Computing the Monge-Kantorovich Distance

F. Mendivil*

Abstract

The Monge-Kantorovich distance gives a metric between probability distributions on a metric space $\mathbb{X}$ and the MK distance is tied to the underlying metric on $\mathbb{X}$. The MK distance (or a closely related metric) has been used in many areas of image comparison and retrieval and thus it is of significant interest to compute it efficiently. In the context of finite discrete measures on a graph, we give a linear time algorithm for trees and reduce the case of a general graph to that of a tree. Next we give a linear time algorithm which computes an approximation to the MK distance between two finite discrete distributions on any graph. Finally, we extend our results to continuous distributions on graphs and give some very general theoretical results in this context.

1 Introduction

The Monge-Kantorovich Distance is a metric between two probability measures on a metric space $\mathbb{X}$. The MK distance is linked to the underlying metric on $\mathbb{X}$ and is related to a mass transportation problem defined by the two distributions. In this paper, we consider this distance when the metric space is the underlying space of a finite connected graph, in which case the MK distance yields the weak-* topology.

Given two probability measures μ, ν on the compact metric space X we define the MK distance to be

$$d_{MK}(\mu,\nu) = \sup\{\int_X f(x)\ d(\mu-\nu)(x)\ :\ f \in Lip_1(X)\} \tag{1}$$

where $Lip(X)$ is the collection of real-valued *Lipschitz* functions on X and $Lip_1(X)$ is the subset of $Lip(X)$ with Lipschitz constant 1.

The MK distance has its origins in the mathematical theory of *mass transportation*. In 1781 Monge first proposed the mathematical problem of optimizing the cost of moving a pile of soil from a given starting configuration to a given ending configuration. In his original formulation the problem was highly nonlinear, thus extremely difficult. In 1942 Kantorovich introduced (independently of

*Department of Mathematics and Statistics, Acadia University, Canada, *franklin.mendivil@acadiau.ca*

Monge) another "relaxed" version of this problem. Kantorovich's formulation is a linear optimization problem over a convex set – this simplification was a revolutionary step. Since that time, variations of the mass transportation problem have found application in numerous areas (see for example [12, 6]).

The MK distance (given above in (1)) arises from a dual formulation of the Kantorovich-Rubinstein problem (another variation of the general mass transportation problem). It turns out that in the case that the cost function is a metric, the Kantorovich-Rubenstien and Monge-Kantorovich formulations coincide (see [12]).

We first became interested in the MK distance as a result of its use in the theory of invariant measures for IFS (where it was called the Hutchinson metric – see [2]). In the context of inverse problems in IFS (see [7]) the MK distance is used as a "fitness function" for an optimization procedure. Another area of application for the MK distance is using the MK distance between color histograms as part of a measure of similarity between two images (see [6]). In this context, the MK distance is related to the *Earth Mover's Distance* [13, 10, 14] which has been popular as a general method of computing a distance between histogram descriptors in the context of image matching and image retrieval. In all these application areas, the ability to efficiently compute the MK distance between discrete probability distributions is of critical importance.

There has been previous work on efficiently computing the MK distance (or related transport-based distances) (for example, see [1, 4, 5, 6, 9, 10, 14]). Our approach is significantly different from any of these but relates most closely to the results in [1, 4, 6]. For instance, [9] derives (again) the exact algorithm for a 1D distribution and then uses this algorithm, along with space-filling curves, to obtain an approximation in the multidimensional setting. On the other hand, [14] uses a wavelet-based approach to compute an approximation to the MK for distributions in $\mathbb{R}^n$ while [10] assumes an ℓ_1 "taxi-cab" ground distance on a graph which is a regular grid to derive their tree-based algorithm.

This paper has three separate but related parts. In Section 2 we give an efficient linear-time algorithm for the exact MK distance on finite trees and then reduce the problem of of the computation on a general graph to the case of a tree. In Section 3 we give an efficient linear-time method which computes an approximation to the MK distance on any finite graph. In addition, we derive bounds on the quality of the approximation. We comment that for many applications computing the exact MK distance is unnecessary and so this approximation is sufficient. Finally, in Section 4 we extend the discussion of Section 2 to computing the MK distance between continuous distributions on graphs and (slight) generalizations. This final section is intended more to provide a theoretical context than as a suggestion of practical algorithms for applications.

2 Monge-Kantorovich metric on finite point sets

In most practical applications, the metric space will be a finite collection of N points. We model this situation as the points being the vertices of a graph

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The Lipschitz condition in (1) can be expressed in terms of the incidence matrix A of the graph. For example, given the graph $\mathcal{G}$ illustrated in Figure 1, we have the following incidence matrix (where we orient each edge in any arbitrary direction)

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}.$$

Notice that $A : V_{\mathcal{G}} \to E_{\mathcal{G}}$ (mapping from the *vertex-space* of $\mathcal{G}$ to the *edge-space* of $\mathcal{G}$). If f is a real-valued function on the vertices of $\mathcal{G}$, the Lipschitz condition on $\mathcal{G}$ becomes $\|Af\|_\infty \leq 1$. Thus, for two probability distributions π and χ on the vertices of $\mathcal{G}$, the MK distance between them is

$$d_{MK}(\pi, \chi) = \sup\{f^T(\pi - \chi) \ : \ \|Af\|_\infty \leq 1\}. \tag{2}$$

