0$ and $l \geq 0$ of each of the four terms in (8) separately. First, we have

$$|\langle \Delta_{P,k}\Delta_{Q,l}P_0Q_0f, T\rangle| = |\langle \Delta_{P,k}\Delta_{Q,l}f, P_0^*Q_0^*T\rangle| \leq 2^{-k\alpha}2^{-l\beta}\,\|P_0^*Q_0^*T\|_1$$

and summing over k and l gives the upper bound $\|P_0^*Q_0^*T\|_1$.

Next, observe that

$$|\langle \Delta_{P,k}\Delta_{Q,l}d_{Q,l}P_0f, T\rangle| = |\langle \Delta_{P,k}\Delta_{Q,l}f, d_{Q,l}^*P_0^*T\rangle| \leq 2^{-k\alpha}2^{-l\beta}\left\|d_{Q,l}^*P_0^*T\right\|_1$$

and summing over k and l gives the upper bound $\sum_{l=0}^{\infty} 2^{-l\beta}\left\|d_{Q,l}^*P_0^*T\right\|_1 = W_{\mathcal{Y}}^{(2)}(T)$. Similarly, the inner product of T with $\Delta_{P,k}\Delta_{Q,l}d_{P,k}Q_0f$ can be bounded above by $W_{\mathcal{X}}^{(2)}(T)$.

Finally, we have the upper bound

$$|\langle \Delta_{P,k}\Delta_{Q,l}d_{P,k}d_{Q,l}f, T\rangle| = |\langle \Delta_{P,k}\Delta_{Q,l}f, d_{P,k}^*d_{Q,l}^*T\rangle| \leq 2^{-k\alpha}2^{-l\beta}\left\|d_{P,k}^*d_{Q,l}^*T\right\|_1$$

and summing over k and l gives the upper bound $\sum_{k,l} 2^{-k\alpha}2^{-l\beta}\left\|d_{P,k}^*d_{Q,l}^*T\right\|_1 = N^{(2)}(T)$. Putting the four bounds together and applying equation (8) yields

$$\left|\left\langle \sum_{k,l} \Delta_{P,k}\Delta_{Q,l}P_kQ_lf, T\right\rangle\right| \lesssim \|T\|_{\Lambda_{\alpha,\beta}^*}^{(2)}$$

and the same estimate applied to each of the four terms on the right side of equation (7) gives

$$\left|\left\langle \sum_{k\geq 0, l\geq 0} \Delta_{P,k}\Delta_{Q,l}f, T\right\rangle\right| \lesssim \|T\|_{\Lambda_{\alpha,\beta}^*}^{(2)}$$

completing the proof that $\|T\|_{\Lambda_{\alpha,\beta}^*} \lesssim \|T\|_{\Lambda_{\alpha,\beta}^*}^{(2)}$.

To prove the reverse inequality, as in the proof of Proposition 13 for a single semigroup we define a function f such that $\|f\|_{\Lambda_{\alpha,\beta}} \simeq 1$ and whose inner product with T achieves the norm $\|T\|_{\Lambda_{\alpha,\beta}^*}^{(2)}$. It is easy to check that the function f defined by

$$\begin{aligned}f = \sum_{k,l\geq 0} 2^{-k\alpha}2^{-l\beta}d_{P,l}d_{Q,l}\operatorname{sgn}(d_{P,k}^*d_{Q,l}^*T) &+ \sum_{k\geq 0} 2^{-k\alpha}d_{P,k}Q_0\operatorname{sgn}(d_{P,k}^*Q_0^*T) \\ &+ \sum_{l\geq 0} 2^{-l\beta}d_{Q,l}P_0\operatorname{sgn}(d_{Q,l}^*P_0^*T) + P_0Q_0\operatorname{sgn}(P_0^*Q_0^*T).\end{aligned}$$

satisfies the necessary conditions; in fact, each of the four terms defining f have mixed Lipschitz norm bounded independently of T, and $\langle f, T\rangle = \|T\|^{(2)}_{\Lambda^*_{\alpha,\beta}}$. We have therefore shown

