DISMA POLITECNICO DI TORINO
IG SEMINAR DECEMBER 6 2019

handout: *Kantorovich distance on a finite metric space* `arXiv:1905.07547`

Luigi Montrucchio · Giovanni Pistone*

December 6, 2019

The Kantorovich distance (1-Wasserstein distance) is of interest in Information Geometry because provides a metric geometry on the probability simplex $\Delta(X)$, X finite set, which is compatible with the affine geometry. That is, it is a distance between probability functions such that the mixture models are metric geodesic. Recall that the probability simplex is a convex set whose affine (tangent) space is the set of ξ such that $\sum_x \xi(x) = 0$.

Long before the full development by Kantorovich, the issue of evaluating the dissimilarity between two probability distribution $\mu, \nu \in \Delta(X)$ has been discussed in terms of a hidden random variable Z along with the two functions x, y for which $x(Z) \sim \mu$ and $y(Z) \sim \nu$. The dissimilarity is then defined as the minimum expected value $\mathbb{E}\left(d(x(Z), y(Z))\right)$. Notice that it is always possible to assume $Z = X \times X$.

1 C. Gini

The problem appeared first in G. Monge (1781) as a transportation problem. One of the first result was published by Gini (1914). Gini was motivated by the idea of study ordered dependence from random variables and provided a solution in a very special case that is interesting because highlights some of the basic features of the theory.

2 L. Kantorovich (1942)

Let X be a set with n points, with generic points $x, y, z, \dots$. In general, we do not number finite sets. $\Delta(X)$ is the probability simplex on X, that is, the set of all probability functions $\mu, \nu, \dots$. The sample space X is endowed with a distance d. Frequently, the distance is provided by a weighted graph.

Given a couple (μ, ν) of probability functions, recall that a joint probability function $\gamma \in \Delta(X \times X)$ is a *coupling*, if μ and ν are the two margins of γ, respectively. The set of all couplings $\mathcal{P}(\mu, \nu)$ is a subset of $\Delta(X \times X)$ defined by the $2(n-1)$ independent affine constraints

$$\sum_{y \in X} \gamma(x, y) = \mu(x) \, , \quad \sum_{x \in X} \gamma(x, y) = \nu(y) \, , \qquad x, y \neq x_0 \, .$$

In particular, it is a polytope, and, being such, it is the convex combination of its vertices.

Definition 1 Given $\mu, \nu \in \Delta(X)$, the *Kantorovich distance* (K-distance) is defined on $\Delta(X) \times \Delta(X)$ to be the value of the Linear Programming problem

$$d(\mu, \nu) = \inf \left\{ \sum_{x, y \in X} d(x, y) \gamma(x, y) \,\middle|\, \gamma \in \mathcal{P}(\mu, \nu) \right\} \, . \tag{1}$$

Luigi Montrucchio
Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino, Italy
E-mail: luigi.montrucchio@unito.it

Giovanni Pistone
de Castro Statistics, Collegio Carlo Alberto, Piazza Vincenzo Arbarello 8, 10122 Torino, Italy
E-mail: giovanni.pistone@carloalberto.org

This statistical index is a distance that extends the ground distance, i.e., $d(\delta_x, \delta_y) = d(x, y)$. Other properties are easily proved in our finite setting and they are still true in the general (non-finite) setting.

Observe first that the mixture model $\mu(t) = (1 - t)\mu + t\nu$, $t \in [0, 1]$, is a metric geodesic linking the two probability functions μ and ν, that is, $d(\mu(s), \mu(t)) = |t - s|\, d(\mu, \nu)$. Second, the metric convergence is equivalent to the weak convergence. This ensures the compatibility of this set-up both with the affine IG and the required convergence.

3 Duality, Arens-Eells spaces and CUT seminorms

A real function u on X is called 1-Lipschitz for the distance d, if $|u(x) - u(y)| \leq d(x, y)$, for all $x, y \in X$. Equivalently, $u(y) \leq d(x, y) + u(x)$ for all $x, y \in X$.

The condition $|u(x) - u(y)| \leq Kd(x, y)$ for some K gives rise to the linear space of Lipschitz functions $\mathrm{Lip}(d)$. The best constant K is a semi-norm, $\|u\|_{\mathrm{Lip}(d)}$. The set of 1-Lipschitz functions will denoted by $\mathrm{Lip}_1(d)$. When the distance is generated by a weighted graph, it is enough to check the Lipschitz condition on edges, as established in the next proposition.

Proposition 1 *Let (X, w) be a weighted graph with associated distance d. A function $u\colon X \longrightarrow \mathbb{R}$ belongs to $\mathrm{Lip}_1(d)$ if, and only if, $|u(x) - u(y)| \leq d(x, y)$ for each edge $xy \in \mathcal{E}$.*

More generally, we can say that $|u(x) - u(y)| \leq Kd(x, y)$ holds for each edge $xy \in \mathcal{E}$ if, and only if, $\|u\|_{\mathrm{Lip}(d)} \leq K$.

Kantorovich duality theorem below is an application of LP duality. For a detailed treatment, see, for example, Santambrogio (2015) and Villani (2008).

