ISSN 2279-9362

Collegio Carlo Alberto

Kantorovich Distance on a Weighted Graph

Luigi Montrucchio and Giovanni Pistone

No. 590
June 2019

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

KANTOROVICH DISTANCE ON A WEIGHTED GRAPH

LUIGI MONTRUCCHIO

Collegio Carlo Alberto, Piazza V. Arbarello 8, 10122 Torino Italy

GIOVANNI PISTONE

de Castro Statistics, Collegio Carlo Alberto, Piazza V. Arbarello 8, 10122 Torino Italy

Abstract. The computation of the Kantorovich distance (1-Wasserstein distance) on a finite state space may be an hard problem in the case of a general distance. In this paper, we derive a simple closed form for the geodesic distance on a tree. Moreover, when the ground distance is defined by an arbitrary graph, we show that the Kantorovich distance is the minimum of the distances on the spanning trees.

Contents

E-mail addresses: luigi.montrucchio@unito.it, giovanni.pistone@carloalberto.org.
Date: 2019-06-20 15:47:44Z.

1. Introduction

Optimal Transport (OT), as formalized by Kantorovich (1942), is currently a very active field of research, both from the mathematical point of view and as a tool in many applications. A number of monographs have been recently published, among which, we like notably to mention C. Villani [19, 18], L. Ambrosio *et al.* [1], and F. Santambrogio [15]. The case of a finite state space is discussed in detail by G. Peyré and M. Cuturi [13]. We refer the reader to those monographs for full references to primary bibliographical sources.

Let us first recall the definition of K-distance and introduce our notations. Let X be a set with n points, and $\Delta(X)$ be the simplex of probability functions on X. We assume that the sample space X is equipped with a distance d.

Given probability functions $\mu, \nu \in \Delta(X)$, the joint probability function $\gamma \in \Delta(X \times X)$ is a *coupling* of μ and ν if μ and ν are the two margins of γ, respectively. Namely, the set of all couplings of μ and ν is the subset of $\Delta(X \times X)$ defined by

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

This set is a nonempty bounded polyhedron in $\mathbb{R}^{X \times X}$, hence it is a polytope i.e., the convex combination of its vertices.

The *Kantorovich distance* (the K-distance) is defined by

$$d(\mu,\nu) = \inf \left\{ \sum_{x,y \in X} d(x,y)\gamma(x,y) \,\middle|\, \gamma \in \mathcal{P}(\mu,\nu) \right\} . \tag{2}$$

Since the set of couplings in Eq. (1) is a polytope, there exists a finite set of extreme points $\widetilde{\gamma}$ which are optimal couplings, i.e., such that $d(\mu,\nu) = \sum_{x,y} d(x,y)\widetilde{\gamma}(x,y)$.

In this paper, we first review the properties of the support of the couplings, that is, the directed graph whose edges are $\{x \to y \,|\, \gamma(x,y) > 0\}$. In our finite space setting, a coupling cannot be realized through a permutation $\sigma \colon X \to X$, unless particular assumptions hold for the margins. This would be the so-called Monge problem, whose solution requires the space X to be continuous. Nevertheless, an extreme coupling, in particular, an optimal one, has a "small" support.

When the two margins μ and ν are both strictly positive, the product coupling $\mu \otimes \nu$ has a support which is a complete graph, but there are always couplings with a much smaller support. For example, if we represent the coupling as a matrix, we can fill the first row, left to right, by setting each $\gamma(1, j)$ to the maximum value compatible with the current marginal constraint, and proceed on, in the same way, to fill top to bottom the rows. This is the algorithm described by [13, §3.4.2] to construct extremal solutions of the polytope $\mathcal{P}(\mu,\nu)$. In general, the support of an extreme coupling contains at most $2n - 1$ edges, see [13, Prop. 3.4].

We focus on the special assumption that the underlying sample space is endowed with the so-called graphic metric defined by a weighted (un-directed) graph. Namely: the sum of weights along a path defines its length; the distance between two points is the length of the shortest path linking them. Intuitively, one expects the optimal transport to flow along the ground "geodesics".

The OT problem is a Linear Programming (LP) problem and, as such, it admits a dual formulation. More precisely, Kantorovich and Rubinstein (1958) found that the dual problem is best expressed as a maximization problem on a class of Lipschitz functions. It follows that the distance induced on probability functions by the solution of the OT problem happens to be a restriction of a norm of $\xi = \mu - \nu$. In turn, this norm is found to be equivalent to a norm defined by R. Arens and J. Eells [3]. We refer to the monograph by N. Weaver [20] for the general theory of the duality of the space of Lipschitz functions.

After introducing Arens-Eells norm, the K-distance $d(\mu, \nu)$ admits the two seemingly different representations:

$$d(\mu, \nu) = \sum_{x,y \in X} d(x, y)\widetilde{\gamma}(x, y) = \sum_{x,y \in X} |\tilde{a}(x, y)|\, d(x, y) \ ,$$

where $\widetilde{\gamma}$ is an optimal coupling and $\widetilde{a}$ is chosen to satisfy $\mu - \nu = \sum_{x,y} \widetilde{a}(x, y)(\delta_x - \delta_y)$. Although there is no obvious relation between them, the issue will be investigated in some detail.

A closed form expression for the K-distance has been known even before the formalization by Kantorovich in the case of the distance induced by a total order. In such a case, the K-distance reduces to the weighted L^1-distance between the cumulative probability functions.

We extend this result to arbitrary trees and provide a generalized closed-form expression. As a matter of fact, three different proofs of this result will be offered. Such a formula has interests in its own, as we shown by providing a couple of applications. K-distance on trees has been considered in B. Kloeckner [11], and by M. Sommerfeld and A. Munk [17].

More importantly, the existence such a closed form solution prompts for an inquiry about the use of spanning trees to compute the K-distance for general graphs. In fact, we state that the distance is the minimum distance among the spanning trees of the given graph. We also provide a few examples to explain the algorithm.

1.1. **Content of the paper.** In order to highlight the specific content of the present paper, some propositions have been promoted to theorems. No valuation of relative importance is intended.

In Sec. 2, we group a few relevant results on the primal problem: the support of optimal solutions and the main properties of the K-distance. Our terminology for graph theory is collected in App. A, while we refer to the monograph by B. Bollobás [5] for further results on this subject. The textbook by A. Barvinok [4] contains topics in convexity used in the present paper.

The following Sec. 3 contains a summary of the Kantorovich-Rubinstein duality theory. In turn, this leads in Sec. 4 to an exposition of the Arens-Eells theory on the space of Lipschitz functions and to the important relation between the K-distance and the Arens-Eells norm.

Sec. 5 concerns with some results on weighted graphs. The main result is Th. 3, which shows the deep relation between weighted graphs and the support as defined through Arens-Eells norms.

Sec. 6 is entirely devoted to the analysis of the class of trees. In Th. 3 a closed-form formulation of K-distance is obtained, while Th. 5 provides results on the optimal solutions. There are also described two applications related to this formula: the computation of the barycentre (Sec. 6.1), and the distance between probabilities on a probability tree (Sec 6.2).

We present in Sec. 7 the main applicable result of the present paper which concerns the use of spanning trees for the computation of the K-distance in arbitrary graphs (see Th. 8). This section ends with a few toy examples that illustrate that theoretical result.

Sec. 8 contains a short discussion of possible developments. The present paper does not deal with the issue of the complexity of searching for the optimal spanning tree.

Finally, a result on the K-distance for countable and locally finite trees is gathered in App. B.

We have tried to offer a self-contained and concise presentation. We do not consider here the so-called p-Wasserstein distance that is, the case in which the cost function is $d(x,y)^p$, $p > 1$. Another important topic we do not discuss here is the so-called dynamical approach, as it is typically done when the sample space is continuous, see e.g. [15, Ch. 4–5]. The reader can consult W. Li and G. Montúfar [12] for a discussion on the dynamical approach based on a graph-distance.

2. Kantorovich distance

The compactness of the polytope $\mathcal{P}(\mu,\nu)$ of Eq. (1) implies that the minimum is reached at some coupling $\widetilde{\gamma}$ namely, $d(\mu,\nu) = \sum_{x,y} d(x,y)\widetilde{\gamma}(x,y)$. More precisely, as we have a minimum of a linear function on a polytope, the set of optimal couplings is a face of $\mathcal{P}(\mu,\nu)$.

Couplings, when they are seen as transport plans, are conveniently represented as transitions, $\gamma(x,y) = \mu(x)P(x,y) = \nu(y)Q(y,x)$, where P and Q are Markov matrices. The Markov matrix P maps the initial probability function μ to a final probability function ν. The mapping cannot be realized by a function of X onto itself, unless $\mu \circ \sigma = \nu$ for some permutation σ. This eventuality remains possible when the sample space is continuous and this problem is called Monge transportation problem, see e.g. [18, p. 22]. Cf. A. Brezis [6] on Monge-Kantorovich on a finite set. Remarkably, the support of an optimal coupling is necessarily small, even if it is not the graph of a function.

Transport plans can be composed by composing the transitions. If $\gamma_1 \in \mathcal{P}(\mu,\zeta)$ and $\gamma_2 \in \mathcal{P}(\zeta,\nu)$ with transitions P_1 and P_2, respectively, then

$$\gamma(x,y) = \mu(x) \sum_z P_1(x,z)P_2(z,y) \tag{3}$$

defines a coupling in $\mathcal{P}(\mu,\nu)$.

2.1. **Properties of the K-distance.**

The following two propositions will be directly established, but another, simpler, proof will be available via duality theory, discussed in the next section.

Proposition 1. *The value of the Kantorovich problem in Eq.* (2) *is a distance that extends the given distance on the sample space.*

Proof. Let us first check the extension property. If the marginal probabilities are Dirac's deltas, $\mu = \delta_x$ and $\nu = \delta_y$, then the plan contains only the element $\delta_x \otimes \delta_y$, so that $d(\delta_x, \delta_y) = d(x,y)$. Let us check the distance axioms. If $d(\mu,\nu) = 0$, then $d(x,y)\widetilde{\gamma}(x,y) = 0$ for all couples (x,y) and so $\widetilde{\gamma}(x,y) = 0$ for all $x \neq y$. Hence, $\widetilde{\gamma}$ is supported by the diagonal of $X \times X$ and $\mu = \nu$. Symmetry is clear by definition.

Regarding triangle inequality, consider three probability functions μ,ν,ζ and let γ_1 and γ_2 be the optimal couplings for $d(\mu,\zeta)$ and $d(\zeta,\nu)$, respectively. Denote by γ the composed coupling obtained by Eq (3). We have

$$d(\mu,\nu) \leq \sum_{x,y} d(x,y)\gamma(x,y) = \sum_{x,y,z} d(x,y)\mu(x)P_1(x,z)P_2(z,y) \leq$$
$$\sum_{x,y,z} (d(x,z)+d(z,y))\mu(x)P_1(x,z)P_2(z,y) =$$
$$\sum_{x,z} d(x,z)\gamma_1(x,z) + \sum_{y,z} d(z,y)\gamma_2(z,y) = d(\mu,\zeta) + d(\zeta,\nu) \ .$$

□

The K-distance on the probability simplex is compatible with the affine mixture geometry on the simplex in the sense that mixtures are metric geodesics.

Proposition 2. *Given two probability functions μ and ν, the mixture curve $\mu(t) = (1-t)\mu + t\nu$, $0 \leq t \leq 1$, is a metric geodesic for the d-distance i.e.,*

$$d(\mu(t),\mu(s)) = (t-s)d(\mu,\nu) \ , \quad 0 \leq s \leq t \leq 1 \ . \tag{4}$$

Moreover, is $\widetilde{\gamma}$ is optimal for $d(\mu,\nu)$, then the transport plan defined by

$$\widetilde{\gamma}(x,y;s,t) = (1-t)\mu(x)(x=y) + s\nu(y) + (t-s)\widetilde{\gamma}(x,y)$$

is optimal for $d(\mu(s),\mu(t))$.