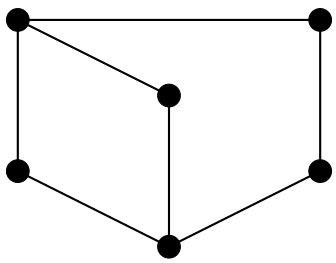

Figure 1: Example graph

In case the distances between adjacent points are not all the same, we have a weighted graph. Continuing with the previous example, suppose the distances along the "outer" cycle are all 2 while those along the two "inner" edges are 3. We then use a diagonal "distance matrix" D, indexed by the set of edges $\mathcal{E}$, which gives the weighted incidence matrix

$$D^{-1}A = \begin{pmatrix} -1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & -1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & -1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & -1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & -1/2 & 0 \\ -1/3 & 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & -1/3 & 0 & 0 & 1/3 \end{pmatrix}$$

and on this graph the Lipschitz condition becomes $\|D^{-1}Af\|_\infty \leq 1$.

Let $\mathbb{1} = (1, 1, 1, \ldots, 1)^T \in \mathbb{R}^N$ and $\mathbb{1}^\perp = \{z \in \mathbb{R}^N : z^T\mathbb{1} = 0\}$ be the space of zero-sum vectors in $\mathbb{R}^N$. The proof of the following is simple (for the first part, notice that A acts as a discrete derivative on vertex functions):

Proposition 1. *Let $\mathcal{G}$ be a graph on N points, D be its distance matrix and A its incidence matrix (as above). Then*

1. $\ker(A) = \{\lambda\mathbb{1} : \lambda \in \mathbb{R}\}$ *and so* $\mathrm{rank(A)} = \mathrm{N} - 1$, *and*
2. $\|x\|_{MK} := \sup\{f^T x : \|D^{-1}Af\|_\infty \leq 1\}$ *defines a norm on* $\mathbb{1}^\perp$.

2.1 Graphs which are trees

In the case where the finite point set $\mathcal{T}$ has the structure of a connected tree, we can easily obtain a linear time algorithm to compute the MK distance on $\mathcal{T}$.

If $\mathcal{T}$ is a tree then the number of edges of $\mathcal{T}$ is one less than the number of vertices of $\mathcal{T}$. Thus, A has full row rank and so there is a matrix B with $AB = I$ (identity on $E_\mathcal{G}$) and $BA|_{\mathbb{1}^\perp} = I_{\mathbb{1}^\perp}$. This means that the MK distance is

$$\begin{aligned}\sup\{f^T(\pi - \chi) : \|D^{-1}Af\|_\infty \leq 1\} &= \sup\{f^T(\pi - \chi) : \|D^{-1}Af\|_\infty \leq 1, f \in \mathbb{1}^\perp\} \\ &= \sup\{(BDg)^T(\pi - \chi) : \|g\|_\infty \leq 1\} \\ &= \|DB^T(\pi - \chi)\|_1. \end{aligned} \tag{3}$$

In this, it is important that $\pi - \chi \in \mathbb{1}^\perp = \ker(A)^\perp$. To make (4) useful, we need to have an explicit expression for B given the tree $\mathcal{T}$.

To motivate the general situation, take the example of a tree $\mathcal{T}$ as shown in Figure 2 with the "top" node chosen as the root. Using a natural labeling of the vertices, we obtain the incidence matrix

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

(notice we have oriented the edges to be going away from the root) and a right inverse

$$B = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

If we think of A as a discrete derivative, then B is a discrete definite integral on $\mathcal{T}$, "accumulating" function values on the edges as we move away from the root node (with the value at the root equal to zero). The statement that $AB = I$ is simply one version of the Fundamental Theorem of Calculus on $\mathcal{T}$. On the other hand, the matrix B^T represents a cumulative sum starting from a given vertex and moving away from the root – all the values from the vertex onwards, so to speak. This is clear by examining B given above and comparing it to the picture of $\mathcal{T}$ in Figure 2.

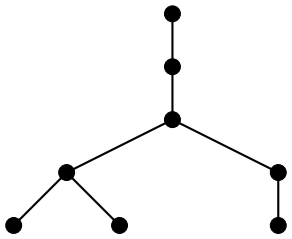

Figure 2: Example of a finite tree

Thus on $\mathcal{T}$ from Figure 2 we have $d_{MK}(\pi, \chi) = \|B^T(\pi - \chi)\|_1$, which is easy to program since we simply accumulate values. The situation for a general tree is much the same.

Theorem 1. *Let $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ be a tree with edge distances $\{d_e\}_{e\in\mathcal{E}}$ and with a node r chosen as the root. Relate the vertices to the edges by saying $e \leq v$ if you must traverse e in order to reach v from the root r. Let π and χ be two probability distributions on the vertices of $\mathcal{T}$. Then the MK distance between π and χ on $\mathcal{T}$ is equal to*

$$d_{MK}(\pi, \chi) = \sum_{e\in\mathcal{E}} d_e \left| \sum_{v\in\mathcal{V}, v>e} \pi(v) - \chi(v) \right|. \tag{4}$$

