Comparison of distances between measures

Jean-Michel Morel [*],
Filippo Santambrogio [†]

April 25, 2006

Abstract

The problem of optimal transportation between a set of sources and a set of wells has become recently the object of new mathematical models generalizing the Monge-Kantorovich problem. These models are more realistic as they predict the observed branching structure of communication networks. They also define new distances between measures. The question arises of how these distances compare to the classical Wasserstein distance obtained by the Monge-Kantorovich problem. In this paper we show sharp inequalities between the d_α distance induced by branching transport paths and the classical Wasserstein distance over probability measures in a compact domain of $\mathbb{R}^m$.

The problem of the optimal mass transportation was introduced by Monge in the 18th century. Kantorovich gave it a first rigorous mathematical treatment. In the Monge-Kantorovich model, two probability measures μ^+ and μ^- (the source and target mass distributions) are given. Each particle of μ^+ travels on a straight line segment onto μ^- and the cost of the transportation to be minimized is the integral of the lengths of the individual paths. This variational model has received a lot of attention because of its remarkable mathematical properties [1], [10].

From the economical viewpoint the Monge-Kantorovich problem is rather unrealistic. In most transportation networks, the aggregation of particles on common routes is preferable to individual straight ones. Thus the local structure of human-designed distribution systems doesn't look as a set of straight wires but rather like a tree. This branching structure is observable in communication networks [5], drainage networks [7], pipelines [4] and in many natural systems like the blood circulation in mammals, the river basins and the trees.

The design of functionals for mass transportation by branched structures was first addressed in [5] as a discrete graph optimization problem with prescribed sources and well points. Recently, continuous models have been proposed for this same setting [9], [8] and [3]. We will describe in the sequel these models in a more detailed way. They all define a cost functional for the transportation between μ^+ and μ^-. The optimal value for this functional yields a distance between μ^+ and μ^-. Our aim here is to compare this new distance with the so called Wasserstein distance associated with the Monge-Kantorovich model.

This distance on probability measures owes its importance to the fact that, on compact domains, it gives a metric to the topology of weak convergence. Given two probability measures μ^+ and μ^- with support in a compact domain $C \subset \mathbb{R}^m$ this distance is obtained by minimizing the Monge-Kantorovich functional

$$\int_{C\times C} c(x,y)\pi(dx,dy)$$

among all probability measures π on $C\times C$ whose marginal measures are exactly μ^+ and μ^-. We denote by $\Pi(\mu^+,\mu^-)$ this set of probabilities

$$\Pi(\mu^+,\mu^-) = \{\pi \in \mathcal{P}(C\times C) : X^+_\sharp \pi = \mu^+ \text{ and } X^-_\sharp \pi = \mu^-\},$$

where $X^\pm$ are the projections of $C\times C$ onto C, i.e. $X^+(x,y) = x$ and $X^-(x,y) = y$. The function $c : C\times C$ is a given cost function and its semicontinuity is sufficient for the existence of an optimal measure $\pi_0 \in \Pi(\mu^+,\mu^-)$ which is called optimal transport plan. When $c(x,y) = |x-y|$ the minimum value of this problem is denoted by $W_1(\mu^+,\mu^-)$ and it defines a distance over the space $\mathcal{P}(C)$ of probability measures on

[*] CMLA, ENS Cachan, 61, Av. du Président Wilson 94235 Cachan Cedex France, morel@cmla.ens-cachan.fr
[†] Scuola Normale Superiore, Piazza dei Cavalieri, 7 56126 Pisa Italy, santambrogio@sns.it

C. The index 1 is due to the fact that one can produce other distances W_p by considering $c(x,y) = |x-y|^p$ and then raising the infimum to the power of $1/p$ [10].

Let us now define more general transport problems modeling branched structures. Given two discrete mass distributions $\mu^+ = \sum_{i=1}^m a_i \delta_{x_i}$ and $\mu^- = \sum_{j=1}^m b_j \delta_{y_j}$ Gilbert [5] considers the minimization problem

$$\inf_G \quad \sum_h w_h^\alpha \mathcal{H}^1(e_h), \tag{1}$$

where the infimum is taken among all weighted oriented graphs G with edges e_h and weights w_h such that at each segment vertex which is not one of the x_i's or y_j's the total incoming mass equals the outcoming, while in each x_i, $a_i + \textit{incoming mass} = \textit{outcoming mass}$ and conversely, in each y_j $\textit{incoming mass} = \textit{outcoming mass} + b_j$. These conditions are nothing but the Kirchhoff law for circuits. The exponent α is a fixed parameter $0 < \alpha < 1$ so that the function $t \mapsto t^\alpha$ is concave and sub-additive and therefore favors the aggregation of routes. This problem was presented in [5] or [6] as an extension of Steiner's minimal length problem. A comparison of the typical structures arising in Gilbert and Monge's models is shown in picture 1.