Theorem 1 (Kantorovich duality) *Let μ and ν be given probability functions on the finite metric space (X, d) and $\mathcal{P}(\mu, \nu)$ be the set of couplings. Then,*

$$0d(\mu, \nu) = \min \left\{ \sum_{x,y \in X} d(x, y)\gamma(x, y) \,\middle|\, \gamma \in \mathcal{P}(\mu, \nu) \right\} = \max \left\{ \sum_{z \in X} u(z)(\mu(z) - \nu(z)) \,\middle|\, u \in \mathrm{Lip}_1(d) \right\} .$$

The second term of the equality in Theorem 1 shows that the distance $d(\mu, \nu)$ depends only on the difference $\xi = \mu - \nu$. For such functions we have $\sum_z \xi(z) = 0$ and $\sum_z |\xi(z)| \leq 2$. Conversely, every function ξ that satisfies $\sum_z \xi(z) = 0$ is the difference of two probability functions if $\sum_z |\xi(z)| \leq 2$. In fact, the assumptions made on ξ imply $\sum_z \xi^+(z) = \sum_z \xi^-(z) \leq 1$. If the strict inequality holds, given any probability function p we can choose $\alpha > 0$ such that both $\xi^+ + \alpha p$ and $\xi^- + \alpha p$ are probability functions.

If we ignore this restriction on the elements ξ, we obtain the vector space $M_0(X)$ of zero-mass measure functions and we can hence define on this space the so-called Kantorovich-Bernstein norm (KB-norm),

$$\|\xi\|_{\mathrm{KB}} = \sup_{u \in \mathrm{Lip}_1(d)} \sum_{z \in X} \xi(z)u(z) \,, \tag{2}$$

so that the K-distance is just the restriction of the KB-norm, i.e., $d(\mu, \nu) = \|\mu - \nu\|_{\mathrm{KB}}$.

3.1 Arens-Eells norm

Let $\mathrm{Lip}^+(d)$ be the space of the Lipschitz functions defined on the pointed metric space (X, d, x_0), namely, the set of Lipschitz functions u for which $u(x_0) = 0$ and where x_0 is a distinguished element of X. It turns out to be a Banach space by norm $\|u\|_{\mathrm{Lip}(d)}$, given by the smallest Lipschitz constant of u. The symbol $\mathrm{Lip}_1^+(d)$ denotes the unit ball and $\operatorname{ext} \mathrm{Lip}_1^+(d)$ the set of its extreme points.

Each difference of delta functions belongs to $M_0(X)$ and the isometric property is verified through the dual formulation

$$d(\delta_x, \delta_y) = \left\|\delta_x - \delta_y\right\|_{\mathrm{KB}} = \sup_{u \in \mathrm{Lip}_1^+(d)} \left\langle \delta_x - \delta_y, u \right\rangle = \sup_{u \in \mathrm{Lip}_1^+(d)} (u(x) - u(y)) = d(x, y) \,. \tag{3}$$

The difference of delta functions spans the whole space, i.e., every $\xi \in M_0(X)$ can be written as

$$\xi = \sum_{x,y} a(x,y)(\delta_x - \delta_y)\,, \quad A = [a(x,y)]_{x,y\in X} \in \mathbb{R}^{X\times X}\,. \tag{4}$$

If we compute the KB-norm of a generic function ξ in eq. (4), we get through eq. (3)

$$\|\xi\|_{\mathrm{KB}} = \left\|\sum_{x,y} a(x,y)(\delta_x - \delta_y)\right\|_{\mathrm{KB}} \leq \sum_{x,y\in X} |a(x,y)|\, d(x,y)\,.$$

The vector space $M_0(X)$ can be endowed with the following norm, in which case it will be called the Arens-Eells space Æ (X). Here, we follow the presentation by Weaver (2018).

Definition 2 The norm $\|\xi\|_{Æ}$ is defined by

$$\|\xi\|_{Æ} = \inf\left\{\sum_{x,y\in X} |a(x,y)|\, d(x,y)\right\}\,, \tag{5}$$

where the inf is made on all the representations of ξ in eq. (4).

Arens-Eells construction has a wider range of application than finite metric spaces. In fact, in a more general setting, Arens-Eells space is defined as the norm-closure of the space of all zero-mass measures with finite support on an arbitrary metric space X. It is also known in literature as a Lipschitz-free space over X and frequently denoted by $\mathcal{F}(X)$.

The Banach space Æ (X) is a predual of $\mathrm{Lip}^+(d)$. More specifically, the linear isometry $T :$ Æ $(X)^* \longrightarrow \mathrm{Lip}^+(d)$ is given by

$$T(\phi)(x) = \phi(\delta_x - \delta_{x_0})$$

for every $\phi \in$ Æ $(X)^*$, $x \in X$ and where x_0 is the distinguished point of X.

Consequently,

$$\|\xi\|_{Æ} = \|\xi\|_{\mathrm{KB}} = \sup\left\{\langle \xi, u\rangle \mid u \in \mathrm{Lip}_1^+(d)\right\} \tag{6}$$

for all $\xi \in$ Æ (X). In addition, there exists a multi-mapping $J\colon$ Æ $(X) \longrightarrow \mathrm{Lip}_1^+(d)$ such that the alignment condition $\langle \xi, J(\xi)\rangle = \|\xi\|_{Æ}$ is satisfied. In our finite-dimensional setting, it holds also the reflexivity property, Æ $(X) = \mathrm{Lip}^+(X)^*$.

Here, we just recall an important property of the Arens-Eells norm (see Weaver (2018) for more details).

Proposition 2 *The norm* $\|\cdot\|_{Æ}$ *is the largest semi-norm on the space* Æ (X) *which satisfies* $\|\delta_x - \delta_y\| \leq d(x,y)$.