Proof. We show that (4) is the appropriate version of $\|DB^T(\pi - \chi)\|_1$ for the tree $\mathcal{T}$. It is routine to check that $AB = I_{V_\mathcal{T}}$ with

$$(Af)(\overrightarrow{v_1v_2}) = \frac{f(v_2) - f(v_1)}{d_{\overrightarrow{v_1v_2}}} \text{ and } (Bg)(v) = \sum_{e\in\mathcal{E}, e\leq v} d_e g(e),$$

where $f \in V_\mathcal{T}$ is a function on the vertices and $g \in E_\mathcal{T}$ is a function on the edges. Then we compute B^T by noting that

$$f \cdot (Bg) = \sum_{v\in\mathcal{V}} f(v) \left(\sum_{e\in\mathcal{E}, e\leq v} d_e g(e) \right) = \sum_{e\in\mathcal{E}} \left(d_e \sum_{v>e} f(v) \right) g(e) = (B^T f) \cdot g.$$

□

Clearly we can compute (4) by using a depth-first traversal of $\mathcal{T}$.

Computations for general graphs

If $\mathcal{G}$ is not a tree, we can still say something about computing the MK distance on $\mathcal{G}$. In fact, we can express the MK distance on $\mathcal{G}$ in terms of the MK distance on a tree $\mathcal{T}$ related to $\mathcal{G}$.

To motivate the general case, we will first discuss the case of computing the MK distance on a cycle on N points. This case has been previously studied in the very nice paper [5], where the authors give a linear time algorithm based in the formula

$$\min_{1\leq s\leq N} \sum_{k}^{N} d_k|\alpha_k - \alpha_s| \qquad \text{with} \qquad \alpha_k = \sum_{i=1}^{k} \pi_i - \chi_i.$$

Here α_k is the difference of the cumulative distributions of the two probability distributions and d_k is the k^{th} edge distance. In case the points are equally spaced, the formula simplifies considerably and becomes

$$\sum_{k=1}^{N} |\alpha_k - \alpha_s|$$

where α_s is the median of the values α_k. In the weighted case, the optimal α_s is the weighted median of the values α_k.

Now a key observation is that this can be viewed as a quotient norm. Consider the graph which is a line on $N+1$ points, call this graph $\mathcal{L}$. We add a point mass of weight β to π at the first point of $\mathcal{L}$ and add a corresponding point mass (also of weight β) to χ at the last point (the extra point) of $\mathcal{L}$. The MK distance between these new $\hat{\pi}$ and $\hat{\chi}$ on the graph $\mathcal{L}$ is

$$\sum_{i=1}^{N} \left| \sum_{k=1}^{i} (\pi_i - \chi_i) - \beta \right|$$

and so when we take the infimum over all possible values of β we get

$$\inf_{\beta} \sum_{i=1}^{N} \left| \sum_{k=1}^{i} (\pi_i - \chi_i) - \beta \right| = \inf_{\beta} \sum_{i=1}^{N} |\alpha_i - \beta| = \inf_{1\leq s\leq N} \sum_{i=1}^{N} |\alpha_i - \alpha_s|$$

which is the formula from [5].

Thus the MK distance on the cycle is obtainable from the MK distance on the line. What one does is add an extra vertex to the cycle and peel it apart to obtain a line, $\mathcal{L}$, with one more vertex. The cycle $\mathcal{G}$ is the quotient of $\mathcal{L}$, where the new point and the first point are identified. We then place a point mass on one end of $\mathcal{L}$ and an offsetting point mass on the other end of $\mathcal{L}$, compute the MK distance on $\mathcal{L}$ and minimize over all possible point masses. The important point is that these point masses are placed at the two points that are identified so that on the quotient graph they cancel.

To see it as a quotient norm think of the edge space of the cycle as being a subspace of the edge space of the line (that subspace where the value of the

function on the first and last vertex are the same) and the set of probability measures on the cycle as being a quotient of the set of probability measures on the line (again, using the quotient map). Then the MK norm on the appropriate set of measures on the cycle is exactly the quotient of the MK norm on the appropriate set of measures on the line.

Recall that if $\mathbb{Y}$ is a closed subspace of a normed space $\mathbb{X}$, then the norm on $\mathbb{X}/\mathbb{Y}$ induced by the norm on $\mathbb{X}$ is

$$\|[x]\| = \inf_{[y]=[x]} \|y\| = \inf_{y\in\mathbb{Y}} \|x+y\| = \inf_{q(z)=x} \|z\| \tag{5}$$

for each equivalence class $[x] \in \mathbb{X}/\mathbb{Y}$, where $q : \mathbb{X} \to \mathbb{X}/\mathbb{Y}$ is the quotient map. By general duality theory, this is also equal to

$$\|[x]\| = \sup\{z \cdot x : z \in \mathbb{X}', z \in \mathbb{Y}^\perp, \|z\| \le 1\}. \tag{6}$$

The condition $z \in \mathbb{Y}^\perp$ is because $(\mathbb{X}/\mathbb{Y})' \cong \mathbb{Y}^\perp$.

