Applied Mathematical Sciences, Vol. 3, 2009, no. 22, 1071 - 1080

A Gradient Descent Solution to the Monge-Kantorovich Problem

Rick Chartrand [1] **and Brendt Wohlberg**

Los Alamos National Laboratory, Theoretical Division
T-5, MS B284, Los Alamos, NM 87545, USA

Kevin R. Vixie

Washington State University, Department of Mathematics
PO Box 643113, 203 Neill Hall, Pullman, WA 99164-3113, USA

Erik M. Bollt

Clarkson University
Department of Mathematics & Computer Science
Potsdam, NY 13699-5815, USA

Abstract

We present a new, simple, and elegant algorithm for computing the optimal mapping for the Monge-Kantorovich problem with quadratic cost. The method arises from a reformulation of the dual problem into an unconstrained minimization of a convex, continuous functional, for which the derivative can be explicitly found. The Monge-Kantorovich problem has applications in many fields; examples from image warping and medical imaging are shown.

Mathematics Subject Classification: 49D10

Keywords: Monge-Kantorovich problem, optimal transport, image warping.

1 Introduction

The original problem, posed by G. Monge [11] in 1781, was to determine the optimal way to move a pile of dirt to a hole of the same volume. Here "optimal"

[1] rickc@lanl.gov

means the total distance that the dirt is moved, one infinitesimal unit of volume at a time, should be minimal.

A modern and suitably generalized version is the following: let μ_1 and μ_2 be compactly supported, absolutely continuous measures on $\mathbb{R}^n$, with supports K_1 and K_2 and densities f_1 and f_2. Assume that the measures have the same total mass. Call a measurable function $s : K_1 \to K_2$ feasible if s is injective off a set of measure zero, and pushes μ_1 forward to μ_2 in the sense that $\mu_1 \circ s^{-1} = \mu_2$. We wish to find the feasible s that minimizes the total-cost functional

$$I(s) = \int_{K_1} c(x, s(x))\, d\mu_1(x), \tag{1}$$

where $c \in C(K_1 \times K_2)$ is a nonnegative function, thought of as measuring the cost of moving a unit of mass from a point in K_1 to a point in K_2.

In addition to being mathematically interesting, this problem has applications in many fields, a few of which are economics, meteorology, astrophysics, and image processing (see [12], [3], and [1] for discussion and references).

The purpose of this paper is to present a simple algorithm for the numerical computation of the optimal mapping s. Other methods in the literature include linear programming [3], computational fluid mechanics [1], and minimizing flows [6]. Linear programming is simple but inefficient; the other two methods are much more complex in their justification and implementation. Once the mapping s has been computed, one can obtain from (1) a measure of the distance between the measures μ_1 and μ_2, known as the Wasserstein distance. However, the mapping s contains much more information about the relationship between the two measures, and any geometric properties of s will likely be of great relevance to the particular application. See Section 4 for a simple example.

The above formulation of the problem assumes that *all* the dirt at a point $x \in K_1$ must be moved to the *same* point $s(x) \in K_2$. This restriction was relaxed by Kantorovich [7], replacing the mapping s with a measure $\pi \in M(K_1 \times K_2)$ that specifies the joint distribution of dirt-hole correspondences. The measure π is called feasible if it has μ_1 and μ_2 as marginal distributions; that is, if

$$\pi(\cdot \times K_2) = \mu_1 \quad \text{and} \quad \pi(K_1 \times \cdot) = \mu_2. \tag{2}$$

The relaxed problem is to find the feasible π that minimizes

$$J(\pi) = \int_{K_1 \times K_2} c\, d\pi. \tag{3}$$

Gangbo and McCann [5] show that if c is strictly convex, the relaxed problem and the original problem have the same, unique solution: one has $\min I(s) = \min J(\pi)$, both functionals have unique minimizers, and the minimizers are related by $\pi(E) = \mu_1\{x \in K_1 : (x, s(x)) \in E\}$ for all $E \subset K_1 \times K_2$.