Proof If $\|\cdot\|$ *is a semi-norm that satisfies the claimed requirements, we have from eq.* (4) *that*

$$\|\xi\| = \left\|\sum_{x,y\in X} a(x,y)(\delta_x - \delta_y)\right\| \leq \sum_{x,y\in X} |a(x,y)|\, d(x,y)$$

is true for any representation of ξ *as a linear combination of differences of Dirac functions. Consequently,* $\|\xi\| \leq \|\xi\|_{Æ}$. □

Another important property is that if X_0 is a nonempty subset of the metric space X, then the identity map takes Æ (X_0) isometrically into Æ (X), see Theorem 3.7 in Weaver (2018).

We we start with a useful notion that refines the adjacency property for points of a graph.

Definition 3 Two vertices x, y of a weighted graph (X, w) are said to be *close* if they are adjacent and, in addition, $d(x,y) = w(x,y)$, i.e., the path xy is one of the shortest paths joining the points themselves.

Adjacent vertices are necessarily close in a tree. So too are all the adjacent pairs in an unweighted graph. Observe further that, in any path $x_1, x_2, \ldots, x_n$ of minimum length, two adjacent vertices are necessarily close. This is the reason why it holds the equality

$$d(x_1, x_n) = \sum_{i=1}^{n-1} d(x_i, x_{i+1}) \tag{7}$$

along points of a path of minimal length.

To see this, suppose not. We would have $w(x_i, x_{i+1}) \geq d(x_i, x_{i+1})$ for all i and $w(x_j, x_{j+1}) > d(x_j, x_{j+1})$ for some j. Therefore

$$d(x_1, x_n) = \sum_{i=1}^{n-1} w(x_i, x_{i+1}) > \sum_{i=1}^{n-1} d(x_i, x_{i+1})$$

which contradicts the triangular inequality.

It is worth remarking that one could replace adjacent points with close points in Proposition 1 too.

3.2 Extreme points

It is known in literature a characterization of the extreme points of the unit ball of the normed space $\mathrm{Lip}^+(d)$, for generic metric spaces, cf. Farmer (1994), Smarzewski (1997).

Here, we are concerned with a useful qualification that holds in finite spaces. For ease of the reader we provide a complete proof which essentially follows Th. 2.59 Weaver (2018).

Theorem 2 *Let (X, x_0) be a pointed finite metric space. A function $f \in \mathrm{Lip}_1^+(d)$ is extremal if and only if for every $x \in X$ there is a path $x_0, x_1, \ldots, x_{n-1}$, with $x_{n-1} = x$, such that*

$$|f(x_i) - f(x_{i-1})| = d(x_i, x_{i-1})$$

for $i = 1, .., n-1$. When the distance is induced by a graph, the path linking x_0 and x can be taken to be a sequence of close points.

Proof Suppose a function f satisfies the stated condition and consider the functions $f \pm u \in \mathrm{Lip}_1^+(d)$. We must show that $u = 0$.

Fixing $x \in X$, by hypothesis there exists a path $x_0, x_1, \ldots, x_{n-1} = x$, with $|f(x_i) - f(x_{i-1})| = d(x_i, x_{i-1})$.

Further, in view of proposition 1, it holds

$$|f(x_i) - f(x_{i-1}) + u(x_i) - u(x_{i-1})| \leq d(x_i, x_{i-1})$$

as well as

$$|f(x_i) - f(x_{i-1}) - u(x_i) + u(x_{i-1})| \leq d(x_i, x_{i-1}).$$

Fixing the index i and setting, for short, $a = f(x_i) - f(x_{i-1})$, $h = u(x_i) - u(x_{i-1})$ and $d = d(x_i, x_{i-1})$, we get the three conditions

$$|a| = d, \ |a+h| \leq d, \ |a-h| \leq d$$

which imply necessarily $h = 0$.

Since $u(x_0) = 0$ it follows that u vanishes along that path and so $u(x) = 0$. In turn, this implies $u(x) = 0$ for each $x \in X$, as desired.

As far as it concerns the necessity condition, we shall treat the case where the finite space is a graph. Assume that the condition stated fails for some $\bar{x}$ and for every path $x_0, x_1, \ldots, x_{n-1} = \bar{x}$ for which the points x_i, x_{i+1} are close.

Define the following function $u : X \longrightarrow \mathbb{R}$

$$u(x) = \min \left[\sum_{i=1}^{m-1} d(z_i, z_{i-1}) - \sum_{i=1}^{m-1} |f(z_i) - f(z_{i-1})| \right] .$$

where the minimum is taken over all the sequences of close vertices from x_0 to x.

Clearly, $u(\bar{x}) > 0$ in that, by construction,

$$\sum_{i=1}^{n-1} |f(x_i) - f(x_{i-1})| < \sum_{i=1}^{n-1} d(x_i, x_{i-1})$$

holds for all the paths linking x_0 and $\bar{x}$.

Take now any pair $x, y \in X$ of close points. Any sequence linking x_0 and x can be extended to a sequence linking x_0 and y, by adding the additional point y preserving the property of being a sequence of close points.

It follows

$$u(y) \leq u(x) + d(x, y) - |f(x) - f(y)|$$

for each pair of close vertices. Switching x and y, we get by some algebra

$$|u(y) - u(x)| + |f(x) - f(y)| \leq d(x, y)$$

that in turn implies the two functions $f \pm u$ are 1-Lipschitz for close points. As already discussed this entails that $f \pm u \in \mathrm{Lip}_1^+(d)$ with $u \neq 0$, which is a contradiction because f was assumed to be extremal. □

Let us look at some specific classes of graphs. In the first one, we are dealing with a straightforward application that does not require a further proof.