Proof. Assume $\mu \neq \nu$, otherwise there is nothing to prove. Note that it is enough to show the inequality

$$d(\mu(t),\mu(s)) \leq (t-s)d(\mu(0),\mu(1)) \ , \quad 0 \leq s \leq t \leq 1 \ .$$

In fact, assume that the strict inequality $d(\mu(t),\mu(s)) < (t-s)d(\mu(0),\mu(1))$ holds for some $s < t$ i.e., $d(\mu(t),\mu(s)) = \alpha d(\mu(0),\mu(1))$ with $0 \leq \alpha < t-s$. In such a case,

$$d(\mu(0),\mu(1)) \leq d(\mu(0),\mu(s)) + d(\mu(s),\mu(t)) + d(\mu(t),\mu(1)) \leq$$
$$(s+\alpha+1-t)d(\mu(0),\mu(1)) < d(\mu(0),\mu(1)) \ ,$$

which is impossible.

We prove now the inequality. For each $\gamma \in \mathcal{P}(\mu,\nu)$ define

$$\gamma_{st}(x,y) = (1-t)\mu(x)(x=y) + s\nu(y)(x=y) + (t-s)\gamma(x,y) \ ,$$

where $(x=y) = 1$ if $x=y$ and 0 otherwise. It is easy to verify that $\gamma_{st} \in \mathcal{P}(\mu(s),\mu(t))$. It follows that

$$d(\mu(s),\mu(t)) \leq$$
$$\inf_{\gamma} \sum_{x,y} d(x,y) \left((1-t)\mu(x)(x=y) + s\nu(y)(x=y) + (t-s)\gamma(x,y)\right) =$$
$$(t-s) \inf_{\gamma} \sum_{x,y} d(x,y)\gamma(x,y) = (t-s)d(\mu,\nu) \ .$$

The last statement follows from a direct check. □

The previous result does not rule out the eventual emergence of multiple geodesics between two points. In fact, it is easy to produce examples of multiple geodesics between two points. The topic of geodesics is related with the so-called dynamical approach to OT. We do not touch at that here.

2.2. **Support of an optimal coupling.** We discuss now necessary conditions for a coupling to be optimal for the K-distance defined in Eq. (2). Given a coupling γ, its support, $\operatorname{Supp}(\gamma)$, is a directed graph with vertices X and such that (x, y) is an edge if, and only if, $\gamma(x, y) > 0$.

Let us call *move* a function $f \colon X \times X \to \{-1, 0, +1\}$ with null margins: $\sum_x f(x, y) = \sum_y f(x, y) = 0$. It is clear that for all coupling $\gamma \in \mathcal{P}(\mu, \nu)$, all move f, and all positive α, it holds $\bar{\gamma} = \gamma - \alpha f \in \mathcal{P}(\mu, \nu)$, as long as $\bar{\gamma} \geq 0$. In such a case, the respective values of γ and $\bar{\gamma}$ are related to each other by

$$\sum_{x,y} d(x,y)\bar{\gamma}(x,y) = \sum_{x,y} d(x,y)\gamma(x,y) - \alpha \left(\sum_{f(x,y)=+1} d(x,y) - \sum_{f(x,y)=-1} d(x,y) \right) ,$$

so that the value of $\bar{\gamma}$ is strictly smaller than the value of γ if

$$\sum_{f(x,y)=+1} d(x,y) > \sum_{f(x,y)=-1} d(x,y) \ . \tag{5}$$

Something like Eq. (5) appears in the following classical definition.

Definition 1. Let d be a distance on X. A set $\Gamma \subseteq X \times X$ is said to be d-cyclically monotone (d-CM) if for every $k \in \mathbb{N}$, every permutation $\sigma \in S_k$ and each finite family of couples $(x_1, y_1), (x_2, y_2), ..., (x_k, y_k) \in \Gamma$, it holds

$$\sum_{i=1}^{k} d(x_i, y_i) \leq \sum_{i=1}^{k} d(x_i, y_{\sigma(i)}). \tag{6}$$

Notice that inequality (6) implies that the value of the move

$$f = \sum_{1}^{k} \delta_{x_i} \otimes \delta_{y_i} - \sum_{1}^{k} \delta_{x_i} \otimes \delta_{y_{\sigma(i)}}$$

is non-positive.

The following proposition is partly expressed in terms of moves.

Proposition 3. *Consider the following properties:*

(0) γ is optimal.
(1) $\operatorname{Supp}(\gamma)$ *is d-cyclically monotone.*
(2) $\operatorname{Supp}(\gamma)$ *does not contain any cycle, that is, it is a forest.*
(3) The number of edges of $\operatorname{Supp}(\gamma)$ *is at most $2n - 1$, where $n = \#X$.*

It holds: (0) $\Leftrightarrow$ (1) $\Rightarrow$ (2) $\Rightarrow$ (3).

Remark. The inverse implication: (2) $\Rightarrow$ (1) does not hold true in general. In fact, one could prove that it remains valid, as long as d is tree-metric.

Proof. Let γ be optimal. Suppose, by contradiction, that the property of d-cyclical monotonicity is violated. Then, there is a sequence of edges in the support for which $\bar{\gamma}$ has a value strictly smaller than γ and where α equals the minimum of the positive values of γ. This proves that (0) $\Rightarrow$ (1).

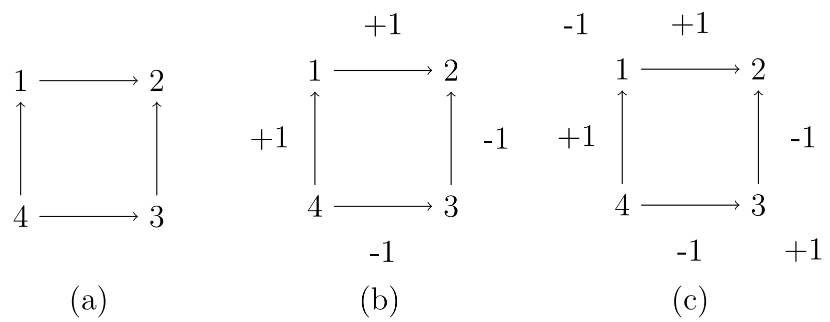


FIGURE 1. Steps in the derivation of Eq. (7)

The converse, (1) ⇒ (0), is well-known to be true. The reader can consult for instance [18, §5].

Now assume (1) to hold. If there were a cycle $x_1, x_2, \ldots, x_{n+1} = x_1$, then, via permutation $i \to i-1, \mod n$, it would give the inequality

$$d(x_1, x_2) + \cdots + d(x_{n-1}, x_1) \leq d(x_1, x_1) + \cdots + d(x_n, x_n) = 0 ,$$

which is absurd and so the implication (1) ⇒ (2) has proven.

Finally, if a graph has no cycle, then there are at most n loops and $n-1$ proper edges, and the last implication is true. □

The following example makes again use of the concept of move in order to state necessary conditions for a coupling to be an extremal point of the polytope.

Example 1. Take the graph with edges $1 \to 2$, $3 \to 2$, $4 \to 3$, $4 \to 1$, as in Fig. 1(a). Choose an arbitrary order e.g., $1 \to 2 \to 3 \to 4 \to 1$. To each arc we associate the value $+1$ if the arc has the chosen order, and the value -1 if not, see Fig. 1(b). Finally, we add loops at 1 and 3 to get alternate signs, see Fig. 1(c).

The algorithm provides a move. Precisely, one takes for each arc (x, y) the term $\delta_x \otimes \delta_y$ with the proper sign. The result is

$$f = -\delta_1 \otimes \delta_1 + \delta_1 \otimes \delta_2 - \delta_3 \otimes \delta_2 + \delta_3 \otimes \delta_3 - \delta_4 \otimes \delta_3 + \delta_4 \otimes \delta_1 \ . \tag{7}$$

The first margin of f is $(-\delta_1 + \delta_4) + (\delta_1 - \delta_3) + (\delta_3 - \delta_4) = 0$. The second margin is $(-\delta_1 + \delta_2) + (-\delta_2 + \delta_3) + (-\delta_3 + \delta_1) = 0$. If γ is a coupling whose support contains the given graph, and in addition both $\gamma(1, 1)$ and $\gamma(3, 3) > 0$, then the function $\gamma + \varepsilon f$ is a coupling for every ε in a neighborhood of 0. It follows that γ cannot be an extreme point of the set of couplings.

The argument illustrated in the above proof that uses a sign function, defined on edges, is a very special application of the theory of cycle and co-cycle spaces of a graph. See [5, §II.3] for un-directed graphs, and G. Pistone and M.-P. Rogantin [14, §3.2] for directed graphs.

It is worth remarking that the argument above might be treated in a slightly different way. Consider the adjacency matrices of the labeled graphs in Fig. 1:

$$\text{(a)} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} , \quad \text{(b)} \begin{bmatrix} 0 & +1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ +1 & 0 & -1 & 0 \end{bmatrix} , \quad \text{(c)} \begin{bmatrix} -1 & +1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & +1 & 0 \\ +1 & 0 & -1 & 0 \end{bmatrix} .$$

A move is a matrix with 0 and ± 1 entries and null row and column sums. The function f of Eq. (7) is a move associated to the matrix (c) above. Adding εf to a coupling produces another coupling, provided positivity constraints are fulfilled. The value of that move is

$$d(1,2) - d(2,3) - d(3,4) + d(1,4)$$

and there is always a choice of sign for ε such that the value of the perturbed coupling $\gamma + \varepsilon f$ does not increase.

There is a rich algebraic theory about moves, with applications in Statistics, see, for example, the monograph of S. Aoki *et al.* [2]. A typical result of such a theory consists in finding generators of the set of moves. For example, our move (c) is the sum of two simpler moves,

$$\begin{bmatrix} -1 & +1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & +1 & 0 \\ +1 & 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} -1 & +1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & +1 & 0 \\ +1 & 0 & -1 & 0 \end{bmatrix} .$$

We will not discuss this topic further.

The following result shows that the directed forest $\operatorname{Supp}(\widetilde{\gamma})$ may contain loops that is, vertices for which $\widetilde{\gamma}(x,x) > 0$.

Proposition 4. *There exist optimal couplings with diagonal terms all strictly positive, provided the margins μ and ν are strictly positive .*

Proof. Assume $\widetilde{\gamma}$ is optimal and that for a vertex, say 1, it holds $\widetilde{\gamma}(1,1) = 0$. Since $\mu(1), \nu(1) > 0$, there exists points, say 2 and 3, for which $\widetilde{\gamma}(1,2), \widetilde{\gamma}(3,1) > 0$. Pick the move

$$f = \delta_1 \otimes \delta_1 - \delta_1 \otimes \delta_2 - \delta_3 \otimes \delta_1 + \delta_3 \otimes \delta_2$$

as well as any number $\alpha \in]0, \widetilde{\gamma}(1,2) \wedge \widetilde{\gamma}(3,1)]$.

It is easily checked that the function $\gamma_\alpha = \widetilde{\gamma} + \alpha f \in \mathcal{P}(\mu,\nu)$ whose value is

$$d(\mu,\nu) - \alpha[d(1,2) + d(3,1) - d(3,2)] \ ,$$

and where $d(1,2)+d(3,1)-d(3,2) \geq 0$ is true by the triangle inequality. The equality must hold, otherwise the value would be strictly smaller than the K-distance. In conclusion, γ_α is an optimal coupling with $\gamma_\alpha(1,1) > 0$ and with all the other diagonal elements equal to those of the original $\widetilde{\gamma}$.

By possibly repeating the argument above, we get hold of an optimal coupling, denoted again by $\widetilde{\gamma}$, with all diagonal elements strictly positive. □

Next proposition asserts that the support of an optimal coupling is generically a connected graph.