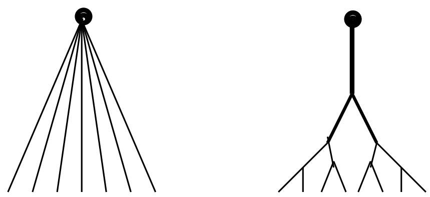

Figure 1: Monge's straight line solution (left) vs Gilbert's branching one (right)

There are several ways to extend this discrete functional to the case of arbitrary measures $\mu^\pm \in \mathcal{P}(C)$. In [9] the problem is presented as an extension of the Monge-Kantorovich case (which corresponds somehow, for $c(x,y) = |x-y|$, to $\alpha = 1$) by a relaxation procedure. The constraint on the incoming and outcoming masses at each vertex may be easily written as $\nabla \cdot \lambda_G = \mu - \nu$, where $\lambda_G = \sum_h w_h [[e_h]]$ is a $1-$current. The term $[[e]]$ denotes the integration on the segment e with orientation given by the direction of e. According to this language, it can be proven that the Gilbert problem becomes in a continuous framework

$$\min_{\partial T = \mu^+ - \mu^-} M^\alpha(T) = \int_M \theta^\alpha \, d\mathcal{H}^1, \tag{2}$$

among all rectifiable currents $T = (M, \theta, \xi)$ with prescribed boundary. The minimum value, which obviously depends on μ^+ and μ^-, will be denoted by $d_\alpha(\mu^+, \mu^-)$. In [9] it is proven that d_α defines a new distance over the space of probability measures $\mathcal{P}(C)$, which induces the weak topology as well, provided $\alpha > 1 - 1/m$. If α is under this threshold it may happen that the infimum is in fact $+\infty$. Other formalizations by means of probabilities on the set of Lipschitz curves in C and yielding an equivalent model may be found in [8] for the case when one of the measures is a single source, say $\mu^+ = \delta_0$, and in [3].

The two distances that we have introduced so far, d_α and W_1, induce the same topology on $\mathcal{P}(C)$, which is the same induced by the weak convergence. It is easily checked [9] that $W_1 \le d_\alpha$. This inequality is optimal as can be checked by taking two close by Dirac masses. The purpose of this note is to give a sharp quantitative estimate of the kind $d_\alpha \le C(W_1)^\beta$. This question was raised as a conjecture by Cedric Villani while reviewing the PhD Thesis [2]. Such an inequality gives an a priori estimate on d^α which is numerically relevant. Indeed W_1 is much easier to compute by linear programming than d_α, which is a non-convex optimization problem.

This estimate, as we avoid using previous results on the topology induced by these distances (i.e. no density argument) gives a direct and quantitative proof of the equivalence between the weak convergence topology and the topology defined by d_α. In fact the only property on d_α we will use is the following: if μ^+ and μ^- are two nonnegative measures on a domain ω with the same total mass M, then their distance d_α (which may easily be extended to positive finite measures) can be estimated through $d_\alpha(\mu^+, \mu^-) \le C_{\alpha,m} M^\alpha \operatorname{diam}(\omega)$, under the important assumption $\alpha > 1 - 1/m$. The proof of this property is easy. It follows from the explicit construction of an irrigation fractal dyadic tree connecting any probability measure on C to a Dirac mass and it can be found for instance in [9].

To fix ideas, we consider two probability measures μ^+ and μ^- with support in a m-dimensional cube C with edge 1, say $C = [0,1]^m$. It is not difficult to scale the result to any bounded domain in $\mathbb{R}^m$.

Proposition 0.1. *The following inequality holds for* $1 > \alpha > 1 - \frac{1}{m}$*:*

$$d_\alpha(\mu^+, \mu^-) \le c W_1(\mu^+, \mu^-)^{m(\alpha-(1-1/m))},$$

where c *denotes a suitable constant depending only on* m *and* α*.*

We shall see in Example 0.1 that this inequality is sharp.