Proposition 3 *In a weighted tree, a function u is extremal in the unit ball of* $Lip_1^+(d)$ *if, and only if,* $|u(x) - u(y)| = d(x, y)$ *for each pair of adjacent vertices.*

Next consider a set X equipped by the discrete distance. In another words, $X = K_n$ is the unweighted complete graph. Observe that the distance admits the realization $d = \frac{1}{2} \sum_{x \in X} \delta_{\{x\}}$.

Proposition 4 *Let d be the discrete distance. The function* $u \in \operatorname{ext} \mathrm{Lip}_1^+(d)$ *if, and only if,* $u = \pm I_Y$, *where* I_Y *is the indicator function of a nonempty subset* $Y \subseteq X \setminus x_0$.

Proof The functions $\pm I_Y$ are surely extremal. Actually, if we pick $x \in Y$, the path $x_0 x$ satisfies the sufficient conditions of Theorem 2. While, if $x \notin Y$, the path $x_0 x_1 x$ does, where x_1 is any point of Y.

To show that there are no others, we put in place the necessary conditions. If $x_0 x_1, ..x_n$ is any sequence claimed in Theorem 2, then $x_1 = \pm 1$. Since $u \in \mathrm{Lip}_1^+(d)$, we infer that either $0 \leq u(x) \leq u(x_1) = 1$ or $-1 \leq u(x) \leq 0$.

Consider the positive case, the other is similar. Suppose by contradiction that the function takes at least three values. Hence

$$0 = u(x_0) < u(\bar{x}) < u(x_1) = 1\ .$$

holds for some $\bar{x} \in X$. But then for any path linking x_0 and $\bar{x}$ the necessary condition of Theorem 2 fails. □

One of the tools useful to study the case of trees relies on the detection of the extreme points of the unit ball of $\mathrm{Lip}^+(d)$, as described in the next proposition.

Proposition 5 *Let* $T = (X, w)$ *be a rooted tree. A function* $u \in \operatorname{ext} \mathrm{Lip}_1^+(d)$ *if and only if it is of the type*

$$u_\epsilon(x) = \sum_{y \leq x} d(y, y^+)\epsilon(y)\,, \tag{8}$$

for all $x \in X \setminus x_0$, *and* $u_\epsilon(x_0) = 0$ *otherwise, where* ϵ *is any given* $\epsilon : X \setminus x_0 \longrightarrow \{-1, 1\}$.

Proof It suffices to remark that the functions u_ϵ are recursively generated by the equation

$$u_\epsilon(y) = u_\epsilon(x) + d(x, y)\epsilon(y), \quad \forall y \in \operatorname{child}(x) \tag{9}$$

with initial condition $u(x_0) = 0$. The desired result is a consequence of Proposition 3. □

3.3 A support property

Theorem 3 *Assume that the distance d is generated by a weighted graph. The class of functions $a(x, y)$, employed in eq. (5), can be restricted to the one satisfying the following two conditions:*

i) if $a(x, y) \neq 0$, then x and y are close.
ii) the graph $\{(x, y) \mid a(x, y) \neq 0\}$ has no cycle.

Proof Item i) Let us first assume that in eq. (4) there is a non-zero term $a(x, y)(\delta_x - \delta_y)$, where x and y are not adjacent. Let $x_1, x_2, \ldots, x_n$ be a geodesic path joining $x = x_1$ to $y = x_n$. By eq. (7),

$$d(x_1, x_n) = \sum_{i=1}^{n-1} d(x_i, x_{i+1}) \quad \text{and} \quad \delta_x - \delta_y = \sum_{i=1}^{n-1} (\delta_{x_i} - \delta_{x_{i+1}})$$

Therefore, if the addendum $a(x, y)(\delta_x - \delta_y)$ is replaced by

$$a(x, y) \sum_{i=1}^{n-1} (\delta_{x_i} - \delta_{x_{i+1}}) \, ,$$

the contribution to the norm will remain the same, since

$$|a(x, y)| \sum_{i=1}^{n-1} d(x_i, x_{i+1}) = |a(x, y)| \, d(x, y).$$

Consequently, the term $a(x, y)(\delta_x - \delta_y)$ may be removed, whenever x and y are not adjacent.

Suppose now that x and y are adjacent but not close. That means that a geodesic path $x_1, x_2, \ldots, x_n$ exists with $x = x_1$ and $y = x_n$, and the strict inequality $d(x, y) > \sum_{i=1}^{n-1} d(x_i, x_{i+1})$ holds. In this case,

$$|a(x, y)| \sum_{i=1}^{n-1} d(x_i, x_{i+1}) < |a(x, y)| \, d(x, y) \, .$$

Once again the term may be removed, if the pair of vertices is not close.

Item ii) The argument will unfold along the following lines. Let

$$\xi = \sum_{x,y \in X} \widetilde{a}(x, y)(\delta_x - \delta_y)$$

be an optimal representation of ξ. That is, let $\|\xi\|_{Æ} = \sum_{x,y \in X} |\widetilde{a}(x, y)| \, d(x, y)$.

In addition, let us suppose that it is a minimal, i.e, it contains the minimum number of non-vanishing coefficients $a(x, y)$. A minimal representation does exist but clearly it is not a unique one. For instance, as $a(x, y)(\delta_x - \delta_y) = -a(x, y)(\delta_y - \delta_x)$, any change of signs for the coefficients produces another optimal minimal representation.

Suppose by contradiction that in a minimal representation of ξ there is a set of non-zero coefficients $\widetilde{a}(x, y)$, $(x, y) \in \mathcal{S}$, whose graph $(S, \mathcal{S})$ is a cycle. We can write

$$\xi = \sum_{(x,y) \in \mathcal{S}} \widetilde{a}(x, y)(\delta_x - \delta_y) + A,$$

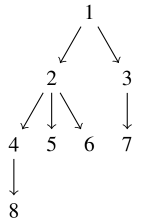
where A includes all the other remaining terms.