Definition 1 (graph quotient map)**.** *A map $q : \mathcal{H} \to \mathcal{G}$ between two graphs is a* quotient map *if it is surjective onto the vertices of $\mathcal{G}$ and, for all adjacent vertices $x, y \in \mathcal{H}$, we have that either $q(x) = q(y)$ or $q(x), q(y)$ are also adjacent.*

We denote the space of signed measures on the vertices of a graph $\mathcal{G}$ with total mass zero by $\mathcal{M}(\mathcal{G})$. A quotient $q : \mathcal{H} \to \mathcal{G}$ induces a linear map $q_* : \mathcal{M}(\mathcal{H}) \to \mathcal{M}(\mathcal{G})$ via $q_*(m)(A) = m(q^{-1}(A))$. It also induces linear maps $q^* : E_\mathcal{G} \to E_\mathcal{H}$ and $q^* : V_\mathcal{G} \to V_\mathcal{H}$ via $q^*(f)(x) = f(q(x))$.

Theorem 2. *Let $q : \mathcal{H} \to \mathcal{G}$ be a quotient map and suppose that for any two adjacent vertices $x, y \in \mathcal{H}$ with $q(x) \neq q(y)$ we have the (distance preserving) property*

$$d^\mathcal{H}_{\overrightarrow{xy}} = d^\mathcal{G}_{\overrightarrow{q(x)q(y)}}. \tag{7}$$

Then for any $x \in \mathcal{M}(\mathcal{G})$,

$$\sup\{g \cdot x : \|D_\mathcal{G}^{-1} A_\mathcal{G} g\|_\infty \le 1\} = \inf_{q_*(y)=x} \sup\{h \cdot y : \|D_\mathcal{H}^{-1} A_\mathcal{H} h\|_\infty \le 1\}. \tag{8}$$

Proof. The identity (8) gives the MK distance on $\mathcal{G}$ as a quotient norm of the MK distance on $\mathcal{H}$ and this is the method of proving the identity. We first note that any $\gamma \in q_*^{-1}(0) \subseteq \mathcal{M}(\mathcal{H})$ has total mass 0 and is supported on the vertices of $\mathcal{H}$ that are identified under q. Furthermore, $q_*^{-1}(0) = q^*(V_\mathcal{G})^\perp \subseteq \mathcal{M}(\mathcal{H})$ since $q^*(V_\mathcal{G}) \subset V_\mathcal{H}$ is the set of functions constant on each $q^{-1}(v)$, $v \in \mathcal{V}_\mathcal{G}$.

In addition, condition (7) implies that

$$\frac{g(q(y)) - g(q(x))}{d^\mathcal{G}_{\overrightarrow{q(x)q(y)}}} = \frac{g(q(y)) - g(q(x))}{d^\mathcal{H}_{\overrightarrow{xy}}}$$

and so the following diagram commutes

$$\begin{array}{ccc} V_{\mathcal{G}} & \xrightarrow{D_{\mathcal{G}}^{-1}A_{\mathcal{G}}} & E_{\mathcal{G}} \\ {\scriptstyle q^*}\downarrow & & \downarrow{\scriptstyle q^*} \\ V_{\mathcal{H}} & \xrightarrow[D_{\mathcal{H}}^{-1}A_{\mathcal{H}}]{} & E_{\mathcal{H}} \end{array}$$

Thus,

$$\inf_{q_*(y)=x} \sup\{h \cdot y : \|D_{\mathcal{H}}^{-1}A_{\mathcal{H}}h\|_\infty \leq 1\} = \inf_{q_*(z)=0} \sup\{h \cdot (z+w) : \|D_{\mathcal{H}}^{-1}A_{\mathcal{H}}h\|_\infty \leq 1\}, \quad (9)$$

where $w \in \mathcal{M}(\mathcal{H})$ is any fixed element such that $q_*(w) = x$. The last expression in (9) is a quotient norm on the space $\mathcal{M}(\mathcal{H})/q_*^{-1}(0)$ and so is equal to

$$\begin{aligned} &\sup\{h \cdot w : \|D_{\mathcal{H}}^{-1}A_{\mathcal{H}}h\|_\infty \leq 1, h \in q_*^{-1}(0)^\perp\} \\ &= \sup\{h \cdot w : \|D_{\mathcal{H}}^{-1}A_{\mathcal{H}}h\|_\infty \leq 1, h \in q^*(V_{\mathcal{G}})\} \\ &= \sup\{(g \circ q) \cdot w : \|D_{\mathcal{G}}^{-1}A_{\mathcal{G}}g\|_\infty \leq 1\} \\ &= \sup\{g \cdot x : \|D_{\mathcal{G}}^{-1}A_{\mathcal{G}}g\|_\infty \leq 1\}. \end{aligned}$$

The last two equalities are justified by the commutativity of the above diagram. □

Clearly any graph $\mathcal{G}$ is the quotient of a tree $\mathcal{T}$. To construct such a quotient, start with a spanning tree $\mathcal{S}$ for $\mathcal{G}$ and extend $\mathcal{S}$ to $\mathcal{T}$ by adding vertices for each edge of $\mathcal{G}$ which is not in $\mathcal{S}$. These "extra" vertices will "disappear" in the quotient $q : \mathcal{T} \to \mathcal{G}$. For each such vertex we choose one of the endpoints of the corresponding edge (in $\mathcal{G}$) and connect the new vertex to this endpoint; thus $\mathcal{T}$ has the same number of edges as $\mathcal{G}$. Each edge of $\mathcal{T}$ is given the same length (distance value) as the corresponding edge in $\mathcal{G}$ in order to ensure that (7) is satisfied.