2 Duality

Kantorovich also formulated a dual problem [8]: maximize

$$K(u,v) = \int_{K_1} u\, d\mu_1 + \int_{K_2} v\, d\mu_2 \tag{4}$$

among $u \in C(K_1), v \in C(K_2)$ satisfying

$$u(x) + v(y) \le c(x,y) \text{ for all } x \in K_1, y \in K_2. \tag{5}$$

(Note that since $M(K_1 \times K_2) = C(K_1 \times K_2)^*$, this should really be called a predual problem.) It is a dual problem in the sense that $\sup K(u,v) = \min J(\pi)$.

For the rest of this paper, we specialize to the case of the quadratic cost $c(x,y) = \frac{1}{2}|x-y|^2$, a strictly convex function. Ideas from convex analysis can be brought into play by substituting $u(x) = \frac{1}{2}|x|^2 - \varphi(x), v(y) = \frac{1}{2}|y|^2 - \psi(y)$ into (4). The resulting problem is to minimize

$$L(\varphi,\psi) = \int_{K_1} \varphi\, d\mu_1 + \int_{K_2} \psi\, d\mu_2 \tag{6}$$

among $\varphi \in C(K_1), \psi \in C(K_2)$ satisfying

$$\varphi(x) + \psi(y) \ge x \cdot y \text{ for all } x \in K_1, y \in K_2. \tag{7}$$

The value of this substitution is the following result, due to Knott and Smith [9] and Brenier [2].

Proposition 2.1. *The functional L has a unique minimizing pair (φ,ψ) of functions, which are convex conjugates:*

$$\psi(y) = \varphi^*(y) := \max_{x \in K_1}(x \cdot y - \varphi(x)) \tag{8}$$

and

$$\varphi(x) = \psi^*(x) := \max_{y \in K_2}(y \cdot x - \psi(y)). \tag{9}$$

Furthermore, $s = \nabla\varphi$ solves the Monge-Kantorovich problem (1).

The mapping $\varphi \mapsto \varphi^*$ defined by (8) is a variant of the Legendre-Fenchel transform, the difference being that the Legendre-Fenchel transform takes extended real-valued functions on a Banach space X to functions on the dual space X^*. The proposition can be phrased in terms of the true Legendre-Fenchel transform for $\mathbb{R}^n$ by defining $\varphi \equiv \infty$ outside K_1 and similarly for ψ, after which one obtains $\psi|_{K_2} = \varphi^*|_{K_2}$ and $\varphi|_{K_1} = \psi^*|_{K_1}$. By this means, one can deduce the following from the corresponding result (see [13, Proposition 11.3]) for the Legendre-Fenchel transform on $\mathbb{R}^n$.

Lemma 2.2. *Let φ, ψ be convex conjugates in the sense of* (8) *and* (9)*. Then*

$$\partial\varphi(x) = \operatorname*{argmax}_{y\in K_2}\big(y\cdot x - \psi(y)\big) \tag{10}$$

for all $x \in K_1$, and for all $y \in K_2$

$$\partial\psi(y) = \operatorname*{argmax}_{x\in K_1}\big(x\cdot y - \varphi(x)\big). \tag{11}$$

In particular, where φ (resp. ψ) is differentiable, there is a unique maximizer for the right side of (10) *(resp.* (11)*).*

Remark 2.3. *It is easily seen from the definition* (8) *that for any function φ on K_1, φ^* (and hence φ^{**}) is convex and Lipschitz. Hence for $\varphi = \varphi^{**}$ to be true requires that φ be convex and Lipschitz. Unlike the case of the Legendre-Fenchel transform on $\mathbb{R}^n$, however, this is not sufficient, as φ^{**} depends on the choice of K_2. It is true, however, that $\varphi^* = \varphi^{***}$ for any function φ, so that φ^* and φ^{**} will be convex conjugates.*

3 A gradient descent iteration

We can now state the main result of the paper: the Monge-Kantorovich problem can be solved by an *unconstrained* problem, for which the derivative can be explicitly found.