Moreover, by arranging signs of coefficients, we can suppose that the cycle is directed, so that we have $\sum_{(x,y) \in \mathcal{S}} (\delta_x - \delta_y) = 0$, then

$$\xi = \sum_{(x,y) \in \mathcal{S}} [\widetilde{a}(x, y) - t](\delta_x - \delta_y) + A$$

holds for any scalar t. It follows that

$$\|\xi\|_{Æ} = \inf_{t \in \mathbb{R}} \sum_{(x,y) \in \mathcal{S}} |\widetilde{a}(x, y) - t| \, d(x, y) + A \, .$$



	1	2	3	4	5	6	7	8
1	1	1	1	1	1	1	1	1
2	0	1	0	1	1	1	0	1
3	0	0	1	0	0	0	1	0
4	0	0	0	1	0	0	0	1
5	0	0	0	0	1	0	0	0
6	0	0	0	0	0	1	0	0
7	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	1

$$\begin{array}{l} \|\xi\|_{Æ} = \\ 0+ \\ w_{21}\,|\xi_2+\xi_4+\xi_5+\xi_6+\xi_8|+ \\ w_{31}\,|\xi_3+\xi_7|+ \\ w_{42}\,|\xi_4+\xi_8|+ \\ w_{52}\,|\xi_5|+ \\ w_{62}\,|\xi_6|+ \\ w_{73}\,|\xi_7|+ \\ w_{84}\,|\xi_8| \end{array}$$

Fig. 1 *Left panel:* A rooted tree in which each vertex is ordered by its distance from the root. *Middle panel:* The adjacency matrix E^* of the descendent relation. *Right panel:* The Arens-Eells norm derived from the adjacency matrix E^*.

On the other hand, the scalar function

$$t \mapsto \sum_{(x,y)\in\mathcal{S}} |\widetilde{a}(x,y) - t|\, d(x,y)$$

is convex and piece-wise linear. Consequently, it attains its minimum value at some point $t = \widetilde{a}(\bar{x},\bar{y})$, with $(\bar{x},\bar{y}) \in \mathcal{S}$. Hence it would be

$$\|\xi\|_{Æ} = \sum_{(x,y)\in\mathcal{S}\setminus(\bar{x},\bar{y})} |\widetilde{a}(x,y) - \widetilde{a}(\bar{x},\bar{y})|\, d(x,y) + A$$

but it contains a smaller number of non-zero coefficients, a contradiction. □

4 Trees

This section is devoted to the specific analysis of the K-distance for weighted trees. Let $T = (X, w)$ be a rooted tree, where x_0 denotes the root. Each vertex $x \in X$ can be classified according to its depth, that is, its (un-weighted) distance from that root.

Define the cumulative function defined by

$$\Xi(x) = \sum_{y\geq x} \xi(y)\,,$$

for $x \in X$ and $\xi \in M_0(X)$. Note that $\Xi(x_0) = \sum_{y\in X}\xi(y) = 0$. See also the right panel of fig. 1. Recall that $w(x,y) = d(x,y)$ holds for the edges xy of a tree. Moreover, the shortest path is unique and it is the same under all weights and for all selections of a vertex as a root.

Theorem 4 *Let $T = (X, w)$ be a weighted rooted tree. Then*

$$\|\xi\|_{Æ} = \sum_{x\in X\setminus x_0} d(x,x^+)\,|\Xi(x)|\ , \tag{10}$$

for every $\xi \in Æ\,(X)$. An equivalent expression of the norm is

$$\|\xi\|_{Æ} = \sum_{y\in X}\xi(y) \sum_{x_0\neq x\leq y} d(x,x^+)\,\mathrm{Sgn}\,\Xi(x)\,, \tag{11}$$

where $\mathrm{Sgn}(\cdot) \in \{-1,1\}$ is a sign function, taking value at zero either $\mathrm{Sgn}(0) = +1$ or $\mathrm{Sgn}(0) = -1$.

Formula eq. (10) was given by Mendivil (2017). The interest of our presentation relies on the methods of proofs. Actually, we provide two different proofs for this theorem, each of them quite instructive in itself. The first one uses an algebraic argument based on the key result of Theorem 3. The alternative proof relies on the characterization of the extremal points of $\mathrm{Lip}_1^+(d)$.

Proof (First proof of Theorem 4) If the graph is a tree, Theorem 3 implies that every $\xi \in Æ(X)$ can be written as

$$\xi = \sum_{x \in X \setminus x_0} a(x, x^+)(\delta_x - \delta_{x^+}) \, . \tag{12}$$

In such a case, the above equation can be uniquely solved for the $a(x, x^+)$'s. Actually, from eq. (12) we get

$$\xi(z) = a(z, z^+) - \sum_{x \in \mathrm{child}(z)} a(x, z)$$

for all $z \neq x_0$ and under the convention $\sum_{x \in \emptyset} a(x, z) = 0$. Hence,

$$\Xi(x) = \sum_{z \succeq x} \xi(z) = \sum_{z \succeq x} a(z, z^+) - \sum_{z \succeq x} \sum_{x \in \mathrm{child}(z)} a(x, z) = a(x, x^+) \, .$$

From this equality and the definition of the Arens-Eells norm, eq. (10) follows.