Proposition 5. *If the support of the optimal coupling $\widetilde{\gamma}$ is a disconnected graph, with connected components $(X_i, \mathcal{S}_i)$, $i = 1, \dots, k$, then $\mu(X_i) = \nu(X_i)$ for all $i = 1, \dots, k$ and $\widetilde{\gamma} = \sum_{i=1}^k \gamma_i$, where each γ_i is supported by $X_i \times X_i$ and is proportional to an optimal coupling for the conditional margins, $\mu|_{X_i}$ and $\nu|_{X_i}$.*

Proof. To reduce the algebra, consider the case $k = 2$. Assume $\operatorname{Supp}(\widetilde{\gamma})$ to have components $(X_1, \mathcal{S}_1)$ and $(X_2, \mathcal{S}_2)$, respectively. Then, for each $x_1, y_1 \in X_1$ it holds

$$\mu(X_1) = \sum_{x_1 \in X_1} \mu(x_1) = \sum_{x_1,y_1 \in X_1} \widetilde{\gamma}(x_1,y_1) = \sum_{y_1 \in X_1} \nu(y_1) = \nu(X_1) \ ,$$

and also $\mu(X_2) = \sum_{x_2,y_2 \in X_2} \widetilde{\gamma}(x_2,y_2) = \nu(X_2)$.

In this case, the cost of Eq. 2 takes the conditional form

$$\sum_{x_1,y_1 \in X_1} d(x_1,y_1)\widetilde{\gamma}(x_1,y_1) + \sum_{x_2,y_2 \in X_2} d(x_2,y_2)\widetilde{\gamma}(x_2,y_2) =$$
$$\sum_{x_1,y_1 \in X_1} \widetilde{\gamma}(x_1,y_1) \sum_{x_1,y_1 \in X_1} d(x_1,y_1) \left.\widetilde{\gamma}\right|_{X_1 \times X_1}(x_1,y_1) +$$
$$\sum_{x_2,y_2 \in X_1} \widetilde{\gamma}(x_2,y_2) \sum_{x_2,y_2 \in X_2} d(x_2,y_2) \left.\widetilde{\gamma}\right|_{X_2 \times X_2}(x_2,y_2) \ .$$

It is easily seen that $\left.\widetilde{\gamma}\right|_{X_i \times X_i}$ is an optimal coupling of $\left.\mu\right|_{X_i}$ and $\left.\nu\right|_{X_i}$. □

3. Duality and Lipschitz functions

In this section, we review the essentials of the Kantorovich-Rubinstein duality theory for the problem of Eq. (2). For a detailed presentation we refer to [15, §1.2 and §3.1.1]. The duality theory teaches us that the K-distance can be found by maximizing a linear function on the unit ball of the space of Lipschitz functions.

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 $\operatorname{Lip}(d)$. The best constant K is a semi-norm, $\|u\|_{\operatorname{Lip}(d)}$. The set of 1-Lipschitz functions will denoted by $\operatorname{Lip}_1(d)$.

Proposition 6. *Let (X, w) be a weighted graph with associated distance d. A function $u \colon X \to \mathbb{R}$ belongs to $\operatorname{Lip}_1(d)$ if, and only if, $|u(x) - u(y)| \leq d(x,y)$ for each edge $xy \in \mathcal{E}$.*

Notice that we are in fact proving, more generally, that $|u(x) - u(y)| \leq Kd(x,y)$ holds for each edge $xy \in \mathcal{E}$ if, and only if, $\|u\|_{\operatorname{Lip}(d)} \leq K$.

Proof. The condition is clearly necessary. Assume it is satisfied and let $x_0 x_1 \cdots x_n$ be a geodesic path connecting x_0 to x_n. Then

$$|u(x_n) - u(x_0)| \leq \sum_{i=1}^{n} |u(x_i) - u(x_{i-1})| \leq \sum_{i=1}^{n} d(x_i, x_{i-1}) = d(x_0, x_n)$$

and the proof is complete. □

Kantorovich duality theorem below is an application of LP to the problem whose value is the K-distance of Eq. (2). It is of interest that the dual problem can be expressed in terms of Lipschitz functions. The argument below summarizes the presentation of [15].

Let $\phi, \psi \colon X \to \mathbb{R}$ be such that $d(x,y) \geq \phi(x) - \psi(y)$. They provide a lower bound for the K-distance. Actually, for all coupling γ of μ and ν, it holds

$$\sum_{x,y \in X} d(x,y)\gamma(x,y) \geq \sum_{x,y \in X} (\phi(x) - \psi(y))\gamma(x,y) = \sum_{x \in X} \phi(x)\mu(x) - \sum_{y \in X} \psi(y)\nu(y) \ ,$$

so that $d(\mu,\nu) \geq \mathbb{E}_\mu[\phi] - \mathbb{E}_\nu[\psi]$. Notice that the lower bound improves as ϕ increases or ψ decreases.

Assume that the equality holds for a couple $\widetilde{\phi}$, $\widetilde{\psi}$. Then, for each optimal coupling $\widetilde{\gamma}$, we have

$$0 = \sum_{x,y \in X} \left(d(x,y) - (\phi(x) - \psi(y))\right) \widetilde{\gamma}(x,y) \ ,$$

which in turn implies $d(x,y) = \widetilde{\phi}(x) - \widetilde{\psi}(y)$ for all $x, y \in \mathrm{Supp}\,(\widetilde{\gamma})$.

Because of that, we are interested in the following construction that connects the lower bound with Lipschitz functions. If we fix ψ in the inequality $\phi(x) \leq d(x,y) + \psi(y)$, the best possible real function ϕ is

$$\psi^d(x) = \inf_{y \in X} d(x,y) + \psi(y) \ , \tag{8}$$

because $\phi \leq \psi^d$ and $d(x,y) \geq \psi^d(x) - \psi(y)$. We can iterate the same argument, by writing $u = \psi^d$ and $-\psi(x) \leq d(x,y) + (-u(x))$ and optimizing $-\psi$ with $(-u)^d$, in order to get $d(x,y) \geq u(x) + (-u)^d(y)$.

In conclusion, given any bound of the form $d \geq \phi \ominus \psi$, there exists a 1-Lipschitz function u such that

$$d(\mu,\nu) \geq \mathbb{E}_\mu\left[u\right] - \mathbb{E}_\nu\left[u\right] \geq \mathbb{E}_\mu\left[\phi\right] - \mathbb{E}_\nu\left[\psi\right] \ . \tag{9}$$

Lemma 1. *$u \in \mathrm{Lip}_1(d)$ if, and only if, $u = u^d$.*

Proof. The function ψ^d defined in Eq. (8) is always 1-Lipschitz. In fact, the triangle inequality gives

$$d(x,z) + \psi(z) \leq d(x,y) + d(y,z) + \psi(z) \ ,$$

so that $\inf_z$, applied to both sides of the inequality, gives $\psi^d(x) \leq d(x,y) + \psi^d(y)$. Hence, $u = u^d$ implies $u \in \mathrm{Lip}_1(d)$.

Conversely, if $u(x) - u(y) \leq d(x,y)$ then $u(x) \leq \inf_y d(x,y) + u(y) = u^d(x)$. This, together with $u(x) = d(x,y) + u(y)|_{y=x} \geq \inf_y d(x,y) + u(y) = u^d(x)$, gives $u = u^d$. □

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

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

Proof. In our finite dimensional setup, the first equality is a straightforward application of LP. See, for example, [4, §IV.6-7]. The second equality follows from Eq. (9). □

Remark 1. Various other proofs appear in the literature. Here is one possible argument. We have already shown that the middle term is smaller than the left-hand term. Moreover, the right-hand term is smaller then the middle term because, by definition, $d(x,y) \geq u(x) - u(y)$ if u is $\mathrm{Lip}_1(d)$. The equality follows from the existence of a $\widetilde{u} \in \mathrm{Lip}_1(X)$ such that $d(x, y) = \widetilde{u}(x) - \widetilde{u}(y)$ on the support of the optimal $\widetilde{\gamma}$. This, in turn, is proved by a construction based on the inequality of Eq. (6), see [15, §3.1.1] and [18].

Remark 2. The duality theorem supplies a new proof of Prop. 1 as well as of the fact that the mixture of measures is a geodesic.

First, d is a distance. Faithfulness and symmetry are clear. Moreover,

$$d(\mu, \nu) = \sup_{u \in \mathrm{Lip}_1(d)} \int u d(\mu - \zeta + \zeta - \nu) \leq$$

$$\sup_{u \in \mathrm{Lip}_1(d)} \int u d(\mu - \zeta) + \sup_{u \in \mathrm{Lip}_1(d)} \int u d(\zeta - \nu) = d(\mu, \zeta) + d(\zeta, \nu) \ .$$

Second, the mixture is a geodesic. In fact,

$$d(\mu(s), \mu(t)) = (t - s) \sup_{u \in \mathrm{Lip}_1(d)} \int u d(\mu - \nu) = (t - s) d(\mu, \nu) \ .$$

4. Arens-Eells spaces

The third term of the equality in Th. 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 on ξ imply $\sum_x \xi^+(z) = \sum_z \xi^-(z) \leq 1$. If the strict inequality holds, given any probability function p we can choose $\alpha > 0$ such that both $\xi^+ + \alpha p$ and $\xi^- + \alpha p$ be probability functions.

If we ignore this restriction on the elements ξ, we have the vector space $M_0(X)$ of zero-mass measure functions. We can introduce on it the so-called Kantorovich-Bernstein norm

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

Hence, K-distance is the restriction of the norm $\|\cdot\|_{KB}$.

If $u \in \mathrm{Lip}_1(d)$, then the convex set $\mathrm{Lip}_1(d)$ contains the straight line $u + k$ with $k \in \mathbb{R}$. Whence, it is an unbounded polyhedron. As $\langle u + k, \mu - \nu \rangle$ does not depend on the constant k, it suffices to restrict the analysis to the polytope consisting of the elements $u \in \mathrm{Lip}_1(d)$ for which $u(x_0) = 0$, where x_0 is a fixed vertex in X.

Define the space $\mathrm{Lip}^+(d)$ of the Lipschitz functions 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. The norm $\|u\|_{\mathrm{Lip}}$ in $\mathrm{Lip}^+(d)$ is given by the smallest Lipschitz constant of u. Let $\mathrm{Lip}^+_1(d)$ denote 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

$$d(\delta_x, \delta_y) = \|\delta_x - \delta_y\|_{KB} = \sup_{u \in \mathrm{Lip}^+_1(d)} \langle \delta_x - \delta_y, u \rangle = \sup_{u \in \mathrm{Lip}^+_1(d)} (u(x) - u(y)) = d(x, y) \ . \tag{10}$$

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

$$\xi = \sum_{x,y} a(x, y)(\delta_x - \delta_y) \ , \quad A = [a(x, y)]_{x,y \in X} \in \mathbb{R}^{X \times X} \ . \tag{11}$$

If $\#X = n$, the vector space $M_0(X)$ has dimension $n-1$, while the dimension the space of matrices is n^2.

Let us briefly discuss the issue of multiple representations of the same $\xi \in Æ(X)$ in Eq. (11). The set of such representations is an affine space whose tangent space is spanned by difference between two representations of the same ξ, i.e., it is given by all functions of the form $\sum_{x,y} a(x,y)(\delta_x - \delta_y) = 0$, whose value at each z is $\sum_y a(z,y) - \sum_x a(x,z) = 0$. All such a span the vector space of real matrices with sum of the z-row equal to the sum of the z-column.

In terms of the matrix $A = [a(x,y)]_{x,y \in X}$, if we define the linear map $T\colon A \mapsto (A - A^t)\mathbf{1}$, where $\mathbf{1}$ is the unit vector, then the kernel of T is the tangent space to the space of the representations of ξ.

For example, if X is pointed at x_0, we can exhaustively write the representations of Eq. (11) as

$$\xi = \sum_{x \neq x_0} (\xi(x) - \xi(x_0))(\delta_x - \delta_{x_0}) + \sum_{x,y \in X} a(x,y)(\delta_x - \delta_y) \ , \quad A \in \ker T \ . \tag{12}$$

The adjoint T^* of T is

$$\langle T(A), u \rangle = \mathbf{1}^t (A^t - A) u = \operatorname{Tr}\left(A^t (u \mathbf{1}^t - \mathbf{1} u^t)\right) \ ,$$

that is, $T^* u = [u(x) - u(y)]_{x,y \in X}$.