Theorem 3.1. *Let $f_1 \in L^1(K_1)$, $f_2 \in L^1(K_2)$. Define M on $C(K_1)$ by*

$$M(\varphi) = \int_{K_1}\varphi f_1 + \int_{K_2}\varphi^* f_2. \tag{12}$$

The functional M is convex, Lipschitz, and Hadamard differentiable. In particular,

$$M'(\varphi) = f_1 - \big(f_2\circ\nabla\varphi^{**}\big)\det\big(D^2\varphi^{**}\big), \tag{13}$$

where the matrix-valued function $D^2\varphi^{**}$ is defined in the Aleksandrov sense. Furthermore, M has a unique, convex minimizer φ, for which $s = \nabla\varphi$ is the solution to the Monge-Kantorovich problem (1)*.*

Remark 3.2. *If φ is such that $\varphi = \psi^*$ for some ψ, then*

$$M'(\varphi) = f_1 - \big(f_2\circ\nabla\varphi\big)\det\big(D^2\varphi\big). \tag{14}$$

Remark 3.3. *The theorem suggests that a potential for the optimal Monge-Kantorovich mapping can be computed by a gradient descent iteration of the form*

$$\varphi_{n+1} = \varphi_n - \alpha_n M'(\varphi_n), \tag{15}$$

where α_n is a stepsize parameter. However, in general $M'(\varphi_n)$ may fail to be continuous. In practice, with discontinuous f_1 and f_2 we find the iteration (15) *to produce a reasonable approximation of the optimal mapping before numerical instabilities occur. A method to improve the performance of the algorithm is described in Section 4.*

Proof. That M has a unique, convex minimizer which is a potential for the solution of the Monge-Kantorovich problem follows immediately from Proposition 2.1.

To show the convexity of M, since the first term of (12) depends linearly on φ, it suffices to show the pointwise convexity of $\varphi \mapsto \varphi^*(y)$:

$$\begin{aligned}\big(t\varphi_1 + (1-t)\varphi_2\big)^*(y) &= \max_{x\in K_1}\big(x\cdot y - t\varphi_1(x) - (1-t)\varphi_2(x)\big)\\ &\le \max_{x\in K_1} t\big(x\cdot y - \varphi_1(x)\big) + \max_{x\in K_1}(1-t)\big(x\cdot y - \varphi_2(x)\big)\\ &= t\varphi_1^*(y) + (1-t)\varphi_2^*(y).\end{aligned} \tag{16}$$

The Lipschitz continuity of M is an immediate consequence of the contractive property of the Legendre-Fenchel transform, namely that $\|\varphi_1^* - \varphi_2^*\|_\infty \le \|\varphi_1 - \varphi_2\|_\infty$. This property is well-known and the proof, a simple consequence of the definitions, is omitted.

The heart of the theorem, and the key to its usefulness, is the differentiability of M. We begin by computing the one-sided directional derivative of M at $\varphi \in C(K_1)$ in the direction of $v \in C(K_1)$. (A similar computation can be found in a paper [4] by W. Gangbo, but in a different context.)

$$D_v M(\varphi) = \lim_{t\to 0^+}\frac{M(\varphi+tv) - M(\varphi)}{t} = \int_{K_1} v f_1 + \int_{K_2}\lim_{t\to 0^+}\frac{(\varphi+tv)^* - \varphi^*}{t} f_2. \tag{17}$$