Proof. Let $\pi_0 \in \mathcal{P}(C)$ be an optimal transport plan between μ^+ and μ^-, where we denote by Ω the product space $C \times C$. We also denote by X^+ and X^- the two projections from Ω onto C, so that $X^+(x,y) = x$, $X^-(x,y) = y$ and $X^\pm_\sharp \pi_0 = \mu^\pm$. In what follows we set $\delta = W_1(\mu^+, \mu^-)$ and

$$\Omega_i = \left\{(x,y) \in C \times C = \Omega,\ (2^i - 1)\frac{\delta}{2} \le |x-y| < (2^{i+1}-1)\frac{\delta}{2}\right\}.$$

We can limit ourselves to consider those indices i which are not too large, i.e. up to $(2^i - 1)\frac{\delta}{2} \le \sqrt{m}$ (being $\sqrt{m}$ the diameter of C). Let I be the maximal index i so that this inequality is satisfied. $\Omega = \cup_{i=0}^I \Omega_i$ is a disjoint union and

$$\sum_{i=0}^{I} (2^i - 1)\frac{\delta}{2}\pi_0(\Omega_i) \le W_1(\mu^+, \mu^-) = \delta \le \sum_{i=0}^{I} (2^{i+1}-1)\frac{\delta}{2}\pi_0(\Omega_i) \tag{3}$$

We call cube with edge e any translate of $[0, e[^m$. For each $i = 0, \cdots, I$, using a regular grid in $\mathbb{R}^m$, one can cover C with disjoint cubes $C_{i,k}$ with edge $(2^{i+1}-1)\delta$. The number of the cubes in the $i-$th covering may be easily estimated by

$$\left(\frac{1}{(2^{i+1}-1)\delta} + 1\right)^m \le \left(\frac{c}{(2^{i+1}-1)\delta}\right)^m = K(i). \tag{4}$$

For each index i, it holds $C \subset \cup_{k=1}^{K(i)} C_{i,k}$ and the cubes are disjoint. Let us set

$$\Omega_{i,k} = (C_{i,k} \times C) \cap \Omega_i, \quad \mu^+_{i,k} = X^+_\#(\pi_0 1\!\!1_{\Omega_{i,k}}) \text{ and } \mu^-_{i,k} = X^-_\#(\pi_0 1\!\!1_{\Omega_{i,k}}).$$

We have just cut μ^+ and μ^- into pieces. Let us call informally μ_i^+ the pieces of μ^+ for which the Wasserstein distance to the corresponding part μ_i^- of μ^- is of order $2^i\frac{\delta}{2}$. Then $\mu^+_{i,k}$ is the part of μ_i^+ whose support is in the cube $C_{i,k}$. What we have now gained is that each $\mu^+_{i,k}$ has a specified diameter of order $2^i\delta$ and is at a distance to its corresponding $\mu^-_{i,k}$ which is of the same order $2^i\delta$ (see picture 2). Let us be a bit more precise. The support of $\mu^+_{i,k}$ is a cube with edge $(2^i - 1)\delta$. By definition of Ω_i, the maximum distance of a point of $\mu^-_{i,k}$ to a point of $\mu^+_{i,k}$ is less than $(2^{i+1}-1)\frac{\delta}{2}$. Thus the supports of $\mu^-_{i,k}$ and $\mu^+_{i,k}$ are both contained in a same cube with edge $6 \cdot 2^i\delta$.

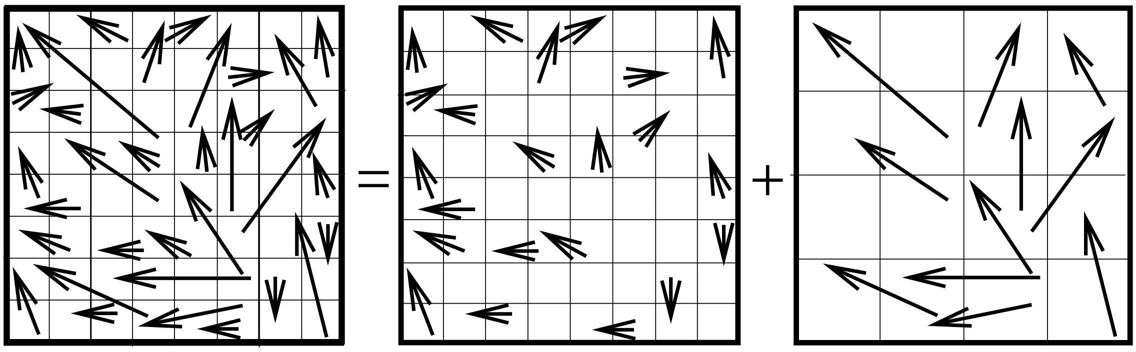

Figure 2: Decomposition of Monge's transportation into the sets $\Omega_{i,k}$

By the scaling properties of the d_α distance we deduce that for some constant c, depending only on α and m, it holds (see [9]):