If we compute the norm of a generic function ξ in Eq. (11), we get through (10)

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

The vector space $M_0(X)$ of zero-mass signed measure functions on X can be endowed with the following norm, in which case it is called the Arens-Eells space $Æ(X)$. Here, we follow the presentation by N. Weaver [20].

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

$$\|\xi\|_{Æ} = \inf \left\{ \sum_{x,y \in X} |a(x,y)| \, d(x,y) \right\} \ , \tag{13}$$

where the inf is made on all the representations of ξ in Eq. (11).

Remark 3. Arens-Eells construction has a wider range of application than finite metric spaces. In fact, in a more general setting, Arens-Eells space is 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 the Lipschitz-free space over X and denoted by $\mathcal{F}(X)$.

In connection to the embedding problem of metric spaces, Arens and Eells [3] have shown that the Banach space $Æ(X)$ is a predual of $\operatorname{Lip}^+(d)$. More specifically, the linear isometry $T : Æ(X)^* \to \operatorname{Lip}^+(d)$ is given by

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

for every $\phi \in Æ(X)^*$ and $x \in X$.

Let us compute the norm in $\operatorname{Lip}^+(d)$, dual of $Æ-norm$. We have

$$\|u\|_{Æ(X)^*} = \sup_{\sum_{x,y}|a(x,y)|d(x,y)\le 1} \sum_z u(z) \sum_{x,y} a(x,y)(\delta_x - \delta_y) =$$

$$\sup_{\sum_{x,y}|a(x,y)|d(x,y)\le 1} \sum_{x\neq y} a(x,y)(u(x)-u(y)) =$$

$$\sup_{\sum_{x,y}|a(x,y)|d(x,y)\le 1} \sum_{x\neq y} a(x,y)d(x,y)\frac{u(x)-u(y)}{d(x,y)} =$$

$$\sup_{x\neq y} \frac{u(x)-u(y)}{d(x,y)} = \|u\|_{\mathrm{Lip}} \ ,$$

where the conclusion follows from the duality between ℓ^1 and ℓ^∞.

Notice that the previous identification of norms provides the relation between the norm as defined in Eq. (13) and the norm that follows from the identification of $\mathrm{Lip}^+(d)$ with $Æ\,(X)^*$. Namely,

$$\|\xi\|_{Æ} = \|\xi\|_{KB} = \sup\left\{\langle \xi, u\rangle \mid u \in \mathrm{Lip}_1^+(d)\right\} \tag{14}$$

for all $\xi \in Æ\,(X)$. In addition, there exists a multi-mapping $J\colon\ Æ\,(X) \to \mathrm{Lip}_1^+(d)$ 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 a property of the Arens-Eells norm we shall use below (see [20] for more details).

Proposition 7. *The norm $\|\cdot\|_{Æ}$ is the largest semi-norm on the space $Æ\,(X)$ which satisfies $\|\delta_x - \delta_y\| \le d(x,y)$.*

Proof. If $\|\cdot\|$ is a semi-norm that satisfies the claimed requirements, we have from Eq. (11) that

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

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

5. Distance induced by a graph

Let us first introduce a useful notion that refines that of adjacency 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 path joining the points themselves. A graph is called *minimal* if for every pair of adjacent vertices xy, the amount $w(x,y)$ is strictly less than the length of any other path from x to y.

In a tree, every pair of adjacent vertices are close. So too are all pairs in an un-weighted graph. Observe further that, in any shortest path $x_1, x_2, ..., x_n$, two adjacent vertices are necessarily close. This is the reason why it holds the equality

$$d(x_1, x_n) = \sum_{i=1}^{n-1} d(x_i, x_{i+1}) \tag{15}$$

along the points of path of minimal length.

Actually, suppose that not all the adjacent points of a "geodesic" are close. 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 in Prop. 6 too one could replace adjacent points with close points.

Remark 4. Despite a graph (X, w) is not necessarily minimal, it can be made into a minimal graph without altering its metric. It suffices to remove, one at a time, edges xy that do not satisfy the stated property. After a finite number of steps, we get a connected minimal graph inducing the same metric (cf. also the related fact (i) of Th. 3).

Recall that every finite metric space can be realized as a graphic metric space. Clearly, if (X, d) is a finite space, take the complete graph (G, w), with $V(G) = X$ and weights $w = d$. Thanks to triangular inequality, the induced distance d_G satisfies $d_G = w = d$. However, this G is not necessarily minimal and a graph with fewer edges can be obtained by the method illustrated above.

More detail about the topic of embedding finite metrics into graphic metrics and related references, we refer to [8, §20.4]. For instance, under a known *four-point condition*, a metric space is realized by a weighted tree (the so-called *tree metrics*). Ultrametrics fall into this class of metrics.

It is known in literature a characterization of the extreme points of the unit ball of $\mathrm{Lip}^+(d)$ for generic metric spaces, cf. [10], [16] and [20, Th. 2.59].

We present here just a simple specification for finite spaces. For ease of the reader we provide a proof which substantially follows [20, Th. 2.59].

Theorem 2. *Let (X, x_0) be a pointed a 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 $x_0, x_1, ..., x_{n-1}$, with $x_{n-1} = x$, such that*

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

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

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 $x_0, x_1, ..., x_{n-1} = x$, with $|f(x_i) - f(x_{i-1})| = d(x_i, x_{i-1})$.

Further, in view of Prop. 6, 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 necessary part of the proof, 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 $x_0, x_1, ..., x_{n-1} = \bar{x}$ in which points x_i, x_{i+1} are close.

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

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

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

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

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

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

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 two specific classes of graphs. In the first one, we are dealing with a straightforward application of the above result and this does not require a further proof.

Proposition 8. *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.*

We shall obtain in Sec. 6 a more transparent representation of these extremal functions, see Eq. (25).

Next consider a set X equipped by the discrete distance, i.e., d is generated by the un-weighted complete graph, usually denoted by K_n.

Proposition 9. *Let d be the discrete distance. The function $u \in \mathrm{ext}\,\mathrm{Lip}_1^+(d)$ if, and only if, $u = \pm I_Y$, where I_Y is the indicator function of a set $\emptyset \neq Y \neq X$.*

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 Th. 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 Th. 2, then $x_1 = \pm 1$. Since $u \in \mathrm{Lip}_1^+(d)$, we infer that either $0 \le u(x) \le u(x_1) = 1$ or $-1 \le u(x) \le 0$.

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 Th. 2 fails. □

Let us now calculate a few K-distances through the extreme points of $\mathrm{Lip}_1^+(d)$. We re-examine two well known examples, where a closed-form expression of the distance is already available.

Example 2 (Discrete distance). In the space K_n, the K-distance, equals the distance in variation:

$$d(\mu, \nu) = \sum_{x \in X} (\mu(x) - \nu(x))^+ = \frac{1}{2} \sum_{x \in X} |\mu(x) - \nu(x)| \ . \tag{16}$$

In view of Eq. (14) and Prop. 9, we have to seek for a function $u = \pm I_Y$ that maximizes $\sum_{x \in X} u(x) \, (\mu(x) - \nu(x))$.

Dealing separately with the two groups of functions (I_Y and $-I_Y$), we get the two amounts to maximize

$$\sum_{x \in Y} (\mu(x) - \nu(x)) \quad \text{and} \quad -\sum_{x \in Y} (\mu(x) - \nu(x)) \, .$$

In the first one the maximum value is attained provided $Y = \{x : \mu(x) \geq \nu(x)\}$. Namely,

$$\sum_{x \in X} (\mu(x) - \nu(x))^+ \ .$$

Regarding the second one, we get the same thing, as being

$$\sum_{x \in X} (\mu(x) - \nu(x))^+ = \sum_{x \in X} (\mu(x) - \nu(x))^- \, .$$

Example 3 (Linear order). Set $X = \{1, 2, .., n\}$ and the weighted graph to be the total ordering $1 \to 2 \to .... \to n$, with positive weights $w_{i,i+1} = d(i, i+1)$.

According to Prop. 8, the extreme points of the unit ball $\mathrm{Lip}_1^+(d)$ are the functions u satisfying the conditions $|u_i - u_{i+1}| = d(i, i+1)$, for $i = 1, 2, .., n-1$.

By using the "discrete integration by part" formula

$$\sum_{i=1}^{n} a_i b_i = -\sum_{i=1}^{n} (b_{i+1} - b_i)(a_1 + a_2 + ... + a_i)$$

which holds for two numerical sequences (a_i), (b_i), $i = 1, 2, ..., n$, and with $b_{n+1} = 0$, the objective function, after having set $\xi = \mu - \nu$, becomes

$$\sum_{i=1}^{n} u_i \, (\mu_i - \nu_i) = \sum_{i=1}^{n-1} (u_i - u_{i+1})(\xi_1 + \xi_2 + ... + \xi_i).$$

Therefore the distance between two probability measures will be given by

$$d\left(\mu,\mu+\xi\right)=\sum_{i=1}^{n-1}d(i,i+1)\left|\xi_1+\xi_2+...+\xi_i\right|.$$

This result reveals that in the linear graphs the K-distance reduces to the ordinary distance between the two cumulate functions. This is of some interest in the present paper, because it will be generalized to trees in Sec. 6 below

5.1. **A support property.** Many elements $a(x,y)$ of the linear combinations (11) are needless. They can be avoided, to the purpose of calculating the inf of (13). The next result is related, but different from what has been discussed in Sec. 2.1. In fact, in those instances the graph is the support of the optimal coupling, while now it is the graph which defines the distance. An application of this theorem will be presented in Sec. 7.1.

Theorem 3. *Assume that the distance d is generated by a weighted graph. The class of functions $a(x,y)$, that are present in Eq.* (13)*, 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 $\{(x,y)\,|\,a(x,y)\neq 0\}$ has no cycle.

Proof: Condition i). Let us first assume that in the representation (11) there is a non-zero term $a(x,y)(\delta_x-\delta_y)$, where x and y are not adjacent. Let $x_1,x_2,...,x_n$ be a geodesic path joining $x=x_1$ to $y=x_n$. By (15)

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

and

$$\delta_x-\delta_y=\sum_{i=1}^{n-1}(\delta_{x_i}-\delta_{x_{i+1}}).$$

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 $\delta_x-\delta_y$ 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 $x_1,x_2,...,x_n$ exists with $x=x_1$ and $y=x_n$, and

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

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 $\delta_x - \delta_y$ may be removed, when the pair is not close.

□

Proof: Condition 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 the vector ξ. That is, let equality

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

hold. We show that if the graph generated by these $\widetilde{a}(x,y)$ contains a cycle, then there is another optimal representation of the same ξ not having cycles. More precisely, if the graph contains $k > 1$ cycles, they can be deleted one at a time.

Assume that in the 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 other possible terms. Since along a cycle we have

$$\sum_{(x,y) \in \mathcal{S}} (\delta_x - \delta_y) = 0,$$

then

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

where t is any scalar number. It follows that

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

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,

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

and so the cycle $\mathcal{S}$ is ruled out.

Once this cycle has been eliminated, we can proceed to eliminate another possible cycle, and so on. After a finite number of steps, we get to a free of cycles representation of ξ. □

5.2. **A decomposition property.** Under proper conditions, the norm associated with a graph may be deduced by decomposing the graph itself into a certain number of sub-graphs.

Following [20], if $\{X_\lambda\}$ is a family of pointed metric spaces, the sum $\coprod X_\lambda$ denotes their disjoint union with all base points identified and metric

$$d(x,y) = d(x,e) + d(e,y),$$

whenever x and y belong to distinct summands and e is the common base point.This operation is also known as *1-sum operation*, see [8, §7.6].