Since φ^* is a convex function, φ^* is differentiable almost everywhere. Fix $y \in K_2$ such that $\nabla\varphi^*(y) = x_0$ exists. Then by Lemma 2.2, x_0 is the unique maximizer of $y\cdot x - \varphi(x)$, the quantity whose maximum is $\varphi^*(y)$. Similarly, for $t > 0$ choose $x_t \in \partial(\varphi+tv)^*(y) = \operatorname{argmax}_{x\in K_1}(x\cdot y - (\varphi+tv)(x))$. Then

$$(\varphi+tv)^*(y) - \varphi^*(y) = x_t\cdot y - \varphi(x_t) - tv(x_t) - x_0\cdot y - \varphi(x_0). \tag{18}$$

Replacing x_0 with x_t in (18) results in a larger quantity, while replacing x_t with x_0 results in a smaller quantity. Rearranging gives

$$0 \leq \frac{(\varphi + tv)^*(y) - \varphi^*(y)}{t} + v(x_0) \leq v(x_0) - v(x_t). \tag{19}$$

Since $tv(x_t)$ converges uniformly to 0, any convergent subsequence of the family (x_t) will converge to a maximizer of $x \cdot y - \varphi(x)$. Since x_0 is the unique such maximizer, it follows that $x_t \to x_0$, hence $v(x_0) - v(x_t) \to 0$. Therefore

$$D_v M(\varphi) = \int_{K_1} v f_1 - \int_{K_2} (v \circ \nabla \varphi^*) f_2. \tag{20}$$

The Hadamard differentiability of M at φ is equivalent (since M is Lipschitz) to the existence of a measure $M'(\varphi) = \sigma \in M(K_1) = C(K_1)^*$ such that $D_v M(\varphi) = \int_{K_1} v \, d\sigma$ for arbitrary $v \in C(K_1)$; when σ is absolutely continous we identify $M'(\varphi)$ with the density function. We obtain this by the change of variables $y = \nabla \varphi^{**}(x)$ in (20), first obtaining

$$D_v M(\varphi) = \int_{K_1} v f_1 - \int_{(\nabla \varphi^{**})^{-1}(K_2)} \big(v \circ \nabla \varphi^* \circ \nabla \varphi^{**}\big)\big(f_2 \circ \nabla \varphi^{**}\big) \det\big(D^2 \varphi^{**}\big). \tag{21}$$

This change of variables is justified by work of McCann [10]: if we denote $\big(v \circ \nabla \varphi^*\big) f_2$ by g, then $\nabla \varphi^{**}$ pushes $\big(g \circ \nabla \varphi^{**}\big) \det D^2 \varphi^{**}$ forward to g, where the Aleksandrov derivative $D^2 \varphi^{**}$ is the absolutely continuous part of the distributionally-defined Hessian of φ^{**}. The purpose of this change of variables is to employ the cancellation property of gradients of convex-conjugate functions. Namely,

$$\nabla \varphi^* \circ \nabla \varphi^{**}(x) = x \tag{22}$$

when the left side exists, a consequence of the uniqueness of the maximizers in (10) and (11). Although the convex functions φ^* and φ^{**} are differentiable almost everywhere, it may be that $\nabla \varphi^{**}$ maps a set of positive measure into the set where $\nabla \varphi^*$ fails to exist. On the other hand, the Aleksandrov derivative will be singular on such a set (see [10]), and so the cancellation property (22) holds on the support of the second integrand in (21).

The two integrals in (21) can be combined, as $\big(\nabla \varphi^{**}\big)^{-1}(K_2) = K_1$. Indeed, at any $x \in K_1$ such that $\nabla \varphi^{**}(x) = y_0$ exists, y_0 is the unique maximizer of $x \cdot y - \varphi^*(y)$. In particular, $y_0 \in K_2$. Combining this with the cancellation property (22), we obtain

$$D_v M(\varphi) = \int_{K_1} \big(f_1 - \big(f_2 \circ \nabla \varphi^{**}\big) \det\big(D^2 \varphi^{**}\big)\big) v. \tag{23}$$

This establishes the existence of the Hadamard derivative (13).

