An Elementary Introduction to Monotone Transportation*

K. Ball

Department of Mathematics, University College London, Gower Street, London WC1E 6BT, UK kmb@math.ucl.ac.uk

Towards the end of the 18^{th} century Gaspard Monge posed a problem concerning the most efficient way to move a pile of sand into a hole of the same volume. In modern language the problem was something like this:

Given two probability measures μ and ν on $\mathbf{R}^2$ (or more generally $\mathbf{R}^n$), find a map $T : \mathbf{R}^2 \to \mathbf{R}^2$ which transports μ to ν and which minimises the total cost

$$\int \|x - Tx\| \, d\mu(x).$$

(Thus the cost of moving a unit mass from x to Tx is just the distance moved.)

The statement that T transports μ to ν means that for each measurable set A,

$$\mu\left(T^{-1}(A)\right) = \nu\left(A\right) \tag{1}$$

or equivalently, that for any bounded continuous real-valued function f

$$\int f\big(T(x)\big) \, d\mu(x) = \int f(x) \, d\nu(x). \tag{2}$$

Monge gave a number of examples to illustrate the difficulties of the problem. Plainly if arbitrary measures are allowed it may not be possible to transport at all. If μ is a single point mass and ν consists of two point masses, each with weight $1/2$, then no map T can "split μ in half". Where transportation maps do exist, there may be many different optimal ones, even for measures on the line. If μ consists of two equal masses at (say) 0 and 1, and ν consists of two equal masses at 2 and 3 then it does not matter whether we choose

$$\begin{matrix} 0 \to 2 \\ 1 \to 3 \end{matrix} \qquad \text{or} \qquad \begin{matrix} 0 \to 3 \\ 1 \to 2 \end{matrix} :$$

* This article is based on a lecture given to the graduate student seminar in analysis at Tel Aviv University.

S. Alesker et al.: LNM 1850, V.D. Milman and G. Schechtman (Eds.), pp. 41–52, 2004.

the total cost is 4 in each case. For general measures, it is difficult to demonstrate the existence of any optimal transportation map. In the late seventies, Sudakov [S] outlined an important new approach to the problem which has recently been implemented in a number of different ways: by Caffarelli, Feldman and McCann [CFM], Trudinger and Wang [TW] and Ambrosio [A] and in particular contexts by Ambrosio, Kirchheim and Pratelli [AKP]. I am indebted to Bernd Kirchheim for his guidance on the history of this problem.

One reason for the difficulty is that the condition (1) is highly non-linear in T. This difficulty can be overcome by relaxing the requirement that μ be transported to ν by a *map*. Instead of insisting that all the μ-measure at x should end up at the same place Tx, we may allow this mass to be "smeared out" over many points. The aim is thus to describe for each *pair* of points (x, y), how much mass is moved from x to y. More formally, we look for a measure γ on the product $\mathbf{R}^n \times \mathbf{R}^n$ whose marginal in the x direction is μ and whose marginal in the y direction is ν: so for measurable sets A and B,

$$\gamma(A \times \mathbf{R}^n) = \mu(A)$$

$$\gamma(\mathbf{R}^n \times B) = \nu(B).$$

A transportation measure is then optimal if it minimises

$$\int \|x - y\| \, d\gamma(x, y).$$

In this form, the problem is a linear program (for γ), whose discrete version is familiar from introductory courses on linear programming. The conversion of the transportation problem to a linear program was first described by Kantorovich. In this setting, the existence of optimal transport measures is not so hard to establish: the difficulty is to decide whether there is such a measure which is concentrated on the graph of a function T.

In the context of linear programming it is customary to consider transportation costs that are much more general than simply the distance $\|x - y\|$. The total cost will be

$$\int c(x, y) \, d\gamma(x, y)$$

if c measures the cost per unit mass of transporting from x to y. If c is a strictly convex function of the distance $\|x - y\|$ then examples of non-unique optimal transportation like the one above, do not occur. In [Br], Brenier explained that there is one particular choice of cost function for which not only is the optimal map unique but also it has a particularly special form, which makes it suitable for a wide range of applications. From the geometric point of view, this cost function is undoubtedly the "right" one. The cost in question is the square of the Euclidean distance: $c(x, y) = \|x - y\|^2$.