A connected graph $G = (X, \mathcal{E})$ is called decomposable if there is a vertex $x_0 \in X$ such that the graph $G \setminus x_0$ is not connected. If $G \setminus x_0$ has $k \geq 2$ components, then the set of vertices X can be partitioned as $X = X_1 \cup X_2 \cup \cdots \cup X_k$, with $X_i \cap X_j = \{x_0\}$. Consequently, if x and y lie into two distinct components X_i and X_j, then

$$d(x,y) = d(x,x_0) + d(x_0,y).$$

Hence, by means of vertex x_0, the graph $G = (X, \mathcal{E})$ splits into k sub-graphs $G_i = (X_i, \mathcal{E}_i)$, $i = 1, \ldots, k$, having in common the vertex x_0, and we can adopt the notation $G = \coprod_{i=1}^k G_i$, where $X = \coprod_{i=1}^k X_i$.

Observe further that any vector $\xi \in M_0(X)$ has a canonical decomposition $\xi = \xi^1 + \xi^2 + \cdots + \xi^k$, with $\xi^i \in M_0(X_i)$. More specifically,

$$\xi^i(x) = \xi(x) \text{ for } x \in X_i \setminus x_0 \text{ and } \xi^i(x_0) = -\sum_{x \in X_i \setminus x_0} \xi(x).$$

After these preliminaries, we can state the following result, whose proof is referred to [20, Prop. 3.9].

Proposition 10. *Assume that a vertex $x_0 \in X$ splits the graph $G = (X, \mathcal{E})$ into k components $G_i = (X_i, \mathcal{E}_i)$. Then*

$$\text{Æ}\left(\coprod_{i=1}^k X_i\right) \cong \bigoplus_i \text{Æ}\,(X_i)\,.$$

More precisely,

$$\|\xi\|_G = \left\|\xi^1\right\|_{G_1} + \left\|\xi^2\right\|_{G_2} + \cdot + \left\|\xi^k\right\|_{G_k}\ .$$

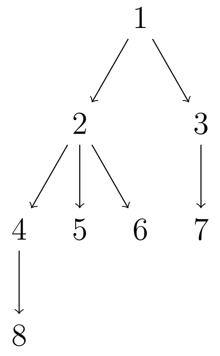
Example 4. In the two-cycles graph of Fig. 3(b), the vertex 2 breaks the graph into 2 components. In view of Example 1, the norm of an element $\xi = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$ will be given by

$$\|\xi\|_{\text{Æ}} = \frac{1}{2}(|\xi_1| + |\xi_5| + |\xi_1 + \xi_5| + |\xi_4| + |\xi_3| + |\xi_4 + \xi_3|).$$

6. Trees

This section is devoted to the specific analysis of the K-distance generated by an arbitrary tree or, which is the same, of its extension to the norm $\|\cdot\|_{\text{Æ}}$ to the Arens-Eells space Æ$\,(X)$. For that, we need some more notation.

Let T_n be a rooted tree of order n, where a distinguished vertex x_0 has been selected to be the root of the tree. Each vertex $x \in X$ can be classified according to its distance, that is, its (un-weighted) distance from the root. Hence, the set X is partitioned as



	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

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

FIGURE 2. *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^*.

$X = \{V_0, V_1, V_2, ..., V_k\}$, with $V_0 = \{x_0\}$ and, recursively, $x \in V_r$ if and only if it is adjacent to an $y \in V_{r-1}$, see Fig. 2.

In this way, we get an oriented tree and we can define a partial order $x \preceq y$ to denote that y is a descendant of x, that is, there is a path from x to y with $d(x_0, x) < d(x_0, y)$ or $x = y$. The dual relation $\succeq$ will be also used so that $x \preceq y$ is equivalent to $y \succeq x$. Finally, for each non-root vertex $x \neq x_0$, x^+ denotes the adjacent predecessor of x i.e., if $x \in V_r$, x^+ is the unique element in V_{r-1}, which is adjacent to x.

Under this notation, we can establish a closed-form equation which provides the Kantorovich distance for weighted rooted trees. Such a closed form is based on the cumulative function $\Xi(x) = \sum_{y \succeq x} \xi(y)$, for $x \in X$ and $\xi \in M_0(X)$. Note that $\Xi(x_0) = \sum_{y \in X} \xi(y) = 0$. See the right panel of Fig. 2.

Recall that $w(x, y) = d(x, y)$ holds for the edges xy of a tree. Moreover, the shortest path is the same under all weights and for all selections of a vertex as a root.

Theorem 4. *Let T be a weighted rooted tree on the set X. Then*

$$\|\xi\|_{Æ} = \sum_{x \in X \setminus x_0} d(x, x^+)\, |\Xi(x)| \ , \quad \text{with} \quad \Xi(x) = \sum_{y \succeq x} \xi(y), \tag{17}$$

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

$$\|\xi\|_{Æ} = \sum_{y \in X} \xi(y) \sum_{x_0 \neq x \preceq y} d(x, x^+)\, \mathrm{Sgn}\, \Xi(x) \ , \tag{18}$$

where $\mathrm{Sgn}(\cdot) \in \{-1, 1\}$ *is the sign function, the value at zero is* $\mathrm{Sgn}(0) = \pm 1$.

Proof. If $n = 2$, formula (17) becomes $\|\xi\|_{Æ} = w(1, 2)\, |\xi(1)|$ which is the right value (see Ex. 3).

By induction, suppose that the representation (17) is true for any tree with $n - 1$ vertices. We must prove that (17) is still true for a tree T_n. Set

$$\Phi(\xi) = \sum_{x \in X \setminus x_0} d(x, x^+) \left| \sum_{y \succeq x} \xi(y) \right| . \tag{19}$$

Clearly, Φ is a semi-norm. Let us check that $\Phi(\delta_x - \delta_y) = d(x, y)$.

Taking any pair of vertices $x, y \in X$, we have one of the following cases: $x \preceq y$, $y \preceq x$, or there is a maximal vertex z such that $z \preceq x$ and $z \preceq y$. We shall treat the last case. The others are similar.

Hence, we can find two distinct paths $x_1, x_2, \cdots, x_r$ and $y_0, y_1, \cdots, y_s$, with $x_1 = x$, $x_r = z = y_0$ and $y_s = y$.

According to Eq. (19),

$$\Phi(\delta_x - \delta_y) = d(x_1, x_2) + \cdots + d(x_{r-1}, z) + d(y_1, z) + \cdots + d(y_s, y_{s-1}) = d(x, y).$$

Thanks to Prop. 7, it remains to prove that Φ is largest semi-norm that satisfies the requirements. Let $\|\cdot\|$ be a semi-norm on Æ(X) for which $\|\delta_x - \delta_y\| \leq d(x, y)$.

Fix a leaf of T_n, say $x = 1$, and consider its adjacent predecessor x^+. Say $x^+ = 2$.

Decompose each element $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$ in Æ(X), in the following manner, $\xi = \xi^1 + \xi^2$, where

$$\xi^1 = (\xi_1, -\xi_1, 0, \cdots, 0) \qquad \xi^2 = (0, \xi_2 + \xi_1, \xi_3, \cdots, \xi_n).$$

As $\|\cdot\|$ is a semi-norm, we have

$$\|\xi\| \leq \left\|\xi^1\right\| + \left\|\xi^2\right\|.$$

Since the semi-norm $\xi^1 \to \|\xi^1\|$ is defined on a one-dimensional space, it must be of the kind $\|\xi^1\| = c\,|\xi_1|$, with $c \leq d(1, 2)$. The second semi-norm is defined over a tree with $n - 1$ vertices. Consequently $\|\xi^2\| \leq \|\xi^2\|_{Æ}$. Hence

$$\|\xi\| \leq d(1, 2)\,|\xi_1| + \left\|\xi^2\right\|_{Æ} = \Phi(\xi)$$

that provides the desired result.

Regarding Eq. (18), it suffices to interchange the order of two summations in Eq. (17). 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. (18). □

Fig. 2 shows an alternative method to find the norm (17). It relies on iterating the adjacency matrix E of the tree, i.e., the matrix whose entries are $a_{x,y} = 1$, if x and y are adjacent with $x \preceq y$, and $a_{x,y} = 0$ otherwise.

Computing the finite sum

$$E^* = (I - E)^{-1} = \sum_{n=0}^{\infty} E^n,$$

the addenda of (17) appear as the rows of the matrix E^*.

An explanation of this fact is that there holds a linear relationship between the distribution ξ and its cumulative distribution Ξ, and it can be formulated through the adjacency matrix. After having labeled the rooted tree and denoted by ξ and Ξ the two resultant column vectors, then the following two equivalent equations hold

$$\xi = (I-E)\ \Xi \quad \Longleftrightarrow \quad \Xi = (I-E)^{-1}\xi = E^*\xi.$$

Indeed, the relation on the left-hand side is the matrix form of the obvious identity

$$\xi(x) = \Xi(x) - \sum_{y\in \mathrm{child}(x)} \Xi(y)\ , \tag{20}$$

for all $x \in X$.

Thanks to the explicit form of the Arens-Eells norm of Eqs. (17) and (18), it is not hard to evaluate functions $a(x, y)$ that minimize the Arens-Eells norm (13) as well as the Lipschitz function that maximizes its dual problem.

Theorem 5. *For any weighted rooted tree, an optimal solution $\widetilde{a}(x, y)$ to the problem*

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

is given by

$$\widetilde{a}(x,y) = \begin{cases} \Xi(x) & \text{if } (x,y) = (x,x^+)\ , \\ 0 & \text{if } y \neq x^+. \end{cases} \tag{21}$$

Of course, such a minimal solution $\widetilde{a}(x, y)$ is not the unique one. A changing the root of the tree delivers a different optimal solution.

Proof. For each given ξ, define

$$\eta = \sum_{x,y\in X} \widetilde{a}(x,y)(\delta_x - \delta_y)\ , \tag{22}$$

with $\widetilde{a}$ defined in Eq. (21). It follows that

$$\eta = \sum_{x\in X, x\neq x_0} \widetilde{a}(x,x^+)(\delta_x - \delta_{x^+}) = \sum_{x\in X, x\neq x_0} \Xi(x)(\delta_x - \delta_{x^+}).$$

Evaluating η at a generic non-root vertex $z \in X$, we obtain (see (20))

$$\eta(z) = \sum_{x\in X, x\neq x_0} \Xi(x)(\delta_x(z) - \delta_{x^+}(z)) = \Xi(z) - \sum_{x\in \mathrm{child}(z)} \Xi(x) = \xi(z)\ . \tag{23}$$

Whence, by Th. 4, η is a minimizing value. In fact, ξ is a linear combination of vectors $\{\delta_x - \delta_{x^+}\}$, with coefficients given by (22). In view of definition of Arens-Eells norm (see Eq. (13)) we get

$$\|\xi\|_{Æ} \le \sum_{x,y\in X} |\widetilde{a}(x,y)|\, d(x,y) = \sum_{x\in X\setminus x_0} \widetilde{a}(x,x^+) d(x,x^+)$$
$$\le \sum_{x\in X\setminus x_0} |\Xi(x)|\, w(x,x^+) = \|\xi\|_{Æ}\ ,$$

which is the desired result. □

Next we will single out a Lipschitz function that maximizes the dual program (cf. Th. 1, and Eq. (14)).

Theorem 6. *For every $\xi \in Æ(X)$ it holds $\langle \xi, \bar{u} \rangle = \|\xi\|_{Æ}$, where $\bar{u} \in \operatorname{ext} \operatorname{Lip}_1^+(d)$ is given by*

$$\bar{u}(y) = \sum_{x \preceq y} d(x, x^+) \operatorname{Sgn} \Xi(x), \quad \forall y \in X. \tag{24}$$

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.

Proof. In view of (18), it remains only to check that such a function $\bar{u}$ is an extreme point of $\operatorname{Lip}_1^+(d)$.