Theorem 6. *The norms $\|T\|_{\Lambda^*_{\alpha,\beta}}$, $\|T\|^{(1)}_{\Lambda^*_{\alpha,\beta}}$, and $\|T\|^{(2)}_{\Lambda^*_{\alpha,\beta}}$ are equivalent on $\Lambda^*_{\alpha,\beta}$.*

References

[1] Bui, H. Q., Duong, X. T., & Yan, L. (2012). Calderón reproducing formulas and new Besov spaces associated with operators. Advances in Mathematics, 229(4), 2449-2502.

[2] Butzer, P. L., & Berens, H. (1967). *Semi-groups of operators and approximation.*

[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] Chavel, Isaac. *Eigenvalues in Riemannian geometry.* Vol. 115. Academic press, 1984.

[5] Coifman, R. R., & Lafon, S. (2006). Diffusion maps. Applied and computational harmonic analysis, 21(1), 5-30.

[6] Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., & Zucker, S. W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. Proceedings of the National Academy of Sciences of the United States of America, 102(21), 7426-7431.

[7] Coifman, R. R., & Maggioni, M. (2006). Diffusion wavelets. Applied and Computational Harmonic Analysis, 21(1), 53-94.

[8] Do Carmo, Manfredo P. *Riemannian geometry.* Springer, 1992.

[9] Dudley, R. M. (2002). *Real analysis and probability* (Vol. 74). Cambridge University Press.

[10] Grigor'yan, Alexander. Heat kernels and function theory on metric measure spaces Contemporary Mathematics 338 (2003): 143-172.

[11] Grigoryan, A., & Liu, L. (2014). Heat kernel and Lipschitz-Besov spaces.

[12] Heinonen, J. (2001). *Lectures on analysis on metric spaces.* Springer.

[13] Indyk, P., & Thaper, N. (2003). Fast image retrieval via embeddings.

[14] Li, Peter, & Shing Tung Yau. On the parabolic kernel of the Schrödinger operator. Acta Mathematica 156.1 (1986): 153-201.

[15] Lieu, L., & Saito, N. (2007, September). Automated discrimination of shapes in high dimensions. In Optical Engineering and Applications (pp. 67011V-67011V). International Society for Optics and Photonics.

[16] Little, A. V., Jung, Y. M., & Maggioni, M. (2009, November). Multiscale Estimation of Intrinsic Dimensionality of Data Sets. In AAAI Fall Symposium: Manifold Learning and Its Applications.

[17] Little, A. V., Maggioni, M., & Rosasco, L. (2012). Multiscale Geometric Methods for Data Sets I: Multiscale SVD, Noise and Curvature.

[18] Marinai, S., Miotti, B., & Soda, G. (2011, September). Using earth mover's distance in the bag-of-visual-words model for mathematical symbol retrieval. In Document Analysis and Recognition (ICDAR), 2011 International Conference on (pp. 1309-1313). IEEE.

[19] Shirdhonkar, S., & Jacobs, D. W. (2008, June). Approximate earth mover's distance in linear time. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on (pp. 1-8). IEEE.

[20] 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.

[21] Sandler, R., & Lindenbaum, M. (2009, June). Nonnegative matrix factorization with earth mover's distance metric. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on (pp. 1873-1880). IEEE.

[22] Stein, E. M. (1970). *Topics in harmonic analysis, related to the Littlewood-Paley theory.* Princeton University Press.

[23] Triebel, H. (1985). *Theory of function spaces.* Birkhäuser.

[24] Wan, X. (2007). A novel document similarity measure based on earth mover's distance. Information Sciences, 177(18), 3718-3730.

[25] Cédric Villani. *Topics in optimal transportation.* No. 58. American Mathematical Soc., 2003.

[26] Zolotarev, Vladimir M. *One-dimensional stable distributions.* Vol. 65. American Mathematical Soc., 1986.