To see why this choice is so special, consider the problem of transporting a probability measure μ on $\mathbf{R}^2$ (which will be assumed to be absolutely continuous with respect to Lebesgue measure) to the probability measure ν which assigns mass 1/2 to each of the points $(1,0)$ and $(-1,0)$. The problem is illustrated in Figure 1: the shaded region on the left is supposed to represent the measure μ.

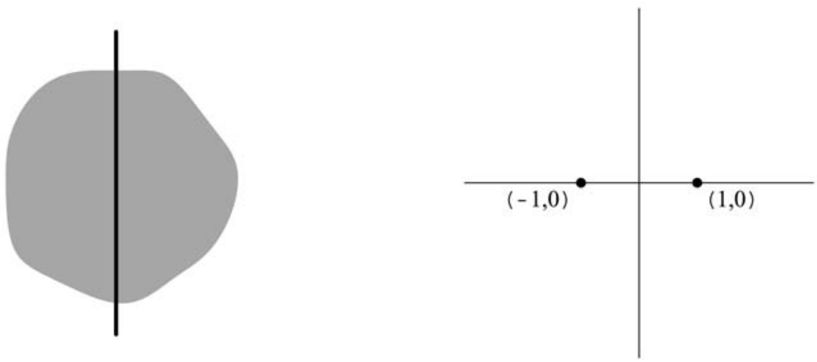

Fig. 1. A simple transport problem

The problem is to find a partition of $\mathbf{R}^2$ into sets A and B with

$$\mu(A) = \mu(B) = 1/2$$

so as to minimise the cost of transporting A to $(-1,0)$ and B to $(1,0)$:

$$\int_A \|x - (-1,0)\|^2 \, d\mu + \int_B \|x - (1,0)\|^2 \, d\mu.$$

I claim that the best thing to do is to divide the measure μ using a line in the direction $(0,1)$, as shown in Figure 1.

To establish the claim, we need to check that given two points (a,u) and (b,v) with $a < b$, it is better to move the leftmost point to $(-1,0)$ and the rightmost point to $(1,0)$, than it is to swap the order. The quadratic cost $\|x - y\|^2$ ensures this because it "keeps separate" the contributions from the first and second coordinates. The costs in the two cases are

$$(a+1)^2 + u^2 + (b-1)^2 + v^2$$

and

$$(a-1)^2 + u^2 + (b+1)^2 + v^2$$

and the only difference arises because $a < b$.

Because the transport function T maps everything to the right of the line, to the point $(1,0)$ and everything to the left of the line to the point $(-1,0)$, the map T is the gradient of a 'V'-shaped function as shown in Figure 2.