Pick two adjacent vertices $y_1, y_2 \in X$. Suppose $y_1 \succeq y_2$. Then

$$\begin{aligned} \bar{u}(y_1) &= d(y_1, y_1^+) \operatorname{Sgn} \Xi(y_1) + \sum_{y_2 \succeq x} d(x, x^+) \operatorname{Sgn} \Xi(x) \\ &= d(y_1, y_1^+) \operatorname{Sgn} \Xi(y_1) + \bar{u}(y_2) \ . \end{aligned}$$

Namely, $\bar{u}(y_1) - \bar{u}(y_2) = \pm d(y_1, y_2)$, and Prop. 8 yields the desired result. □

Remark 5. Th. 6 suggests another method to find out the Arens-Eells norm, based on the extreme points of the unit ball of $\operatorname{Lip}^+(d)$. In fact, it is not difficult to understand that every function $u \in \operatorname{ext} \operatorname{Lip}_1^+(d)$ can be realized by

$$u_\epsilon(y) = \sum_{y \succeq x} d(x, x^+) \epsilon(x) \tag{25}$$

where ϵ is an assigned function $\epsilon : X \setminus x_0 \to \{-1, 1\}$ and the base point, at which u_ϵ vanishes, agrees with the root x_0 of the tree.

It suffices to observe that such functions are generated by the recursive equation

$$u_\epsilon(y) = u_\epsilon(x) + d(x, y)\epsilon(y), \quad \forall y \in \operatorname{child}(x) \tag{26}$$

with initial condition $u(x_0) = 0$ and so the desired result follows from Prop. 8.

We have thus to maximize the functional

$$\sum_{y \in X} \xi(y) u_\epsilon(y) = \sum_{y \in X} \xi(y) \sum_{y \succeq x} d(x, x^+) \epsilon(x), \tag{27}$$

over the functions ϵ. On the other hand, by interchanging the order between the two summation operators, we get

$$\sum_{y \in X} \xi(y) \sum_{y \succeq x} d(x, x^+) \epsilon(x) = \sum_{x \in X} d(x, x^+) \epsilon(x) \Xi(x) \tag{28}$$

and so the maximum is attained when $\epsilon(x) = \operatorname{Sgn} \Xi(x)$ for all $x \in X$.

The structure of the extreme points, as well as of the whole space $\operatorname{Lip}^+(d)$, is easily understood through the next result. Associate to every function ϕ defined on $X \setminus x_0$, the following Kantorovich potential

$$u_\phi(y) = \sum_{y \succeq x} d(x, x^+) \phi(x) \tag{29}$$

defined on the vertices of the tree.

Proposition 11. *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*

$$(\Delta u)(x) = \frac{u(x) - u(x^+)}{d(x, x^+)} \tag{30}$$

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,

$$u_\phi(y_1) = d(y_1, y_2)\phi(y_1) + u_\phi(y_2). \tag{31}$$

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

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 (31) implies the equality $\|u_\phi\|_{Lip} = \|\phi\|_\infty$. We have so proved that the mapping is an injective isometry.

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

$$(\Psi \circ \Delta)u(y) = \sum_{y \succeq x \neq x_0} d(x, x^+) \frac{u(x) - u(x^+)}{d(x, x^+)} = u(y). \tag{32}$$

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 11 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 $\|x\| \le 1$ of $l_\infty(X \setminus x_0)$.

iii) Interestingly, the inverse map Δ of $\phi \to u_\phi$ is closely related to De Leeuw's map which is the map that associates with a Lipschitz function $f : X \to \mathbb{R}$, the function

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

defined for $x \neq y \in X$ (see [20]).

Another fact of interest is the differentiability of the Æ-norm, which is a direct consequence of the representation (18).

Proposition 12. *The norm-function $f(\xi) = \|\xi\|_{\text{Æ}}$ is differentiable at every $\xi \in \text{Æ}(X)$ such that $\Xi(x) \neq 0$ for all $x \neq x_0$. Its gradient $\nabla \|\xi\|_{\text{Æ}} \in \mathrm{ext}\,\mathrm{Lip}_1^+(d)$ is given by*

$$\nabla \|\xi\|_{\text{Æ}} = \bar{u}_\xi \tag{33}$$

where $\bar{u}_\xi$ is aligned with ξ, i.e., $\langle \xi, \bar{u}_\xi \rangle = \|\xi\|_{\text{Æ}}$. More generally, the directional derivative of f at ξ, into the direction $\eta \in \text{Æ}(X)$, is

$$f'(\xi; \eta) = \sum_{y \in X} \eta(y) \sum_{y \preceq x, x \in V} d(x, x^+)\,\mathrm{Sgn}\,\Xi(x) + \sum_{y \in X} \eta(y) \sum_{y \preceq x, x \in V^0} d(x, x^+)\,\mathrm{Sgn}(\sum_{x \preceq y} \eta(y)).$$

where $V^0 \subseteq X$ are the vertices x for which $\Xi(x)$ vanishes and $V = X \setminus V^0$.

Proof. If $\Xi(x) \neq 0$ for all $x \neq x_0$, then the functions $\mathrm{Sgn}\,\Xi(x)$ are constant in a neighborhood of ξ. By (18) it follows that the function $\xi \to \|\xi\|_{\text{Æ}}$ is locally linear. Th. 6 provides the desired gradient.

Tedious algebra leads to the directional derivatives too. □

The importance of knowing an explicit solution to the dual problem, as it has been established in Th. 6, is that this fact, in turn, furnishes information about the solutions to the primal problem. This happens since the support of the optimal plan is contained in the set of the points for which $\bar{u}(y) - \bar{u}(x) = d(x, y)$. An alternative way of thinking is that of resorting to the well-known complementary slackness conditions which are valid between the primal and the dual LP, see [13, §3.3].

In this light, we have seen that

$$\operatorname{Sgn} \Xi(x) < 0 \iff \bar{u}(x^+) - \bar{u}(x) = d(x, x^+), \tag{34}$$

for any $x \neq x_0$.

It is however difficult to give a general explicit solution to the primal problem. The next result presents a positive result, which is valid under restriction (6) below.

Theorem 7. *Consider a rooted tree in X and let $\xi = \mu - \nu$. An optimal plan $\gamma^* \in \mathcal{P}(\mu, \nu)$ is given by*

$$\begin{aligned} \gamma^*(x, x^+) &= [\Xi(x)]^+ , \quad \gamma^*(x^+, x) = [\Xi(x)]^- , \\ \gamma^*(x, x) &= \mu(x) - [\Xi(x)]^+ - \sum_{u \in \text{child}(x)} [\Xi(u)]^- , \end{aligned}$$

and $\gamma^*(x, y) = 0$ otherwise, provided

$$\mu(x) \geq [\Xi(x)]^+ + \sum_{u \in \text{child}(x)} [\Xi(u)]^- \tag{35}$$

is true for every $x \in X$. A sufficient condition for (35) to hold is that $\mu \gg 0$ and $\|\mu - \nu\|_{l_1}$ is sufficiently small.

Remark 6. Another set of optimal solutions can be easily established by changing the role of the two probability functions. Therefore we also have:

$$\begin{aligned} \gamma^*(x, x^+) &= [\Xi(x)]^- , \quad \gamma^*(x^+, x) = [\Xi(x)]^+ , \\ \gamma^*(x, x) &= \nu(x) - [\Xi(x)]^+ - \sum_{u \in \text{child}(x)} [\Xi(u)]^- \\ \text{if} \quad \nu(x) &\geq [\Xi(x)]^- + \sum_{u \in \text{child}(x)} [\Xi(u)]^+ . \end{aligned}$$

Proof. By construction, $\gamma^*(x, y) \geq 0$. Let us show that the plan γ^* is feasible. Actually,

$$\sum_{y \in X} \gamma^*(x, y) = \gamma^*(x, x) + \gamma^*(x, x^+) + \sum_{u \in \text{child}(x)} \gamma^*(x, u) =$$

$$\gamma^*(x, x) + [\Xi(x)]^+ + \sum_{u \in \text{child}(x)} [\Xi(u)]^- = \mu(x),$$

while

$$\sum_{x\in X}\gamma^*(x,y) = \gamma^*(y,y)+\gamma^*(y^+,y)+\sum_{u\in\text{child}(y)}\gamma^*(u,y) =$$

$$= \mu(y) - [\Xi(y)]^+ - \sum_{u\in\text{child}(y)}[\Xi(u)]^- + [\Xi(y)]^- + \sum_{u\in\text{child}(y)}[\Xi(u)]^+$$

$$= \mu(y) - \Xi(y) + \sum_{u\in\text{child}(y)}\Xi(u) = \mu(y)-\xi(y) = \nu(y) \ ,$$

where the last line comes from Eq. (23). So we have checked that γ^* is a feasible plan.

Regarding its optimality, we have

$$\begin{aligned}\sum_{x,y\in X} d(x,y)\gamma^*(x,y) &= \sum_{x\in X}\sum_{y\in X} d(x,y)\gamma^*(x,y)\\ &= \sum_{x\in X}[d(x,x^+)\gamma^*(x,x^+) + \sum_{u\in\text{child}(x)} d(x,u)\gamma^*(x,u)]\\ &= \sum_{x\in X}[d(x,x^+)[\Xi(x)]^+ + \sum_{u\in\text{child}(x)} d(x,u)[\Xi(u)]^-]\\ &= \sum_{x\in X} d(x,x^+)[\Xi(x)]^+ + \sum_{x\in X}\sum_{u\in\text{child}(x)} d(x,u)[\Xi(u)]^-.\end{aligned}$$

Under the usual interchanging of summation order, the last addendum becomes

$$\sum_{x\in X}\sum_{u\in\text{child}(x)} d(x,u)[\Xi(u)]^- = \sum_{x\in X}\sum_{u\in X} d(x,u)[\Xi(u)]^- I(x,u) =$$

$$\sum_{u\in X}[\Xi(u)]^-\sum_{x\in X} d(x,u)I(x,u) = \sum_{u\in X}[\Xi(u)]^- d(u^+,u) \ .$$

At last, we get

$$\sum_{x,y\in X} d(x,y)\gamma^*(x,y) = \sum_{x\in X} d(x,x^+)\,|\Xi(x)| = \|\xi\|_{\text{Æ}}$$

which is the desired result.

To conclude, since condition (35) is fulfilled if $\mu(x) > 0$ and $\xi = \mu - \nu = 0$, it continues to be true under our specified assumptions. □

Feasibility condition (35) imposes severe restrictions on the actual extent of the previous result. For instance, the case of two Dirac functions, $\mu = \delta_x$ and $\nu = \delta_y$, is not covered by it. More generally, adding Eq. 35 with respect to the x variable, we get the relation

$$\sum_{x\in X}|\mu(x)-\nu(x)| \le 1 \tag{36}$$

that obliges the two distributions to be sufficiently close to each other.

A more exact condition than that declared in the proposition is

$$\operatorname{Sgn}\Xi(x)\cdot\operatorname{Sgn}\Xi(u) < 0, \quad \forall u \in \text{child}\,(x)\,, \quad \forall x \in X\setminus x_0,$$

that, in view of Eq. (34), entails the support of optimal plans to be concentrated only on the pairs of vertices of kind (x,x), (x,x^+) and (x^+,x). It is also easy to check that the above condition implies Eq. 35.

Our purpose now is to clarify further the relation between the representation of the distance of Eq. (17) and the trees on X. For instance, it is not clear whether the AE-norm admits a unique representation or not.

Let us introduce the following class of semi-norms in Æ(X).

A semi-norm is called of class (T) if it is of the form

$$\|\xi\| = \sum_{j=1}^{m} A_j \, |a_j \cdot \xi| \tag{37}$$

where a_j are Boolean functions i.e., $a_j \in \{0,1\}^X$, $a_j \cdot \xi = \sum_{x \in X} a_j(x)\xi(x)$ and the A_j's are non-negative coefficients.

While in Eq. (17) the index m is the number $n-1$ of the edges, here there is no restriction. This allows for the possibility to take into account repeated terms too.

Since we are concerned about uniqueness, a caveat is an order. The Kantorovich distance in the case of two points graph, is $d(\mu, \mu + \xi) = |\xi_1| = |\xi_2|$. Clearly the two functions $|\xi_1|$ and $|\xi_2|$ are distinct, though they agree on the space Æ(X). Therefore, uniqueness will have to be understood up to the relation

$$\xi_1 + \xi_2 + ... + \xi_n = 0. \tag{38}$$

Proposition 13 (Uniqueness)**.** *Consider a weighted tree in X and let $\|\cdot\|$ be a seminorm of class (T) that induces the tree, e.i., $\|\delta_x - \delta_y\| = d(x,y)$ holds for each pair $x, y \in X$. Then, $\|\xi\| = \|\xi\|_{Æ}$ on Æ(X).*

Proof. This statement is true for $n = 2$. By induction, suppose that this assertion holds for trees with $n-1$ vertices and that the semi-norm $\|\cdot\|$ induces a tree of order n.