The quotient q is defined to be the identity on each vertex of $\mathcal{S} \subset \mathcal{T}$ and maps each "extra" vertex to the other endpoint of the corresponding edge (from $\mathcal{G}$) – the endpoint to which it is not already connected in $\mathcal{S}$. Using this construction, we obtain the following corollary to the theorem.

Corollary 1. *Let $\mathcal{G}$ be a graph and $\mathcal{T}$ and $q : \mathcal{T} \to \mathcal{G}$ be constructed as above. Then the MK distance between two probability measures π, χ on the vertices of $\mathcal{G}$ is equal to*

$$\inf_{\beta \in q_*^{-1}(0)} \sum_{e \in \mathcal{E}_{\mathcal{T}}} d_e \left| \sum_{v \in \mathcal{V}_{\mathcal{T}}, v > e} \hat{\pi}(v) - \hat{\chi}(v) + \beta(v) \right|, \quad (10)$$

for any fixed elements $\hat{\pi} \in q_*^{-1}(\pi)$ and $\hat{\chi} \in q_*^{-1}(\chi)$.

As a simple example, Figure 3 shows the original graph $\mathcal{G}$, then the spanning tree $\mathcal{S}$ and then the "extended" tree $\mathcal{T}$. Notice how only vertices get identified in the quotient; all the edges of $\mathcal{G}$ are present in the tree $\mathcal{T}$. This means that the number of "extra" vertices we must add is equal to $|E_{\mathcal{G}}| - |V_{\mathcal{G}}| + 1$, which is the number of "independent loops" in $\mathcal{G}$ (or the number of generators of the first level homology of $\mathcal{G}$).

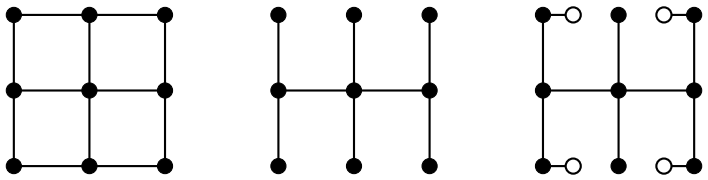

Figure 3: Example of a graph as a quotient of a tree

3 Approximation using the pseudo-inverse $A^\dagger$

The computation in (4) depends on the fact that the incidence matrix A has a right inverse when $\mathcal{G}$ is a tree. For a connected graph, this is the only time when A has full row rank and thus the only time it has a right inverse. In [6], the authors suggested using an approximate inverse to obtain an approximation to the MK distance. In this section we give a brief discussion of this idea and give a refinement of their error bound.

We first define a generalization of the MK-norm on $\mathbb{1}^\perp$ for the graph $\mathcal{G}$.

Definition 2. *Let $1 \leq p \leq \infty$. Define the norm $\|\cdot\|_{H,p}$ on $\mathbb{1}^\perp$ by*

$$\|x\|_{H,p} = \sup\{f^T x : \|D^{-1}Af\|_p \leq 1\}. \tag{11}$$

For $1 \leq p \leq q \leq \infty$ we have $\|y\|_p \geq \|y\|_q$ and thus $\|x\|_{H,p} \leq \|x\|_{H,q}$.

Given a matrix A, we denote by $A^\dagger$ the *pseudo-inverse* (or *Moore-Penrose pseudo-inverse*) of A [3].. The following is a simple extension of a result mentioned in [6].

Proposition 2. *For any $x \in \mathbb{1}^\perp$ we have that*

$$\|x\|_{H,p} \leq \|(A^\dagger D)^T x\|_q \leq \|D^{-1}AA^\dagger D\alpha\|_p \, \|x\|_{H,p} \tag{12}$$

for any $\alpha \in \{z : \|z\|_p \leq 1, z^T(A^\dagger D)^T x = \|(A^\dagger D)^T x\|_q\}$. In particular, for the special case of $p = 2$, we have

$$\|x\|_{H,2} = \|(A^\dagger D)^T x\|_2.$$

Proof.

$$\begin{aligned}\|x\|_{H,p} &= \sup\{f^T x : \|D^{-1}Af\|_p \le 1\} = \sup\{f^T x : f \in \mathbb{1}^\perp, \|D^{-1}Af\|_p \le 1\} \\ &= \sup\{(A^\dagger Dg)^T x : \|g\|_p \le 1, g \in Range(D^{-1}A)\} \\ &\le \sup\{(A^\dagger Dg)^T x : \|g\|_p \le 1\} = \|(A^\dagger D)^T x\|_q,\end{aligned}$$

which shows the first inequality. Now, let $z = D^{-1}AA^\dagger D\alpha$. Then $z \in Range(D^{-1}A)$, $\|z\|_p = \|D^{-1}AA^\dagger D\alpha\|_p$ and

$$A^\dagger Dz = A^\dagger AA^\dagger D\alpha = A^\dagger D\alpha \quad \Rightarrow \quad (A^\dagger Dz)^T x = (A^\dagger D\alpha)^T x = \|(A^\dagger D)^T x\|_q.$$

Thus,

$$\begin{aligned}\|(A^\dagger D)^T x\|_q &= \sup\{(A^\dagger Dg)^T x : \|g\|_p \le 1\} \\ &\le \sup\{(A^\dagger Dz)^T x : z \in Range(D^{-1}A), \|z\|_p \le \|D^{-1}AA^\dagger D\alpha\|_p\} \\ &\le \|D^{-1}AA^\dagger D\alpha\|_p \sup\{(A^\dagger Dz)^T x : z \in Range(D^{-1}A), \|z\|_p \le 1\} \\ &= \|D^{-1}AA^\dagger D\alpha\|_p \|x\|_{H,p}.\end{aligned}$$