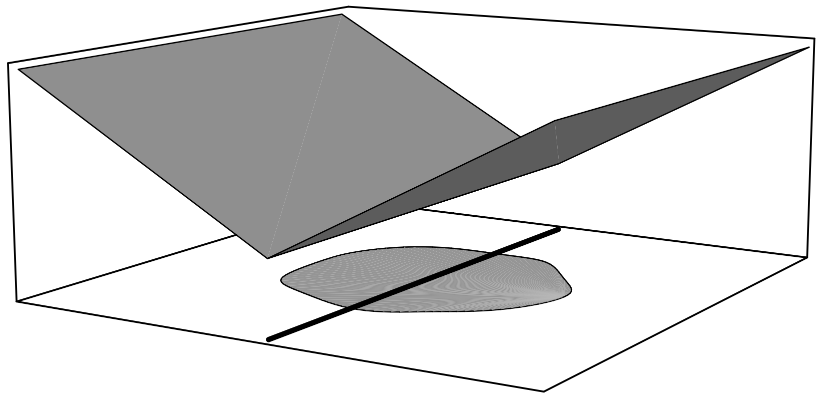

Fig. 2. Representing the transport as a gradient

More generally, the special feature of the Brenier map is that whenever we find a transport map which is optimal with respect to the quadratic cost, it will be the gradient of a convex function. Let's see why such a gradient should be optimal. Suppose that $\phi : \mathbf{R}^n \to \mathbf{R}^n$ is convex, $T = \nabla\phi$ and we have points $x_1, x_2, \ldots, x_k$ whose images under T are $y_1, y_2, \ldots, y_k$ respectively. The aim is to show that if $u_1, u_2, \ldots, u_k$ is a permutation of the x_i then

$$\sum \|x_i - y_i\|^2 \le \sum \|u_i - y_i\|^2 :$$

that the total cost goes up if we map the points any other way than by T. Plainly the inequality amounts to

$$\sum \langle x_i, y_i \rangle \ge \sum \langle u_i, y_i \rangle. \tag{3}$$

Now, since ϕ is convex, we have that for every x and u

$$\phi(u) \ge \phi(x) + \langle u - x, \nabla\phi|_x \rangle.$$

Hence

$$\sum \langle u_i - x_i, y_i \rangle \le \sum \big(\phi(u_i) - \phi(x_i)\big)$$

and the expression on the right is zero, since the u_i are a permutation of the x_i. This establishes (3).

Brenier demonstrated the existence of such optimal transport maps under certain conditions on the measures μ and ν: his result was generalised by McCann in [Mc1]. The Brenier map is sometimes called a monotone transportation map by analogy with the 1-dimensional case, in which the derivative of a convex function is monotone increasing. The rest of this article is organised as follows. In the next section we will demonstrate the existence of the

Brenier map under fairly general conditions on the measures, using an argument which is somewhat in the spirit of the example above. In the last section we shall explain how the special structure of the Brenier map fits neatly with a number of geometric inequalities.

There are a number of excellent surveys relating to mass transportation and monotone maps, for example the lecture notes of Villani [V].

1 A Construction of the Brenier Map

The aim of this section is to outline a proof of the following theorem:

Theorem 1. *If μ and ν are probability measures on $\mathbf{R}^n$, ν has compact support and μ assigns no mass to any set of Hausdorff dimension $(n-1)$, then there is a convex function $\phi : \mathbf{R}^n \to \mathbf{R}$, so that $T = \nabla\phi$ transports μ to ν.*

The proof has two main steps. We first establish the existence of ϕ in the case that the second measure ν is discrete and then pass to general measures by approximating them weakly by discrete ones. Suppose then that ν is a convex combination of point masses

$$\nu = \sum_1^m \alpha_i \delta_{u_i}.$$

The convex function we want has the form

$$\phi(x) = \max\{\langle x, u_i\rangle - s_i\}$$

for some suitable choice of real numbers $s_1, s_2, \ldots, s_m$. This function partitions $\mathbf{R}^n$ into m pieces according to whichever linear function is biggest. If A_i is the set where ϕ is given by

$$x \mapsto \langle x, u_i\rangle - s_i$$

then we want to juggle the s_i so as to arrange that $\mu(A_i) = \alpha_i$ for each i. In order to use a fixed point theorem we shall define a map H on the simplex of points $t = (t_1, t_2, \ldots, t_m)$ with non-negative coordinates satisfying $\sum t_i = 1$, by considering the function