Let x be a leave of T_n, namely, x be a vertex such that $T_n \setminus \{x\}$ is still a tree. Let y be the vertex adjacent to x. For sake of simplicity, label the two vertices as $x = 2$ and $y = 1$, respectively.

Thanks to the relation (38) we can rule out the variable ξ_1 by setting $\xi_1 = -\sum_{i=2}^{n} \xi_i$. With abuse of notation we still denote by $\|(\xi_2, \xi_3, ..., \xi_n)\|$ this new semi-norm.

We can write

$$\|(\xi_2, \xi_3, ..., \xi_n)\| = \sum_{i \in J} A_i \, |\xi_2 + ...| + \Psi(\xi_3, ..., \xi_n). \tag{39}$$

Here the notation $|\xi_2 + ...|$ means that there could be some more variables within the absolute value. From the fact that $d(1,2) = w_{12}$, it follows that $\sum_{i \in J} A_j = w_{12}$. We are going to show that every term $|\xi_2 + ...|$ is actually $|\xi_2|$.

Suppose that a certain variable ξ_k, with $3 \le k \le n$, be present is some addend different from those in Ψ.

More specifically, set $J = J_k \cup J_{nk}$. Where $i \in J_k$ if $A_i \, |\xi_2 + \xi_k + ...|$. While $i \in J_{nk}$ means that the addend $A_i \, |\xi_2 + ...|$ does not contain the variable ξ_k.

Clearly,

$$d(2,k) = d(2,1) + d(1,k) = w_{12} + d(1,k). \tag{40}$$

In view of (39), we have

$$\sum_{i \in J_{nk}} A_i + p = w_{12} + \sum_{i \in J_k} A_i + p \tag{41}$$

where p is the contribution of the addenda in Ψ.

Therefore

$$\sum_{i\in J_{nk}} A_i = \sum_{i\in J_k\cup J_{nk}} A_i + \sum_{i\in J_k} A_i. \tag{42}$$

That is, $\sum_{i\in J_k} A_i = 0$. Hence, $J_k = \emptyset$, a contradiction. It follows that

$$\|(\xi_2, \xi_3, ..., \xi_n)\| = w_{12}\,|\xi_2| + \Psi(\xi_3, ..., \xi_n) \tag{43}$$

where the function Ψ is still a semi-norm of class (T).

On the other hand, by Th. 4, under the choice of vertex $\bar{x} = 1$, we get

$$\|\xi\|_{Æ} = w_{12}\,|\xi_2| + \|\xi\|_{Æ}^{n-1}, \tag{44}$$

where $\|\cdot\|_{Æ}^{n-1}$ is the distance generated by the tree $T_n \setminus \{2\}$.

Therefore, the equality $\|\cdot\| = \|\cdot\|_{Æ}$ is equivalent to $\Psi = \|\cdot\|_{Æ}^{n-1}$, which in turn is true by induction hypothesis. □

Example 5. A semi-norm of class (T), see (37), does not necessarily generate a metric which is compatible with any graph in X. Let $X = \{1, 2, 3, 4\}$ and

$$\|\xi\| = |\xi_2 + \xi_3| + |\xi_3 + \xi_4| + |\xi_2|\,.$$

It is a norm and so induces a true distance on $\Delta(X)$. Nonetheless, $d(1, 2) = 2$ and there is no point z for which $d(1, z) = 1$ and $d(z, 2) = 1$.

6.1. **Barycentre.** Given a probability function λ on the vertices X of a weighted tree, a barycentre is a vertex $\hat{x} \in X$ such that the K-distance between λ and the delta function of that vertex is minimal, namely

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

Baycentres of probability measures on metric spaces are used in various statistical applications. See [9] for the specific example of weighted trees.

If we take $\bar{x}$ as the root of the tree, Eq. (17) yields

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

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

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

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,

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

$$\sum_{y\in X} \lambda(y) \sum_{x\preceq y} d(x, x^+) = \sum_{y\in X} \lambda(y) d(y, \bar{x}) = \mathbb{E}_\lambda\left[d(x, \bar{x})\right]\ .$$

Consequently, the baricenter x_B will be that root that minimizes the λ-mean distance of vertices from the root itself, i.e.,

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

6.2. **Probability trees.** A tree $(X, \mathcal{E})$ is a probability tree if it is a rooted directed tree with weights $p(x,y) > 0$ such that $\sum_{y \in \mathrm{ch}(x)} p(x,y) = 1$. We essentially follow the set-up of [7, Ch. 3] where each vertex is thought of as a *situation* or as a state-of-affairs. On such a tree we can define a specially adapted distance as follows.

Given any two vertices x and y, consider their least common ancestor z and the minimal path from x to z namely, $x = x_l, \dots, x_0 = z = y_0, \dots, y_m = y$. The product of transition probabilities along the path, that is $\prod_{i=1}^l p(x_{i-1}, x_i) \prod_{j=1}^m p(y_{j-1}, y_j)$, is the probability that two independent individuals moving down the tree hold situation x and situation y, respectively, conditional to the fact they have shared all situations up to z. The $-\log$-score of such conditional probability provides a distance between x and y based on the weights $w(x,y) = -\log p(x,y)$, $xy \in \mathcal{E}$.

Under such a distance, a function u is 1-Lipschitz if on all edges $\mathrm{e}^{-|u(y)-u(x)|} \le p(x,y)$, cf. Prop. 6. Moreover, from Prop. 8, u is extreme if, and only if, equality holds, $\mathrm{e}^{-|u(y)-u(x)|} = p(x,y)$.

A probability tree defines a probability on the set of root-to-leaf paths $w = x_0, \dots, x_l$. If u is an extreme 1-Lipschitz function, then the following representation of probabilities holds,

$$p(w) = \prod_{j=1}^{l} p(x_{j-1}, x_j) = \exp\left(\sum_{j=1}^{l} |u(x_j) - u(x_{j-1})|\right) .$$

Let us consider now the dynamic picture. The probability characterizes the evolution of an individuals in a population along state-of-affairs. Each individual waits for a random time, then moves to the next state-of-affairs with the given conditional probabilities. A probability on states-of-affairs X is a distribution of such a population at a given time consisting of individuals at different stages of evolution.

In this case, the Kantorovich distance is given by Eq. (17),

$$d(\mu,\nu) = \sum_{x \neq x_0} -\log p(x^+, x) \left|\sum_{z \succeq x} (\mu(z) - \nu(z))\right| = -\log\left(\prod_{x \neq x_0} p(x^+, x)^{\left|\sum_{z \succeq x}(\mu(z)-\nu(z))\right|}\right) .$$

It is remarkable that the negative exponential of the K-distance is a monomial in the transition probabilities,

$$\mathrm{e}^{-d(\mu,\nu)} = \prod_{x \neq x_0} p(x^+, x)^{\left|\sum_{z \succeq x}(\mu(z)-\nu(z))\right|}$$

so that any computation of sensitivity of the distance with respect to the transition probabilities is feasible.

7. Spanning trees

An almost natural extension of the approach undertaken so far consists in computing the K-distance by using the set of spanning trees of a given arbitrary graph. We will here show that the Æ-norm of a graph turns out to be the envelope of the Æ-norms of its spanning trees.

If $G = (X, \mathcal{E})$ is a graph, a spanning tree of G is a tree $T_i = (X, \mathcal{E}_i)$ where $\mathcal{E}_i \subset \mathcal{E}$. In other words T_i is a tree covering the graph G with the minimum number of edges. See [5, §1.2] for some more about spanning trees.

The inclusion relation $\mathcal{E}_i \subset \mathcal{E}$ implies that $d \le d_i$, where d and d_i are the distance in X induced by G and T_i, respectively. Hence, it holds the obvious relation $\|\xi\|_G \le \|\xi\|_{T_i}$, with regard to the two Arens-Eells norms defined on the space Æ(X). Therefore, denoting by $ST \equiv ST(G) = \{T_i\}$ the totality of the spanning trees of the graph G, it follows that

$$\|\xi\|_G \leq \min_{T_i \in ST} \|\xi\|_{T_i} .$$

In order to show that the above inequality is in fact an equality, we put before the following lemma.

Lemma 2. *Let G be a connected graph and F be a forest contained in G. There exists a spanning tree of G which extends F.*

Proof. Let the forest F have k components $T_1, T_2, \dots, T_k$, where each T_i is a tree. Let us enumerate these trees so that $d(T_1, T_2) \leq d(T_1, T_i)$ for $i = 2, \dots, k$. Pick vertices $x_1 \in V(T_1)$ and $x_2 \in V(T_2)$ for which $d(x_1, x_2) = d(T_1, T_2)$. Since G is connected, there exists a geodesic path $y_1, y_2, ..., y_r$ from $y_1 = x_1$ to $y_r = x_2$. By construction, the points $\{y_1, y_2, ..., y_r\}$ are not connected, in G, to the trees T_i, $i > 2$. Moreover, if we add to $T_1 \cup T_2$ the (bridge) path $y_1, y_2, ..., y_r$, we get a tree T_{12}. We have thus obtained a new forest $T_{12}, T_3, \dots, T_k$ with $k - 1$ components. Continuing this procedure, we get finally to a tree T which extends the forest F. If $\#V(T) = \#V(G)$, we are done, and T is the desired spanning tree. Otherwise, we have still to extend the tree T. The method is similar to the previous one. Take a vertex $x \in V(G)$ such that $d(x, T) = 1$ By adding this vertices and the corresponding edge we obtain a tree with one more vertices (for this method consult also [5, p. 10]), and once again, after a finite number of steps, we get the desired spanning tree of G. □

Theorem 8. *The Arens-Eells norm of any connected graph G is the envelope of the norms of its spanning trees. That is,*

$$\|\xi\|_G = \min_{T_i \in ST} \|\xi\|_{T_i} . \tag{45}$$

Proof. Th. 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 $F = (Y, \mathcal{F})$ is an acyclic subgraph of $G = (X, \mathcal{E})$. In other words, $(Y, \mathcal{F})$ is a forest. By Lemma 2, 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 $\mathcal{T}$, by assigning the value $a(x,y) = 0$, outside $\mathcal{F}$, we can re-write

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

where the last equality follows from point (i) of Th. 3, as the pair of vertices x and y is close, as long as $a(x,y) \neq 0$.

To conclude,

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

that proves our assertion. □

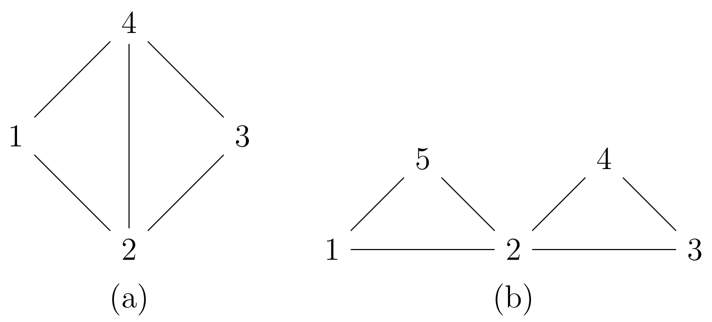


Figure 3. (a) Graph of Example 7. (b) Graph of Example 4

7.1. **A worked out example: cycle graph.** It should be of some interest to elaborate by hand a few examples that illustrate the properties stated by Th. 8.

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.

The weight matrix W determines a metric d on the set of vertices X. The distance $d(i,j)$, between the two vertices i and j, will be denoted by d_{ij}.

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 the Arens-Eells norm $\|\cdot\|_{C_n}$ of C_n as well as, by a constructive way, the above mentioned envelope property.