$$d_\alpha(\mu^+_{i,k}, \mu^-_{i,k}) \leq c2^i \delta \pi_0(\Omega_{i,k})^\alpha.$$

From this last relation, the sub-additivity of d_α, Hölder inequality, (3) and the bound on $K(i)$ given in (4), one obtains in turn

$$\begin{aligned}
d_\alpha(\mu^+, \mu^-) &\leq \sum_{i,k} d_\alpha(\mu^+_{i,k}, \mu^-_{i,k}) \\
&\leq \sum_{i,k} c2^i \delta \pi_0(\Omega_{i,k})^\alpha = c \sum_{i,k} (2^i \delta \pi_0(\Omega_{i,k}))^\alpha (2^i \delta)^{1-\alpha} \\
&\leq c(\sum_{i,k} (2^i \delta \pi_0(\Omega_{i,k})))^\alpha (\sum_{i,k} 2^i \delta)^{1-\alpha} \\
&\leq c(\sum_{i} (2^i \delta \pi_0(\Omega_{i})))^\alpha (\sum_{i=0}^{I} K(i) 2^i \delta)^{1-\alpha} \\
&\leq c(\delta)^\alpha \left(\sum_{i=0}^{I} \left(\frac{c}{(2^{i+1}-1)\delta} \right)^m 2^i \delta \right)^{1-\alpha} \\
&\leq c\delta^{\alpha+(1-m)(1-\alpha)} \left(\sum_{i=0}^{I} 2^{i(1-m)} \right)^{1-\alpha} \\
&\leq c\delta^{\alpha m-(m-1)} = cW_1(\mu^+, \mu^-)^{\alpha m-(m-1)},
\end{aligned}$$

where c denotes various constants depending only on m and α and where the last two inequalities are valid if $m \geq 2$ so that the series $\sum_{i=0}^\infty 2^{i(1-m)}$ is convergent.

In the case $m = 1$ a different proof is needed. In this case we know how does an optimal transportation for $d_\alpha(\mu^+, \mu^-)$ look like. We refer to the formulation in (2), which in the one-dimensional setting gives

$$d_\alpha(\mu^+, \mu^-) = \int_0^1 |\theta(x)|^\alpha dx.$$

The function θ plays the role of the multiplicity and it is given by

$$\theta(x) = \mu([0, x]), \quad \mu := \mu^+ - \mu^-,$$

as a consequence of its constraint on the derivative. Hence we have

$$d_\alpha(\mu^+, \mu^-) = \int_0^1 |\mu([0, x])|^\alpha dx \leq \left[\int_0^1 |\mu([0, x])| dx \right]^\alpha,$$

where the inequality comes from Jensen inequality. Then we set $A = \{x \in [0, 1] : \mu([0, x]) > 0\}$ and $h(x) = 1\!\!1_A(x) - 1\!\!1_{[0,1]\setminus A}(x)$ and we have

$$\begin{aligned}
\int_0^1 |\mu([0, x])| dx &= \int_0^1 \mu([0, x]) h(x) dx = \int_0^1 h(x) dx \int_0^1 1\!\!1\{t \leq x\} \mu(dt) \\
&= \int_0^1 \mu(dt) \int_t^1 h(x) dx = \int_0^1 u(t) \mu(dt) \leq W_1(\mu^+, \mu^-),
\end{aligned}$$

where $u(t) = \int_t^1 h(x) dx$ is a Lipschitz continuous function whose Lipschitz constant does not exceed 1 as a consequence of $|h(x)| \leq 1$. Thus the last inequality is justified by the duality formula (see [10], Theorem 1.14, page 34):

$$W_1(\mu^+, \mu^-) = \sup_{v \in Lip_1} \int_0^1 v \, d(\mu^+ - \mu^-).$$

Hence it follows easily $d_\alpha(\mu^+, \mu^-) \leq W_1(\mu^+, \mu^-)^\alpha$, which is the thesis for the one dimensional case. $\square$