$$\phi_t(x) = \max\left\{\langle x, u_i\rangle - \frac{1}{t_i}\right\}.$$

If t has non-zero coordinates, $H(t)$ will be the point $(\mu(A_1), \ldots, \mu(A_m))$ whose coordinates are the measures of the sets on which ϕ is linear. As t_i approaches 0, the i^{th} linear function drops to negative infinity and so $\mu(A_i)$ decreases to zero. In view of the hypothesis on μ, H can be defined continuously on the simplex and maps each face of the simplex into itself. It is a well-known consequence (or reformulation) of Brouwer's fixed point theorem, that such a map is surjective. (If the map omits a point, say in the interior of the simplex, then we can follow it by a projection of the punctured simplex onto its boundary, to obtain a continuous map from the simplex to its boundary which fixes each face. Now by cycling the coordinates of the simplex we obtain a continuous map with no fixed point.) So there is a choice of t for which $H(t)$ is $(\alpha_1, \ldots, \alpha_m)$.

Now suppose that we have a general probability ν. Approximate it weakly by a sequence (ν_k) of discrete measures and choose convex functions ϕ_k whose gradients transport μ to these approximants. Assume that the ϕ_k are pinned down by (for example) $\phi_k(0) = 0$. We may assume that all ν_k are supported on the same compact set from which it is immediate that the ϕ_k are equicontinuous. So we may assume that they have a locally uniform limit, ϕ, say. The function ϕ is convex.

A standard result in convex analysis guarantees that outside a set of Hausdorff dimension $n-1$, ϕ and all the ϕ_k are differentiable: (for the ϕ_k the differentiability is obvious by their construction). It is quite easy to check that except on the exceptional set,

$$\nabla\phi_k \to \nabla\phi.$$

Let $T_k = \nabla\phi_k$ for each k and $T = \nabla\phi$. By the condition on the support of ϕ, $T_k \to T$, μ-almost everywhere. We want to conclude that T transports μ to ν. This is a standard argument in weak convergence: if f is a bounded continuous function then

$$\begin{aligned} \int f\, d\nu &= \lim \int f\, d\nu_k \\ &= \lim \int f \circ T_k\, d\mu \\ &= \int f \circ T\, d\mu\ . \end{aligned}$$

Hence T transports μ to ν. □

The preceding theorem establishes the existence of a monotone transportation map T. Since T is the gradient of a convex function ϕ, its derivative is the Hessian of ϕ, so T' is positive semi-definite symmetric. Therefore $\det(T')$ is non-negative. This means that T is essentially 1-1. Therefore, if μ and ν have densities f and g respectively, the condition (1) for the Brenier map, has a local formulation as the familiar change of variables formula

$$f(x) = g(Tx).\det\big(T'(x)\big). \tag{4}$$

This version of the measure transportation property is particularly useful in a number of geometric applications such as those below.

2 The Brunn–Minkowski Inequality

The aim of this section and the following one is to illustrate why the Brenier map is the perfect tool for a number of geometric applications. The first such applications were found by McCann [Mc2] who gave a proof of the Brunn–Minkowski inequality using displacement convexity, which depends upon mass transportation, and Barthe [B] who used the Brenier map to give a very clear proof of the Brascamp-Lieb inequality [BL], and its generalisations proved in [L], and also of a reverse inequality that had been conjectured by the present author. There have been several other, striking, geometric applications, for example those in [CNV].

In this article I have chosen two somewhat simpler applications: to results which were originally proved in other ways but which illustrate quite well, just why monotone maps are so appropriate. The first is the Brunn–Minkowski inequality.

If A and B are non-empty measurable sets in $\mathbf{R}^n$ and $\lambda \in (0,1)$ then we define the set

$$(1-\lambda)A + \lambda B$$

to be

$$\big\{(1-\lambda)a + \lambda b : a \in A, b \in B\big\}.$$

The Brunn–Minkowski inequality states that

$$\mathrm{vol}\big((1-\lambda)A + \lambda B\big)^{1/n} \geq (1-\lambda)\mathrm{vol}(A)^{1/n} + \lambda \mathrm{vol}(B)^{1/n}.$$

By using the arithmetic/geometric mean inequality, we can deduce a multiplicative version of the Brunn–Minkowski inequality, which has a number of advantages (one of them being that it admits a number of simple proofs)

$$\mathrm{vol}\big((1-\lambda)A + \lambda B\big) \geq \mathrm{vol}(A)^{1-\lambda}\mathrm{vol}(B)^{\lambda}.$$

For fixed A, B and λ, this latter inequality is weaker than the Brunn–Minkowski inequality, but it is not hard to check that its truth for *all* A, B and λ implies the formally stronger statement.