Proposition 14. *Let C_n be the cycle graph of order n, and*

$$\Phi(t) = \sum_{i=1}^{n} |t + \xi_1 + \xi_2 + \cdots + \xi_i| \, d_{i,i+1}$$

be the scalar function with domain $t \in \mathbb{R}$, $\xi \in Æ(X)$, and $n+1 \equiv n$. Then

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

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

Proof. Thanks to Th. 3, the restriction of set of elements $a(x,y)$ leads to the representations

$$\xi = a_{12}(\delta_1 - \delta_2) + a_{23}(\delta_2 - \delta_3) + \cdots + a_{n1}(\delta_n - \delta_1).$$

By inverting the previous relation and introducing the parameter $t = a_{n1}$, we get easily

$$a_{i,i+1} = t + \xi_1 + \xi_2 + \cdots + \xi_i$$

for $i = 1, 2, \ldots, n$. This implies that every vector ξ admits ∞^1-many representations, and formula (13) for the Arens-Eells 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 $t = -\xi_1 - \xi_2 - \cdots - \xi_i$, $(i = 1, 2, ..., n)$, 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 $-\xi_1-\xi_2-...-\xi_j$, then

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

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. □

Example 6. Prop. 14 entails that the norm associated with the cycle graph C_3

$$1\to 2\to 3\to 1,$$

is given by $\|\xi\|_{C_3}=\min_{i=1,2,3}\|\xi\|_{T_i}$, where

$$\|\xi\|_{T_1}=|\xi_2|\,d_{23}+|\xi_1|\,d_{31},\quad \|\xi\|_{T_2}=|\xi_2|\,d_{12}+|\xi_3|\,d_{31},\ \|\xi\|_{T_3}=|\xi_1|\,d_{12}+|\xi_3|\,d_{23}.$$

It is easy to check that $\|\cdot\|_{T_1}$ is the norm generated by the linear tree $1\to 3\to 2$, while $\|\cdot\|_{T_3}$ is by $1\to 2\to 3$.

Example 7. The method adopted in Prop. 14 of calculating the norm, may be duplicated for any graph for which it holds $\#X=\#\mathcal{E}=n$, such as cycle graphs. However, the case in which $\#\mathcal{E}>n$ is more interesting and clearly the function Φ will be no longer a scalar one.

By way of example, consider the two-cycles graph of Fig. 3(a), where $\#X=4$ and $\#\mathcal{E}=5$. If for sake of simplicity we assume that the graph is un-weighted, the function to minimize turns out to be

$$\Phi(t,u)=|u|+|t|+|\xi_1-u|+|\xi_1+\xi_2-t+u|+|u-t-\xi_4|$$

with $(t,u)\in\mathbb{R}^2$. After tedious algebra, the norm of the vector $\xi=(\xi_1,\xi_2,\xi_3,\xi_4)$ turns out to be the minimum of the following 8 functionals

$$\begin{gathered}|\xi_1|+|\xi_2|+|\xi_1+\xi_4|\\ |\xi_1|+|\xi_4|+|\xi_1+\xi_2|\\ |\xi_1|+|\xi_3|+|\xi_1+\xi_4|\\ |\xi_1|+|\xi_3|+|\xi_1+\xi_2|\\ |\xi_2|+|\xi_3|+|\xi_1+\xi_2|\\ |\xi_3|+|\xi_4|+|\xi_1+\xi_4|\\ |\xi_1|+|\xi_3|+|\xi_4|\\ |\xi_1|+|\xi_2|+|\xi_3|\end{gathered}$$

which are just the norms of the 8 spanning trees of the graph.

8. Conclusion

We have considered both the primal and the dual theory of the Kantorovich distance for finite metric spaces in which the distance is generated by a weighted graph. Under general conditions, the support is a looped tree. This "smallness" property is a reminders of the existence of a Monge solution in the continuous case.

The Arens-Eells theory shows that the K-distance is actually the restriction of a distance associated to a norm, and provides precise results about the extreme points of the set of optimal couplings. In particular, this was used to extend the well known case of closed form expression of the distance from linear graphs to trees. A possible application of interest could be the computation of the distance between probabilities on a given probability tree.

The result on spanning trees suggest an algorithm to compute the K-distance consisting in generating all spanning trees of the given graph, followed by the application of the closed form equation for the distance on a tree. This has been illustrated with small examples. The actual comparison between the computational effort of known solution methods versus the spanning tree method was not done here. This is possibly worth further research in the future.

Acknowledgments

The authors thank Luigi Malagò (RIST Cluj-Napoca) for suggesting relevant references. G. Pistone acknowledges the support of de Castro Statistics and Collegio Carlo Alberto. He is a member of INdAM-GNAMPA.

References

1. Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows in metric spaces and in the space of probability measures*, second ed., Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2008. MR 2401600 (2009h:49002)
2. Satoshi Aoki, Hisayuki Hara, and Akimichi Takemura, *Markov bases in algebraic statistics*, Springer Series in Statistics, Springer, New York, 2012. MR 2961912
3. Richard F. Arens and James Eells, Jr., *On embedding uniform and topological spaces*, Pacific J. Math. **6** (1956), 397–403. MR 0081458
4. Alexander Barvinok, *A course in convexity*, Graduate Studies in Mathematics, vol. 54, American Mathematical Society, Providence, RI, 2002.
5. Béla Bollobás, *Modern graph theory*, Graduate Texts in Mathematics, vol. 184, Springer-Verlag, 1998. MR MR1633290 (99h:05001)
6. Haïm Brezis, *Remarks on the Monge-Kantorovich problem in the discrete setting*, C. R. Math. Acad. Sci. Paris **356** (2018), no. 2, 207–213. MR 3758722
7. Rodrigo A. Collazo, Christiane Görgen, and Jim Q. Smith, *Chain event graph*, Computer Science and Data Analysis Series, CRC Press, 2018.
8. Michel Marie Deza and Monique Laurent, *Geometry of cuts and metrics*, Algorithms and Combinatorics, vol. 15, Springer-Verlag, Berlin, 1997. MR 1460488
9. Steven N. Evans and Frederick A. Matsen, *The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **74** (2012), no. 3, 569–592. MR 2925374
10. Jeff D. Farmer, *Extreme points of the unit ball of the space of Lipschitz functions*, Proc. Amer. Math. Soc. **121** (1994), no. 3, 807–813. MR 1195718
11. Benoît R. Kloeckner, *A geometric study of Wasserstein spaces: ultrametrics*, Mathematika **61** (2015), no. 1, 162–178. MR 3333967
12. Wuchen Li and Guido Montúfar, *Natural gradient via optimal transport*, Information Geometry **1** (2018), no. 2, 181–214, arXiv:1803.07033.
13. Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Foundations and Trends in Machine Learning **11** (2019), no. 5–6, 355–607.

14. Giovanni Pistone and Maria Piera Rogantin, *The algebra of reversible Markov chains*, Ann. Inst. Statist. Math. **65** (2013), no. 2, 269–293. MR 3011623
15. Filippo Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Birkhäuser, 2015.
16. Ryszard Smarzewski, *Extreme points of unit balls in Lipschitz function spaces*, Proc. Amer. Math. Soc. **125** (1997), no. 5, 1391–1397. MR 1389537
17. Max Sommerfeld and Axel Munk, *Inference for empirical Wasserstein distances on finite spaces*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **80** (2018), no. 1, 219–238. MR 3744719
18. C. Villani, *Optimal transport: Old and new*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, 2008.
19. Cédric Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003. MR 1964483
20. Nik Weaver, *Lipschitz algebras*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2018, Second edition. MR 3792558

Appendix A. Graphs

A weight w on the finite set X is a symmetric mapping from $X \times X$ to non-negative reals. The support of a weight defines a (un-directed, simple) graph $G = (X, \mathcal{E})$ with edges $\mathcal{E} = \{\{x, y\} \mid w(x, y) > 0\}$. We write also $V(G) = X$ and $E(G) = \mathcal{E}$. It is customary to present the transition weight as a weight matrix $W = [w(x, y)]_{x,y \in X}$. The structure (X, W) is a weighted graph. Conversely, a (un-directed, simple) graph without loops i.e., edges of the form $\{x, x\}$, can be seen as a weighted graph in which all the weights on edges are equal to 1. In such a case, the weight matrix E equals the adjacency matrix of the graph. Sometimes, we denote the edge $\{x, y\}$ by $xy = yx$.

A path between x and y is a sequence of at least two vertices $x = x_0, \dots, x_n = y$ such that $x_{i-1}x_i$ is an edge, $i = 1, \dots, n$. A cycle is a path with $x_0 = x_n$. The length of the path is $\sum_{i=1}^n w(x_{i-1}, x_i)$. All graphs we consider are connected, that is, there is always a path connecting x to y.

Given a weighted graph, the distance $d(x, y)$ is defined to be the length of the shortest path connecting x and y. It is easy to see that it is actually a distance. If confusion can arise we will also write d_G or $d_{G,w}$.

A graph without cycles is called a tree if it is connected, a forest otherwise. In a tree, the length of a two-point path equals its weight. It is frequently useful to select a distinguished vertex, the root of the tree, and to provide each edge with a direction in such a way that the distance from the root vertex increases in that direction. In such a case, we write the edge as $x \to y$. Given a vertex x, the set of all y such that $x \to y$ (the children of x) is denoted $\operatorname{child}(x)$. The unique parent of a non-root vertex x is denoted by x^+. The partial order induced by a rooted tree is denoted by $x \preceq y$ or $y \succeq x$.

Since a coupling $\gamma \colon X \times X \to \mathbb{R}_+$ is not symmetric, we want to consider directed graphs. To each directed graph we can associate the graph whose adjacency matrix has elements with value equal one if and only if the corresponding weight is positive. A rooted tree, provided with the natural partial ordering is a directed graph.

Appendix B. Countable trees

We extend here our representation to locally finite countable trees, under the assumption that the diameter of X is finite.

Proposition 15. *For a countable tree*

$$\|\xi\|_{Æ} = \sum_{x \in X \setminus x_0} d(x, x^+) \, |\Xi(x)| = \sum_{y \in X} \xi(y) \sum_{y \succeq x} d(x, x^+) \operatorname{Sgn} \Xi(x), \tag{46}$$

holds for every $\xi = \mu - \nu$.

Proof. Let us aside at a moment the issue if all the differences between probability functions, $\xi = \mu - \nu$, lie in the space Æ(X). Anyway, also in this case, the elements of $\operatorname{ext}\mathrm{Lip}_1^+(d)$ are the functions u that satisfy

$$u(y) = \sum_{y \succeq x} d(x, x^+)\epsilon(x)$$

with $\epsilon : X \setminus x_0 \to \{-1, 1\}$. Therefore we have to maximize

$$\sum_{y \in X} \xi(y) \sum_{y \succeq x} d(x, x^+)\epsilon(x).$$

In order to interchange the order of the two summations we can here invoke Fubini theorem. It will be sufficient to check that the double sum

$$\sum_{x,y \in X} \xi(y) d(x, x^+)\epsilon(x) I(x, y)$$

is summable, where $I(x, y) = 1$ if $y \succeq x$. Actually,

$$\left| \sum_{x,y \in X} \xi(y) d(x, x^+)\epsilon(x) I(x, y) \right| \le \sum_{x,y \in X} |\xi(y)|\, d(x, x^+) \le 2K$$

where K is the diameter of X. Whence, we get to the desired relations (46).

Let us now complete the proof by checking that all the differences between probability measures are in Æ(X).

In fact, the relation $\mathrm{Lip}^+(d) \simeq l_1(X \setminus x_0)^*$ remains true also for countable trees. Hence, $l_1(X \setminus x_0)$ is a predual of $\mathrm{Lip}^+(d)$. It has been proved [20, Th. 3.26] that the predual of $\mathrm{Lip}^+(d)$ is unique, provided the diameter is finite. Therefore, under our conditions Æ(X) is isometrically isomorphic to $l_1(X \setminus x_0)$. It is then immediately seen that every $\xi = \mu - \nu$ lies in Æ(X). □