The second conclusion follows by the definition of $A^\dagger$ since $D^{-1}AA^\dagger D$ is an orthogonal projection onto the range of $D^{-1}A$. □

A trivial corollary of this result is

$$\|x\|_{H,p} \le \|(A^\dagger D)x\|_q \le \|D^{-1}AA^\dagger D\|_p \|x\|_{H,p}. \tag{13}$$

For a given $x \in \mathbb{1}^\perp$, the estimate (12) is almost always tighter than that given in (13). Since $\|x\|_\infty \le \|x\|_2 \le \sqrt{N}\|x\|_\infty$ we have also the (extremely weak but uniform) bound $\|D^{-1}AA^\dagger D\|_\infty \le \sqrt{N}$.

Example 1 (**The cycle on N points**)**.** Let G be a cycle on N equally spaced vertices. In this case we calculate that $AA^\dagger = I - \frac{1}{N}\mathcal{O}$, where $\mathcal{O}$ is the $N \times N$ matrix of all ones. Thus, for example $\|AA^\dagger\|_\infty = 2 - 2/N = \|AA^\dagger\|_1$. Additionally we see that $Range(A) = \mathbb{1}^\perp$.

Now if α is as in the proposition for $p = \infty$, then $\alpha_i = \pm 1$ is the sign of the ith component of $(A^\dagger)^T x$ and

$$\|AA^\dagger\alpha\|_\infty = 1 + \frac{1}{N}\left|\sum_i \alpha_i\right|. \tag{14}$$

Notice that at least one component of α must be positive and at least one negative since all the vectors in the range of A have zero sum. We guess that we will see all of the allowed sign patterns of $(A^\dagger)^T x$ equally likely and thus $|\sum_i \alpha_i|$ ranges from 0 to $N-2$.

Simulation shows, however, that the upper bound given in (14) is too pessimistic for large N. The simulation was accomplished by drawing π_1, π_2 as two uniformly random N-point probability distributions and setting $x = \pi_1 - \pi_2$.

Then all three of $\|x\|_{H,\infty}$, $\|(A^\dagger)^T x\|_1$ and $\|AA^\dagger\alpha\|_\infty$ were computed. As N increased, the average of the ratio

$$\frac{\|(A^\dagger)^T x\|_1}{\|x\|_{H,\infty}}$$

seemed to decrease with a limit of one. This indicates that for large N "most" of the time $\|(A^\dagger)^T x\|$ is a very good estimate for $\|x\|_{H,\infty}$ and much better than (14) would indicate.

If we have a distance matrix D, then

$$\|D^{-1}AA^\dagger D\|_\infty = \frac{1}{N\min_j d_j}\sum_i d_i - \frac{1}{N} \quad \text{and} \quad \|D^{-1}AA^\dagger D\|_1 = \frac{\max_i d_i}{N}\sum_j \frac{1}{d_j} - \frac{1}{N}.$$

The expression for $\|D^{-1}AA^\dagger D\alpha\|_\infty$ is not particularly illuminating and so we do not give it. $\diamond$

Remark 1. The above arguments use the fact that $AA^\dagger$ is a projection onto $range(A)$ but it is not necessary that it be an orthogonal projection. Perhaps another "inverse" to A would be provide a tighter bound for $\|D^{-1}AA^\dagger D\|_p$?

4 Continuous distributions

In this section we briefly outline some extensions of our results to the case of continuous (as opposed to discrete) distributions.

4.1 Dendrites

Our aim is to extend the discussion in section 2.1 to the continuous setting. We only consider compact spaces, and so we specialize our definitions to this case (see [11, Chapter 10]).

Definition 3 (**dendrite**). *A* dendrite *is a compact, connected, locally connected metric space which contains no simple closed curves.*

Let $\mathcal{T}$ be a dendrite. Then $\mathcal{T}$ can be decomposed into a set of *endpoints* $\mathcal{E}$ and a countable set of arcs $\{\mathcal{C}_i\}$ and these arcs only meet at *branch points*; unlike for trees, the *order* of a branch point can be infinite. We choose one branch point $r \in \mathcal{T}$ to be the *root* of the tree. Any two points $x, y \in \mathcal{T}$ are connected by a unique path and we say that $x \leq y$ if x lies on the path from r to y. Dendrites are models for the underlying topological space of an infinite tree when this space is compact.

We need to assume recifiability in order to have an arclength measure for a Fundamental Theorem of Calculus. For a path $\sigma : [0,1] \to \mathcal{T}$, the *length* of the path is defined (as usual) to be

$$\ell(\sigma) := \sup \sum_i d(\sigma(t_i), \sigma(t_{i+1})) \tag{15}$$

where the supremum is over all finite partitions $0 = t_0 < t_1 < t_2 < \ldots < t_n = 1$ of $[0, 1]$.

We assume that for any two $x, y \in \mathcal{T}$ the length of the arc between them is equal to the distance between them, i.e. that $\mathcal{T}$ is a *length space*. Since each arc $\mathcal{C}_i$ is homeomorphic to $[0, 1]$, at all non-branch points there is a well-defined arclength measure λ. For simplicity we assume that $\lambda(\mathcal{T}) < \infty$. We comment that $f \in Lip_1(\mathcal{T})$ if and only if $|f'(x)| \leq 1$ for λ-almost all points x.