Using the Brenier map it is easy to give a short explanation of the multiplicative form of the Brunn–Minkowski inequality, although to make the argument formal one requires regularity results for the Brenier map which are not so easy to derive. (Cafarelli [C1] and [C2] has obtained a powerful regularity theory for these maps which is more than sufficient for the applications described here.)

We may assume that both A and B have finite non-zero measure. Let μ and ν be the restrictions of Lebesgue measure to A and B respectively, rescaled by the volumes of these sets so as to have total measure 1. Let T be the Brenier map transporting μ to ν.

Equation (4) for the transportation shows that for each $x \in A$,

$$\frac{1}{\mathrm{vol}(A)} = \frac{1}{\mathrm{vol}(B)} . \det\left(T'(x)\right).$$

Now, let T_λ be the map given by

$$x \mapsto (1-\lambda)x + \lambda T(x).$$

This map transports μ to a measure, of total mass 1, supported on $(1-\lambda)A + \lambda B$. The density f_λ of this measure satisfies

$$\frac{1}{\mathrm{vol}(A)} = f_\lambda\left(T_\lambda(x)\right) . \det\left(T'_\lambda(x)\right).$$

In order to prove that $(1-\lambda)A + \lambda B$ is large, it suffices to check that the density is small: to be precise, that

$$f_\lambda\big(T_\lambda(x)\big) \le \frac{1}{\mathrm{vol}(A)^{1-\lambda}\mathrm{vol}(B)^\lambda}.$$

This will follow if

$$\begin{aligned} \det\left(T'_\lambda(x)\right) &\ge \left(\frac{\mathrm{vol}(B)}{\mathrm{vol}(A)}\right)^\lambda \\ &= \big(\det\big(T'(x)\big)\big)^\lambda . \end{aligned}$$

This says that for each x,

$$\det\big((1-\lambda)I + \lambda T'(x)\big) \ge \big(\det\big(T'(x)\big)\big)^\lambda$$

where I is the identity map on $\mathbf{R}^n$.

Now we use the fact that the Brenier map is monotone. At each x, $T'(x)$ is the Hessian of a convex function ϕ and so is a positive semi-definite symmetric matrix. With respect to an appropriate orthonormal basis, it is diagonal: let's say its diagonal entries are $t_1, t_2, \ldots, t_n$. The problem is to show that for these positive t_i

$$\prod (1-\lambda+\lambda t_i) \ge \left(\prod t_i\right)^\lambda .$$

This inequality is the special case of the Brunn–Minkowski inequality in which the set A is a unit cube and the set B is a cuboid with sides $t_1, \ldots, t_n$, aligned in the same way as the cube. It is immediate because the arithmetic/geometric mean inequality shows that for each i

$$(1-\lambda) + \lambda t_i \ge t_i^\lambda. \qquad \square$$

As long ago as 1957, Knothe [K] gave a proof of the Brunn–Minkowski inequality which involved a kind of mass transportation. His argument used induction upon dimension and corresponded, in effect, to the construction of a transportation map whose derivative is upper triangular rather than symmetric, as is the derivative of the Brenier map. Since such a choice of map is basis dependent, the proof is geometrically less elegant, but it has the advantage that regularity is not a serious issue.

Recently Alesker, Dar and Milman [ADM] used monotone transportation to give a proof of the Brunn–Minkowski inequality which was "constructive" in the following slightly exotic way. Normally each point of the combination $(1-\lambda)A + \lambda B$ can be realised as a combination of points (one from A and one from B) in many different ways. In [ADM] the authors construct a map which associates each point of the convex combination with a particular combination that generates it, and which allows them to compare directly the sets themselves, rather than just their volumes.