□

4 Examples

In this section, we present two examples in which we use the result of Theorem 3.1 to compute optimal mappings for warping images. The first demonstrates the effectiveness of the algorithm using a standard pair of images from the image processing literature. The second is a simple example from medical imaging of how geometric properties of the optimal mapping between two images contain information about the relationship between the two images.

In the context of image warping, f_1 and f_2 are discretely approximated by the intensity values of the pixels in two greyscale images. In practice, we find it works as well to replace the iteration (15) with

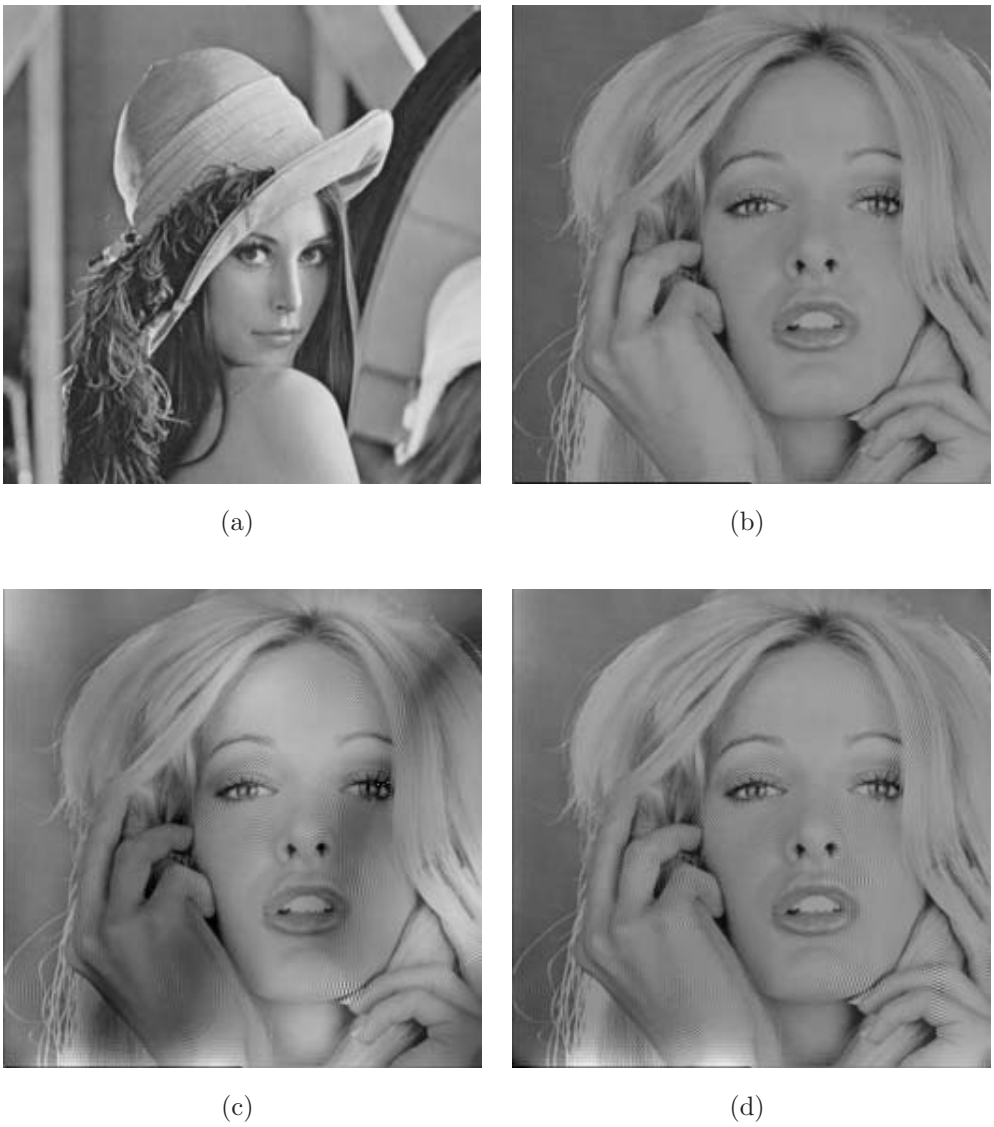
$$\varphi_{n+1} = \varphi_n - \alpha_n \Big(f_1 - \big(f_2 \circ \varphi_n\big) \det\big(D^2 \varphi_n\big) \Big), \tag{24}$$

which amounts to approximating φ_n^{**} with φ_n.