With this setup, we have the following formula for the MK distance on $\mathcal{T}$. It is exactly what one would expect as the continuous analogue of Theorem 1.

Theorem 3. *Let μ and ν be two probability measures on $\mathcal{T}$. Then the MK distance between μ and ν is given by*

$$\int_{\mathcal{T}} |F_\eta(x)| \; d\lambda(x) \tag{16}$$

where $\eta = \mu - \nu$, λ is length measure along $\mathcal{T}$ and $F_\eta(x) = \eta(\{y : y > x\})$.

Proof. Let $f : \mathcal{T} \to \mathbb{R}$ have Lipschitz factor 1. Then f is differentiable almost everywhere (with respect to λ) and $|f'(t)| \leq 1$. Let $g = f'$ where it exists and 0 elsewhere. Define

$$B = \{(x, y) \in \mathcal{T} \times \mathcal{T} \; : \; x < y\}$$

and $h(x, y) = g(x)$ a function on B. Finally, define $\rho = \lambda \otimes \eta$ on $\mathcal{T} \times \mathcal{T}$. Then the integral

$$\int \int_B h(x, y) \; d\rho(x, y) \tag{17}$$

is equal to (by Fubini's Theorem)

$$\int_{\mathcal{T}} g(x)\eta(\{y : y > x\}) \, d\lambda(x) = \int_{\mathcal{T}} g(x) \left(\int_{y>x} d\eta(y) \right) \, d\lambda(x) = \int \int_B h(x, y) \, d\eta(y) \, d\lambda(x).$$

However, (17) is also equal to (integrating in the other order)

$$\int_{y \in \mathcal{T}} \int_{x<y} h(x, y) \, d\lambda(x) \, d\eta(y) = \int_{y \in \mathcal{T}} \left(\int_{x<y} g(x) \; d\lambda(x) \right) \, d\eta(y) = \int_{y \in \mathcal{T}} f(y) \, d\eta(y).$$

Here we have assumed that the value of f at the root is zero. This does not matter since $\eta(\mathcal{T}) = 0$, and so we can add an arbitrary constant f in the integral

$$\int_{y \in \mathcal{T}} f(y) \; d\eta(y)$$

and not change the value. Therefore, we obtain that

$$\int f(x) \; d\eta(x) = \int f'(x) F_\eta(x) \; d\lambda(x) \tag{18}$$

for all f with Lipschitz factor one. Now the supremum of the right hand side of (18) over all Lipschitz f is simply the 1-norm of the function F_η. Thus the MK distance is equal to

$$\int_{\mathcal{T}} |F_\eta(x)| \; d\lambda(x)$$

as desired. □

Comparing this result with the version for discrete distributions on the vertices trees we see that the matrix B computes the "distribution function" $\eta(\{y : y \leq x\})$ while B^T computes the "complementary distribution function" $F_\eta(x) = \eta(\{y : y > x\})$.

4.2 Quotients of dendrites

First we have the following abstract generalization of Theorem 2.

Theorem 4. *Let $q : \mathbb{X} \to \mathbb{Y}$ be a quotient map between compact metric spaces. Suppose that $q^*(C(\mathbb{Y})) \bigcap Lip_1(\mathbb{X}) = q^*(Lip_1(\mathbb{Y}))$. Then for any two probability measures π, χ on $\mathbb{Y}$ we have that*

$$d_{MK}^{\mathbb{Y}}(\pi, \chi) = \inf\{d_{MK}^{\mathbb{X}}(\hat{\pi}, \hat{\chi}) : q_*(\hat{\pi}) = \pi, q_*(\hat{\chi}) = \chi\}. \tag{19}$$

Proof. We see that

$$\begin{aligned}
&\inf_{\eta \in q_*^{-1}(\pi - \chi)} \sup\{\int_{\mathbb{X}} f(x) \; d\eta(x) : f \in Lip_1(\mathbb{X})\} \\
&= \inf_{\eta \in q_*^{-1}(0)} \sup\{\int_{\mathbb{X}} f(x) \; d(\theta + \eta) : f \in Lip_1(\mathbb{X})\} && (20) \\
&= \sup\{\int_{\mathbb{X}} h(x) \; d\theta(x) : h \in Lip_1(\mathbb{X}) \cap q^*(C(\mathbb{Y}))\} && (21) \\
&= \sup\{\int_{\mathbb{X}} h(x) \; d\theta(x) : h \in q^*(Lip_1(\mathbb{Y}))\} && (22) \\
&= \sup\{\int_{\mathbb{X}} g(q(x)) \; d\theta(x) : g \in Lip_1(\mathbb{Y})\} && (23) \\
&= \sup\{\int_{\mathbb{Y}} g(y) \; d(\pi - \chi)(y) : g \in Lip_1(\mathbb{Y})\}. && (24)
\end{aligned}$$

Here $\theta \in \mathcal{M}(\mathbb{X})$ is some fixed element in $q_*^{-1}(\pi - \chi)$. The equality of (20) and (21) is by general properties of quotient norms and duality (equations (5) and (6)). The equality of (21) and (22) is by our assumption on q. The equality of (22) and (23) is by the definition of q^*. Finally, the equality of (23) and (24) is by a change-of-measure and the fact that $q_*(\theta) = \pi - \chi$. □