3 The Marton–Talagrand Inequality

In [M] and subsequent articles, Marton described a new method for demonstrating probabilistic deviation inequalities based upon a comparison between transportation cost and entropy. Her main focus was on Markov chains where the existing methods did not give any estimates and in this setting she considered transportation costs based on an L_1-type cost function rather than the L_2-type discussed in this article. In his article [T], Talagrand pointed out that Marton's method could be applied very elegantly to the isoperimetric inequality for Gaussian measure provided one uses the quadratic cost function. The aim of this section is to explain the Marton–Talagrand inequality using the Brenier map. As in the previous section, the argument will treat regularity of the maps naively. The fact that the Brenier map fits well with the Marton–Talagrand inequality was noticed independently by several people. The first available references seem to be [Blo] and [OV].

Let γ be the standard Gaussian measure on $\mathbf{R}^n$ with density

$$g(x) = \frac{1}{(\sqrt{2\pi})^n} e^{-|x|^2/2}.$$

For a density f on $\mathbf{R}^n$ we define the relative entropy of f (relative to the Gaussian) to be

$$\mathrm{Ent}(f\|\gamma) = \int_{\mathbf{R}^n} f \log(f/g).$$

The Marton–Talagrand inequality compares the relative entropy of f with the cost of transporting γ to the probability with density f. This cost is

$$C(g, f) = \int |x - T(x)|^2 \, d\gamma$$

where T is the Brenier map transporting γ to the measure with density f.

Theorem 2. *With the notation above*

$$\frac{1}{2}C(g,f) \le Ent(f\|\gamma).$$

Proof. Let T be the Brenier map transporting γ to the probability with density f. Then for each x

$$g(x) = f\big(T(x)\big)T'(x).$$

The relative entropy is

$$\int f(y)\log\big(f(y)/g(y)\big)\,dy$$

and after the change of variables $y = T(x)$ this becomes

$$\int f\big(T(x)\big)\log\left(\frac{f(T(x))}{g(T(x))}\right)T'(x)\,dx = \int g(x)\log\left(\frac{g(x)}{g(T(x))T'(x)}\right).$$

The latter simplifies to

$$\int g(x)\left[-|x|^2/2 + |T(x)|^2/2 - \log T'(x)\right]$$

and we want to show that this expression is at least

$$\frac{1}{2}\int g(x)|x-T(x)|^2.$$

That amounts to showing that

$$\int g(x)\log T'(x) \le \int g(x)\langle x, T(x)-x\rangle.$$

As in the Brunn–Minkowski argument above, for each x, the derivative $T'(x)$ is a positive semi-definite symmetric matrix with eigenvalues $t_1,\dots,t_n$ (say). Therefore $\log T'(x) = \sum \log t_i$ and this is at most $\sum(t_i-1)$ which is the trace of the matrix $T'-I$. This in turn is the divergence (at x) of the map

$$x \mapsto T(x) - x.$$

Using integration by parts and the fact that $\nabla g(x) = -xg(x)$ we can conclude as follows:

$$\begin{aligned}\int g(x)\log T'(x) &\le \int g(x)\mathrm{div}(T(x)-x)\\ &= -\int\langle\nabla g(x), T(x)-x\rangle\\ &= \int g(x)\langle x, T(x)-x\rangle.\end{aligned}$$

□

The original argument of Talagrand, like that of Marton, used induction on dimension, much as in the proof of the Brunn–Minkowski inequality found by Knothe.