With natural images, we find the iteration (24) can produce good-quality warpings. An example with two 256×256-pixel images is shown in Figure 1. A Lax-type numerical scheme was used. Values of φ were computed at pixel vertices. Derivatives of φ were computed at pixel centers using centered-differencing. The second term of (24) was computed at the pixel centers, then φ_{n+1} was updated from φ_n at each vertex by averaging over surrounding centers. The resulting warp of the Lena image is shown in Figure 1(c), where 190 iterations with stepsize parameter 1 were used, starting with the potential of the identity mapping as the initial function φ_0. Numerical artifacts are just beginning to appear where the residue of the mirror of the Lena image meets Tiffany's eyelashes. These artifacts worsen upon further iteration, precluding iterating further to remove the mirror residue and dark remnants of Lena's hair.

To improve the quality of the warp, we employ a multiresolution approach. We begin with smoothed versions of the two images, obtained by convolution with a Gaussian kernel. We run the algorithm to obtain a warping potential. We then repeat the algorithm with the images having been smoothed to a lesser degree, using the final potential from the previous step as the initial function. We repeat this procedure, then at some point use the warping potential obtained as the initial function for the unsmoothed images. Figure 1(d) shows the result of using just one step of this iterated smoothing procedure, using a Gaussian kernel of size 16 and width 4. The resulting warp of the Lena image is almost identical to the Tiffany image. Compare with the corresponding images in [6].

An example is presented in Figure 2 of a way that geometric information contained in the optimal Monge-Kantorovich mapping can be used to infer something about the relationship between two images. In Figures 2(a) and 2(b)