As we announced, the result in Proposition 0.1 is sharp as far as estimates of d_α in terms of W_1 are concerned. The assumption $\alpha > 1 - 1/m$ cannot be removed since, for $m \geq 2$, if we remove this assumption, the quantity d_α could be infinite while W_1 is always finite. In dimension 1 the only uncovered case is $\alpha = 0$. In this case d_α is in fact always finite but, for instance if $\mu^+ = \delta_0$ and $\mu^- = (1-\varepsilon)\delta_0 + \varepsilon\delta_1$ it holds $d_\alpha(\mu^+, \mu^-) = 1$ while $W_1(\mu^+, \mu^-) = \varepsilon$. As ε is as small as we want, this excludes any desired inequality. Hence we get back to the m-dimensional case where the result cannot be improved as far as α is concerned. On the other hand the exponent $m(\alpha - (1 - 1/m))$ cannot be improved as can be seen from the following example.

Example 0.1. *There exists a sequence of pairs of probability measures (μ_n^+, μ_n^-) on the cube C such that*

$$d_\alpha(\mu_n^+, \mu_n^-) = cn^{-m(\alpha-(1-1/m))} \text{ and } W_1(\mu_n^+, \mu_n^-) = c/n.$$

Proof. It is sufficient to divide the cube C into n^m small cubes of edge $1/n$ and to set $\mu_n^+ = \sum_{i=1}^{n^m} \frac{1}{n^m}\delta_{x_i}$ and $\mu_n^- = \sum_{i=1}^{n^m} \frac{1}{n^m}\delta_{y_i}$, where each x_i is a vertex of one of the n^m cubes (let us say the one with minimal sum of the $m-$coordinates) and the corresponding y_i is the center of the same cube. In this way y_i realizes the minimal distance to x_i among the y_j's. Thus the optimal configuration both for d_α and W_1 is given by linking any x_i directly to the corresponding y_i. In this way we have

$$\begin{aligned} d_\alpha(\mu_n^+, \mu_n^-) &= n^m \left(\frac{1}{n^m}\right)^\alpha \frac{c}{n} = cn^{-m(\alpha-(1-1/m))} \\ W_1(\mu_n^+, \mu_n^-) &= n^m \frac{1}{n^m}\frac{c}{n} = \frac{c}{n}, \end{aligned}$$

where $c = \frac{\sqrt{m}}{2}$. □

One can deduce easily inequalities between d_α and W_p by using standard inequalities between W_1 and W_p, namely $CW_p^p \leq d_\alpha \leq CW_p^{m(\alpha-(1-1/m))}$. The right hand inequality is sharp by using again example 0.1. It is not clear instead whether the left-hand inequality is optimal.

References

[1] L. Ambrosio: Lecture Notes on Optimal Transport Problems. Lecture Notes in Mathematics "Mathematical aspects of evolving interfaces" *CIME Series, Madeira 2000*. P. Colli and J.F. Rodrigues Eds., pp. 1-52, 2003.

[2] M. Bernot, Optimal transport and irrigation, *PhD Thesis*, ENS Cachan, 2005, available at `http://perso.crans.org/~bernot`.

[3] M. Bernot, V. Caselles and J.-M. Morel, Traffic plans, *Publicacions Matemàtiques* (49), no. 2, 417–451, 2005.

[4] S. Bhashkaran and F. J. M. Salzborn, Optimal design of gas pipeline networks, *J. Oper. Res. Society* (30), 1047–1060, 1979.

[5] E. N. Gilbert, Minimum cost communication networks, *Bell System Tech. J.* (46), 2209–2227, 1967.

[6] E. N. Gilbert and H. O. Pollak, Steiner minimal trees. *SIAM J. Appl. Math.* (16), 1–29, 1968.

[7] D. H. Lee, Low cost drainage networks, *Networks* (6), 351–371, 1976.

[8] F. Maddalena, S. Solimini and J.-M. Morel, A variational model of irrigation patterns. *Interfaces and Free Boundaries* (5), 391–415, 2003.

[9] Q. Xia, Optimal Paths related to Transport Problems. *Comm. Cont. Math.* (5), no. 2, 251–279, 2003.

[10] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics, AMS, 2003.