To see how the Marton–Talagrand inequality provides a deviation estimate, consider a measurable set B and form a density by restricting the Gaussian density g to B and dividing by $\gamma(B)$: call it f_B. So

$$f_B(x) = \mathbf{1}_B(x)g(x)/\gamma(B).$$

Then the relative entropy of f_B is just

$$\log\left(\frac{1}{\gamma(B)}\right).$$

So the cost of transporting γ to f_B is at most $-2\log\gamma(B)$. Now suppose that A is a set of fairly large Gaussian measure (1/2 perhaps) and that all points of B are at least distance ϵ from A. Then in transporting γ to f_B, we must transport all the measure in A, at least a distance ϵ. So the cost is at least $\gamma(A)\epsilon^2$.

Hence

$$\gamma(A)\epsilon^2 \leq -2\log\gamma(B)$$

which implies that

$$\gamma(B) \leq e^{-\gamma(A)\epsilon^2/2}.$$

Thus the ϵ-neighbourhood of a set of fairly large measure has Gaussian measure, close to 1.

References

[ADM] Alesker, S., Dar, S., Milman, V.: A remarkable measure-preserving transformation between two convex bodies in $\mathbf{R}^n$. Geom. Dedicata, **74**, 201–212 (1999)

[A] Ambrosio, L.: Lecture notes on optimal transport problems. CIME Springer Lecture Notes, to appear

[AKP] Ambrosio, L., Kirchheim, B., Pratelli, A.: Existence of optimal transport maps for crystalline norms. Preprint

[B] Barthe, F.: Inégalités de Brascamp-Lieb et convexité. C.R. Acad. Sci. Paris, **324**, 885–888 (1997)

[Blo] Blower, G.: The Gaussian isoperimetric inequality and transportation. Positivity, **7**, 203–224 (2003)

[BL] Brascamp, H.J., Lieb, E.H.: Best constants in Young's inequality, its converse, and its generalization to more than three functions. Advances in Math., **20**, 151–173 (1976)

[Br] Brenier, J.: Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math., **44**, 375–417 (1991)

[C1] Cafarelli, L.: Boundary regularity of maps with a convex potential. Comm. Pure Appl. Math., **45**, 1141-1151 (1992)

[C2] Cafarelli, L., Regularity of mappings with a convex potential. J.A.M.S., **5**, 99–104 (1992)

[CFM] Cafarelli, L., Feldman, M., McCann, R.J.: Constructing optimal maps for Monge's transport problem as a limit of strictly convex costs. J.A.M.S., **15**, 1–26 (2002)

[CNV] Cordero-Erausquin, D., Nazaret B., Villani, C.: A mass transportation approach to sharp Sobolev and Gagliardo–Nirenberg inequalities. Advances in Math., to appear

[K] Knothe, H.: Contributions to the theory of convex bodies. Michigan Math. J., **4**, 39–52 (1957)

[L] Lieb, E.H.: Gaussian kernels have only Gaussian maximizers. Invent. Math., **102**, 179–208 (1990)

[M] Marton, K.: Bounding $\overline{d}$-distance by informational divergence: A method to prove measure concentration. The Annals of Probability, **24**, 857–866 (1996)

[Mc1] McCann, R.J.: Existence and uniqueness of measure preserving maps. Duke Math. J., **80**, 309–323 (1995)

[Mc2] McCann, R.J.: A convexity principle for interacting gases. Adv. Math., **228**, 153–179 (1997)

[OV] Otto, F., Villani, C.: Generalisation of an inequality by Talagrand and links with the logarithmic Sobolev inequality. J. Funct. Anal., **173**, 361–400 (2000)

[S] Sudakov, V.N.: Geometric problems in the theory of infinite-dimensional distributions. Proc. Steklov Inst. Math., **141**, 1–178 (1979)

[T] Talagrand, M.: Transportation cost for Gaussian and other product measures. Geometric and Functional Analysis, **6**, 587–599 (1996)

[TW] Trudinger, N.S., Wang, X.J.: On the Monge mass transfer problem. Calc. Var. PDE, **13**, 19–31 (2001)

[V] Villani, C.: Topics in optimal transportation. Graduate Studies in Mathematics, **58**, Amer. Math. Soc. (2003)