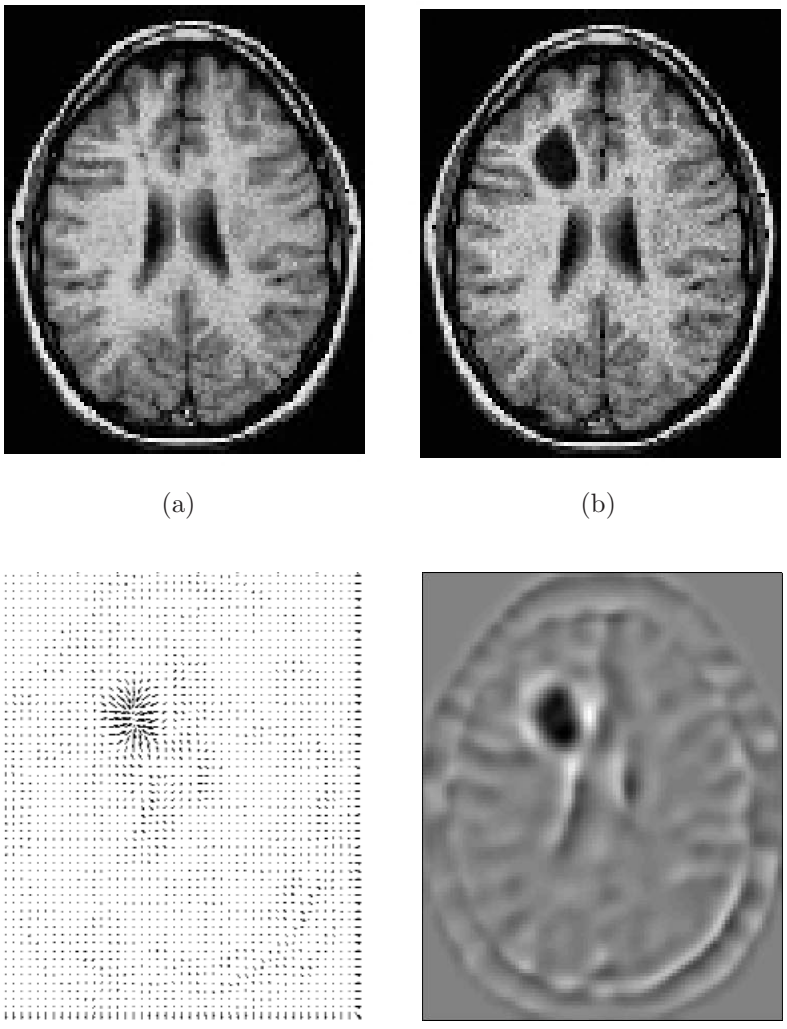


(a) (b)

(c) (d)

Figure 1: (a) Lena image. (b) Tiffany image. (c) Lena to Tiffany warp, without smoothing. (d) Lena to Tiffany warp, using smoothed images first to compute the initial function.

are images of the same brain slice, before and after a tumor has developed. In Figure 2(c) is a plot of the vector field $s(x) - x$, where s is the Monge-Kantorovich mapping. One can see that most of the deformation of the domain is in the tumor region. This region can be identified using the divergence of s, displayed in Figure 2(d). The dark color corresponds to a negative divergence, as the mapping s compresses the domain to create the dark tumor region. The nearby light areas show where surrounding tissues have been compressed by the growth of the tumor.



(a) (b)

(c) (d)

Figure 2: (a) Healthy brain. (b) Same brain with tumor. (c) Vector-field plot of mapping $s(x) - x$. (d) Divergence of optimal mapping.

References

[1] J.-D. Benamou and Y. Brenier, A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem, Numer. Math., **84** (2000), 375–393.

[2] Y. Brenier, Décomposition polaire at réarrangement monotone des champs de vecteurs, C. R. Acad. Sci. Paris Sér. I Math., **305** (1987), 805–808.

[3] L. C. Evans, Partial differential equations and Monge-Kantorovich mass transfer, in Current developments in mathematics, 1997 (Cambridge, MA), Int. Press, Boston, MA, 1999, 65–126.

[4] W. Gangbo, An elementary proof of the polar factorization of vector-valued functions, Arch. Rational Mech. Anal., **128** (1994), 381–399.

[5] W. Gangbo and R. J. McCann, The geometry of optimal transportation, Acta Math., **177** (1996), 113–161.

[6] S. Haker and A. Tannenbaum, On the Monge-Kantorovich problem and image warping, in Mathematical methods in computer vision, vol. 133 of IMA Vol. Math. Appl., Springer, New York, 2003, 65–85.

[7] L. V. Kantorovich, On the translocation of masses, Dokl. Akad. Nauk SSSR, **37** (1942), 227–229.

[8] ——, On a problem of Monge, Uspekhi Mat. Nauk, **3** (1948), 225–226.

[9] M. Knott and C. S. Smith, On the optimal mapping of distributions, J. Optim. Theory Appl., **43** (1984), 39–49.

[10] R. J. McCann, A convexity principle for interacting gases, Advances in Math., **128** (1997), 153–179.

[11] G. Monge, Mémoire sur la théorie des déblais et des remblais, Histoire de l'Académie Royale des Sciences de Paris, (1781), 666–704.

[12] S. T. Rachev and L. Rüschendorf, Mass transportation problems. Vol. II: Applications, Probability and its Applications (New York), Springer-Verlag, New York, 1998.

[13] R. T. Rockafellar and R. J.-B. Wets, Variational analysis, vol. 317 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Springer-Verlag, Berlin, 1998.

Received: November, 2008