The condition on the quotient map q expresses the fact that q somehow preserves the Lipschitz classes in mapping from $\mathbb{X}$ to $\mathbb{Y}$ (notice that this is the same condition in Theorem 2). We point out that if q is non-expanding then

$q^*(Lip_1(\mathbb{Y})) \subseteq Lip_1(\mathbb{X}) \cap q^*(C(\mathbb{Y}))$. However, the reverse condition need not hold. To see this, consider $\mathbb{X} = [0, 2/5]$, $\mathbb{Y} = [0, 4/25]$ and $q(x) = x^2$, so that q is strictly contractive. Consider $f : \mathbb{X} \to \mathbb{R}$ given by $f(x) = x$. Then clearly $f \in Lip_1(\mathbb{X})$. However, there is no $g \in Lip(\mathbb{Y})$ so that $x = f(x) = g(q(x)) = g(x^2)$, since the only solution to this is $g(y) = \sqrt{y}$ which is not in $Lip(\mathbb{Y})$.

The condition $q^*(Lip_1(\mathbb{Y})) = Lip_1(\mathbb{X}) \cap q^*(C(\mathbb{Y}))$ is both necessary and sufficient for the equality (21) to (22) to hold for all π, χ. To see this just notice that both $q^*(Lip_1(\mathbb{Y}))$ and $Lip_1(\mathbb{X}) \cap q^*(C(\mathbb{Y}))$ are closed and convex and so if they are not equal, the Hahn-Banach theorem would give a measure θ which separates them. One then easily uses a scaled version of θ to obtain probability measures π and χ for which the equality does not hold.

We conclude with a simple extension of Corollary 1. We leave the details of the proof to the reader. For our current purposes we only need to identify endpoints of the dendrite, and so the conditions on q are given in terms of this.

Proposition 3. *Let $\mathcal{T}$ be a dendrite (as above) and $q : \mathcal{T} \to \mathbb{X}$ be a quotient map which identifies only endpoints of $\mathcal{T}$ and is an isometry on all of the arcs of $\mathcal{T}$. Assume that the distance on $\mathbb{X}$ is given by the length of the shortest path. Then $q^*(Lip_1(\mathcal{T})) = q^*(C(\mathcal{T})) \cap Lip_1(\mathbb{X})$ and so*

$$d_{MK}^{\mathbb{X}}(\pi, \chi) = \inf\{d_{MK}^{\mathcal{T}}(\hat{\pi}, \hat{\chi}) : q_*(\hat{\pi}) = \pi, q_*(\hat{\chi}) = \chi\}.$$

Acknowledgments

The author would like to thank Steve Demko for a very large number of illuminating discussions on this topic and explaining to me his work in [6]. This work is partially supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (238549).

References

[1] F. Barahona, C. A. Cabrelli, and U. M. Molter, "Computing the Hutchinson distance by network flow methods," *Random and Comput. Dynam.*, **1**, pp. 117-129, 1992.

[2] Barnsley, Michael, *Fractals Everywhere*, Academic Press, New York, 1988.

[3] T. Boullion and P. Odell, *Generalized Inverse Matrices*, Wiley-Interscience, 1971.

[4] J. Brandt, C. Cabrelli, and U. Molter, "An algorithm for the computation of the Hutchinson distance," *Inform. Process. Lett.*, **40**, 113-117, 1991.

[5] C. A. Cabrelli and U. M. Molter, "The Kantorovich metric for probability measures on the circle," *J. Comput. Appl. Math.*, **57**, pp. 345-361, 1995.

[6] K. Chen, S. Demko and R. Xie, "Similarity-based retrieval of images using color histograms," Storage and Retrieval for Image and Video Databases VII, Yeung, Yeo and Bouman, Editors, *Proceedings of SPIE* vol 3656, pp. 643-652, 1998.

[7] A. Deliu, F. Mendivil and R. Shonkwiler, "Genetic Algorithms for the 1-D Fractal Inverse Problem," Proceedings of the 4th International Conference on Genetic Algorithms, San Diego, July 1991.

[8] J. Hutchinson, "Fractals and self-similarity," *Indiana Univ. Math. J.*, **30**, pp. 713-747, 1981.

[9] M. Jang, S. Kim, C. Faloutsos and S. Park, A linear-time approximation of the earth mover's distance, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pp. 504-514, 2011.

[10] H. Ling and K. Okada, An efficient Earth Mover's distance algorithm for robust histogram comparison, *IEEE Transactions on PAMI*, **29**, no. 5, pp. 840 - 853, 2006.

[11] S. Nadler, *Continuum theory: An introduction*, Dekker, New York, 1992.

[12] S. T. Rachev and L. Ruschendorf, *Mass Transportation Problems volume I: Theory*, Springer-Verlag, 1998.

[13] Y. Rubner, C. Tomasi, and J. Guibas, The Earth Mover's distance as a metric for image retrieval, *Int. J. Comput. Vis.* **40**, no. 2, pp. 99-121, 2000.

[14] S. Shirdhonkar and D. Jacobs, Approximate Earth Mover's distance in linear time, in *Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition*, CVPR, 2008.