With regard to eq. (11), it suffices to interchange the order of two summations in eq. (10). More precisely,

$$\|\xi\|_{Æ} = \sum_{x \in X \setminus x_0} d(x, x^+) \, \mathrm{Sgn}(\Xi(x)) \sum_{y \succeq x} \xi(y) = \sum_{x \in X \setminus x_0} d(x, x^+) \, \mathrm{Sgn}(\Xi(x)) \sum_{y \in X} \xi(y) I_A(x, y) \, ,$$

where I_A is the indicator function: $I_A(x, y) = 1$ if $y \succeq x$ and $I_A(x, y) = 0$, otherwise.

Therefore

$$\|\xi\|_{Æ} = \sum_{x \in X \setminus x_0} \sum_{y \in X} d(x, x^+) \;\; \mathrm{Sgn}(\Xi(x)) \xi(y) I_A(x, y) = \sum_{y \in X} \xi(y) \sum_{x \in X \setminus x_0} d(x, x^+) \;\; \mathrm{Sgn}(\Xi(x)) I_A(x, y) \, ,$$

which is eq. (11). □

Proof (Second proof of Theorem 4)

Thanks to the characterization of the extreme points of $\mathrm{Lip}_1^+(d)$ stated in Proposition 5, we want to maximize the functional

$$\sum_{y \in X \setminus x_0} \xi(y) u_\epsilon(y) = \sum_{y \in X \setminus x_0} \xi(y) \sum_{y \succeq x} d(x, x^+) \epsilon(x),$$

over all $\epsilon : X \setminus x_0 \to \{-1, 1\}$.

On the other hand, by interchanging the order between the two summations,

$$\sum_{y \in X \setminus x_0} \xi(y) \sum_{y \succeq x} d(x, x^+) \epsilon(x) = \sum_{x \in X \setminus x_0} d(x, x^+) \epsilon(x) \Xi(x)$$

and so the maximum value is attained when $\epsilon(x) = \mathrm{Sgn}\, \Xi(x)$ for all $x \in X \setminus x_0$. Hence the maximum value will be given by eq. (10). □

Remark 1 By inspecting the second proof, we find easily the dual elements aligned to the points $\xi \in Æ(X)$. Namely, for every $\xi \in Æ\,(X)$ it holds $\langle \xi, \bar{u} \rangle = \|\xi\|_{Æ}$, where $\bar{u} \in \mathrm{ext}\, \mathrm{Lip}_1^+(d)$ is given by

$$\bar{u}(y) = \sum_{x \preceq y} d(x, x^+) \, \mathrm{Sgn}\, \Xi(x), \quad \forall y \in X. \tag{13}$$

Multiple solutions to the alignment condition $\langle \xi, u \rangle = \|\xi\|_{Æ}$ will be due to the indeterminacy of the Sgn function for the vertices x at which $\Xi(x)$ vanishes.

The construction of the extreme points for trees made in Proposition 5 suggests the following extension.

Associate with every function ϕ defined on $X \setminus x_0$, the following Kantorovich potential

$$u_\phi(y) = \sum_{x \preceq y} d(x, x^+) \phi(x)$$

defined on the vertices of the tree.

Proposition 6 *The mapping* $\phi \mapsto u_\phi$*, sending* $l_\infty(X \setminus x_0)$ *onto* $\mathrm{Lip}^+(d)$*, is an isometric isomorphism. Its inverse,* $\Delta : \mathrm{Lip}^+(d) \mapsto l_\infty(X \setminus x_0)$ *is given by*

$$(\Delta u)(x) = \frac{u(x) - u(x^+)}{d(x, x^+)}$$

with $(\Delta u)(x_0) = 0$.

Proof Clearly the map is linear. Let us check that it is an isometry. Consider adjacent vertices y_1, y_2, with $y_1 \succeq y_2$. Then,

$$u_\phi(y_1) = d(y_1, y_2)\phi(y_1) + u_\phi(y_2). \tag{14}$$

Hence, $u_\phi(y_1) - u_\phi(y_2) = d(y_1, y_2)\phi(y_1)$, and so $\left\|u_\phi\right\|_{Lip} \leq \|\phi\|_\infty$.

On the other hand, if y_1 is an element in $X \setminus x_0$ for which $\phi(y_1) = \pm \|\phi\|_\infty$, then the relation eq. (14) implies the equality $\left\|u_\phi\right\|_{\mathrm{Lip}} = \|\phi\|_\infty$. We have so proved that the mapping is an injective isometry.

Denoting by Ψ the direct map $\phi \mapsto u_\phi$, we have

$$(\Psi \circ \Delta)u(y) = \sum_{y \succeq x \neq x_0} d(x, x^+)\frac{u(x) - u(x^+)}{d(x, x^+)} = u(y).$$

Consequently, Ψ is onto with inverse given by Δ. □

Let us outline a few consequences that can be derived from the construction of the previous map.

i) Proposition 6 provides a simple proof that $\mathrm{Lip}^+(d)$ is a dual space. Actually, $\mathrm{Lip}^+(d) \simeq l_1(X \setminus x_0)^*$.
ii) For every tree with n vertices, 2^{n-1} is the number of the extreme points of the unit ball of $\mathrm{Lip}^+(d)$. Actually, it is the image of the unit cube $\|x\| \leq 1$ of $l_\infty(X \setminus x_0)$.
iii) Interestingly, the inverse map Δ of $\phi \longrightarrow u_\phi$ is closely related to De Leeuw's map (see Weaver (2018)) which associates with every Lipschitz function $f : X \longrightarrow \mathbb{R}$, the function

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

defined for $x \neq y \in X$.

Example 1 (Barycenter)

If μ is a probability function defined on the vertices X of a weighted tree, a barycentre is a vertex $\hat{x} \in X$ such that the K-distance between μ and the delta function of that vertex is minimal, namely

$$\|\delta_{\hat{x}} - \mu\|_{\text{Æ}} = \min_{x \in X} \|\delta_x - \mu\|_{\text{Æ}} \ .$$

Barycenters of probability measures on metric spaces are used in various statistical applications. See Evans and Matsen (2012) for the specific example of weighted trees.

If $\bar{x}$ denotes the root of the tree, eq. (10) yields

$$\|\delta_{\bar{x}} - \mu\|_{\text{Æ}} = \sum_{x \in X \setminus \bar{x}} d(x, x^+) \sum_{x \preceq y} \mu(y),$$

that can be simplified by interchanging the two summations. More specifically, we have

$$\|\delta_{\bar{x}} - \mu\|_{\text{Æ}} = \sum_{x \in X \setminus \bar{x}} \sum_{y \in X} d(x, x^+)\mu(y)I_A(x, y),$$

where I_A is the indicator function with $I_A(x, y) = 1$ if $x \preceq y$ and $I_A(x, y) = 0$ otherwise.

Therefore,

$$\|\delta_{\bar{x}} - \mu\|_{\text{Æ}} = \sum_{y \in X} \mu(y) \sum_{x \in X \setminus \bar{x}} d(x, x^+)I_A(x, y) = \sum_{y \in X} \mu(y) \sum_{x \preceq y} d(x, x^+) = \sum_{y \in X} \mu(y)d(y, \bar{x}) = \mathbb{E}_\mu\left[d(\cdot, \bar{x})\right] \ .$$

Consequently, the barycenter x_B will be that vertex that minimizes the μ-mean distance of vertices from itself, i.e.,

$$x_B = \arg\min_{\bar{x} \in X} \mathbb{E}_\mu\left[d(\cdot, \bar{x})\right].$$

5 Beyond trees

In this last section, we study a few extensions along two distinct lines of research both suggested by the previous results on trees.

A first extension is based on computing the K-distance through the spanning trees of a given arbitrary graph. Mendivil (2017) has suggested a different approach, based on starting from a single spanning tree and then reach the full graph via a quotient map.

The distance induced by trees is a rigid ℓ_1-embeddable metric. A second development is thus related to the study the extend to which the results for trees can be generalized to other types of ℓ_1-embeddable metrics.

5.1 Spanning trees

If $G = (X, \mathcal{E})$ is a connected graph, a spanning tree of G is a tree $T = (X, \mathcal{T})$ with $\mathcal{T} \subset \mathcal{E}$. In other words, T is a sub-graph of G with the same vertex set as G and with the minimum number of edges that allows connection. See, for example, § 1.2 of Bollobás (1998).

The inclusion relation $\mathcal{T} \subset \mathcal{E}$ implies the inequality $d \le d_T$ between the two distances d and d_T induced by G and T, respectively. Hence, $\|\xi\|_G \le \|\xi\|_T$ holds for the two graphs.

Denoting by $\mathrm{ST}(G)$ the totality of the spanning trees of G, it follows that

$$\|\xi\|_G \le \min_{T \in \mathrm{ST}(G)} \|\xi\|_T \ .$$

In order to show that the above inequality is in fact an equality, we need the following lemma about growing a forest to a tree. The result is provided by the Kruskal algorithm, Kruskal (1956). See also (Bollobás 1998, p. 10).

Lemma 1 *Let G be a connected graph and F be a forest contained in G. There exists a spanning tree of G which extends F.*

Theorem 5 *The Arens-Eells norm of any connected graph G is the envelope of the norms of its spanning trees. That is,*

$$\|\xi\|_G = \min_{T \in \mathrm{ST}(G)} \|\xi\|_T \ .$$

Proof Theorem 3 implies that the Arens-Eells norm of ξ is

$$\|\xi\|_{Æ} = \sum_{(x,y)\in\mathcal{F}} |a(x,y)|\, d(x,y)$$

where $F = (X, \mathcal{F})$ is an a-cyclic subgraph of $G = (X, \mathcal{E})$. In other words, $(X, \mathcal{F})$ is a forest. By Lemma 1, there is a spanning tree $T = (X, \mathcal{T})$ extending such a forest. If we enlarge the domain of the functions $a(x, y)$ to $(x, y) \in \mathcal{T}$, by assigning the value $a(x, y) = 0$, outside $\mathcal{F}$, we can re-write the equation above as

$$\|\xi\|_{Æ} = \sum_{(x,y)\in\mathcal{T}} |a(x,y)|\, d(x,y) = \sum_{(x,y)\in\mathcal{T}} |a(x,y)|\, w(x,y),$$

where the last equality follows from item i) of Theorem 3, since the pair of vertices x and y are close, as long as $a(x, y) \neq 0$.

To conclude,

$$\|\xi\|_{Æ} = \sum_{(x,y)\in\mathcal{T}} |a(x,y)|\, w(x,y) \ge \|\xi\|_T \ge \min_{T \in \mathrm{ST}(G)} \|\xi\|_T$$

that proves our assertion. □

5.2 A worked out example: cycle graphs

It should be of some interest to solve by hand a few examples of what was stated in Theorem 5. The cyclic case has already been analyzed by Cabrelli and Molter (1995) as well as Mendivil (2017) but by quite different techniques.

A labelled weighted cyclic graph (or circuit) of order n, denoted by C_n, is the graph

$$1 \to 2 \to \cdots \to n \to 1$$

consisting of a unique cyclic path. Set $d_i = d(i, i+1)$ the distance between the two adjacent vertices i and $i+1$.

Clearly the cycle graph C_n admits n spanning trees $\{T_i\}_{i=1}^n$ obtained by ruling out each single edge of C_n.

Next proposition provides explicitly the Arens-Eells norm $\|\cdot\|_{C_n}$ for the cycle C_n as well as a constructive proof of the envelope property.

Proposition 7 *Let C_n be the cycle graph of order n with vertices $1, 2, \ldots, n$ and edges $\{i, i+1\}$, $n+1 = 1$. Define the real function*

$$\Phi(t) = \sum_{i=1}^{n} |t - \xi_1 - \xi_2 - \cdots - \xi_i|\, d_i\,, \quad t \in \mathbb{R}\,.$$

Then, for each $\xi \in \text{Æ}(X)$,

$$\|\xi\|_{C_n} = \min_{t\in\mathbb{R}} \Phi(t) = \min_{i=1,2,..,n} \Phi(\xi_1 + \xi_2 + \cdots + \xi_i) = \min_{i=1,2,..,n} \|\xi\|_{T_i}$$

where $T_i \in \mathrm{ST}(C_n)$. Specifically, $\Phi(\xi_1 + \xi_2 + \cdots + \xi_i)$ is the norm for the tree obtained by removing the edge $\{i-1, i\}$.

As observed by Mendivil (2017), the value t that minimizes Φ is the weighted median value of the distribution ξ.

Proof Thanks to Item i) of Theorem 3, the restriction of the elements $a(x, y)$ leads to the representations

$$\xi = a_1(\delta_1 - \delta_2) + a_2(\delta_2 - \delta_3) + \cdots + a_n(\delta_n - \delta_1)$$

with $a_i \in \mathbb{R}$. By inverting the previous relation and introducing the parameter $t = -a_n$, we get easily that

$$a_i = t - \xi_1 - \xi_2 - \cdots - \xi_i$$

for $i = 1, 2, \ldots, n$. This implies that every vector ξ admits ∞^1-many representations, and Arens-Eells formula eq. (5) for the norm becomes $\inf_{t\in\mathbb{R}} \Phi(t)$.

Of course, this piece-wise linear and convex function Φ reaches the minimum value at one of the n points $t = \xi_1 + \xi_2 - \cdots + \xi_i$, $(i = 1, 2, \ldots, n)$, and so also the second formula is checked.

It remains to show that the values $\Phi(\xi_1 + \xi_2 + \cdots + \xi_i)$ are nothing but the Arens-Eells norms of the spanning trees of C_n.

Fix an index j and evaluate the function Φ at the point $\xi_1 + \xi_2 + \cdots + \xi_j$, then

$$\Phi(\xi_1 + \xi_2 + \cdots + \xi_i) = \sum_{i=2}^{i=j} \left|\xi_i + \cdots + \xi_j\right| d_{i-1} + \sum_{k=1}^{k=n-j} \left|\xi_{j+1} + \cdots + \xi_{j+k}\right| d_{j+k}$$

If now we get rid of variable ξ_j, by means of the relation $\xi_j = -\sum_{i\neq j} \xi_i$, it is not difficult to check that $\Phi(\xi_1 + \xi_2 + \cdots - \xi_i)$ turns out to be the norm of the linear tree

$$j \to j+1 \to \cdots \to n \to 1 \to \cdots \to j-1$$

by taking $j-1$ as root. □

The method employed in Proposition 7, might be duplicated for other graphs for which $\#X = \#\mathcal{E} = n$, like in the cycle graphs. However, the case $\#\mathcal{E} > n$ is more interesting and clearly the function Φ, in this case, will be no longer a scalar one.

References

Bollobás B (1998) Modern graph theory, Graduate Texts in Mathematics, vol 184. Springer-Verlag

Cabrelli CA, Molter UM (1995) The Kantorovich metric for probability measures on the circle. J Comput Appl Math 57(3):345–361, DOI 10.1016/0377-0427(93)E0213-6, URL https://doi.org/10.1016/0377-0427(93)E0213-6

Evans SN, Matsen FA (2012) The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. J R Stat Soc Ser B Stat Methodol 74(3):569–592, DOI 10.1111/j.1467-9868.2011.01018.x, URL https://doi.org/10.1111/j.1467-9868.2011.01018.x

Farmer JD (1994) Extreme points of the unit ball of the space of Lipschitz functions. Proc Amer Math Soc 121(3):807–813, DOI 10.2307/2160280, URL https://doi.org/10.2307/2160280

Gini C (1914) Di una misura della dissomiglianza di due gruppi di quantità e delle sue applicazioni allo studio delle relazioni statistiche. Atti R Ist Veneto Sc Lett Arti LXXIV:185–213

Kruskal JB Jr (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. Proc Amer Math Soc 7:48–50, DOI 10.2307/2033241, URL https://doi.org/10.2307/2033241

Mendivil F (2017) Computing the Monge-Kantorovich distance. Comput Appl Math 36(3):1389–1402, DOI 10.1007/s40314-015-0303-7, URL https://doi.org/10.1007/s40314-015-0303-7

Santambrogio F (2015) Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling. Birkhäuser

Smarzewski R (1997) Extreme points of unit balls in Lipschitz function spaces. Proc Amer Math Soc 125(5):1391–1397, DOI 10.1090/S0002-9939-97-03866-5, URL https://doi.org/10.1090/S0002-9939-97-03866-5

Villani C (2008) Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, URL https://books.google.it/books?id=hV8o5R7_5tkC

Weaver N (2018) Lipschitz algebras. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, second edition