Data-driven Optimal Transport

Esteban G. Tabak — *Courant Institute*
Giulio Trigila — *TU München*

August 27, 2014

Abstract

The problem of optimal transport between two distributions $\rho(x)$ and $\mu(y)$ is extended to situations where the distributions are only known through a finite number of samples $\{x_i\}$ and $\{y_j\}$. A weak formulation is proposed, based on the dual of Kantorovitch's, with two main modifications: replacing the expected values in the objective function by their empirical means over the $\{x_i\}$ and $\{y_j\}$, and restricting the dual variables $u(x)$ and $v(y)$ to a suitable set of test functions adapted to the local availability of sample points. A procedure is proposed and tested for the numerical solution of this problem, based on a fluid-like flow in phase space, where the sample points play the role of active Lagrangian markers.

1 Introduction

This article is concerned with finding a map $y(x)$ that transports a probability density $\rho(x)$ into another one $\mu(y)$ while minimizing the integral of a transportation cost $c(x, y)$. This problem was first formulated by Monge in 1781 [22], extended by Kantorovich in 1942 [18] and studied in various settings since (see [27] and more specific references provided below.) We focus here on the frequently occurring scenario where in lieu of the distributions $\rho(x)$ and $\mu(y)$, one is provided with independent samples $\{x_i, i \in [1, m]\}$ and $\{y_j, j \in [1, n]\}$ from each. Examples of applications are:

- **Aggregation of data gathered in various laboratories into a single database:** Since some methodological elements necessarily vary among laboratories, each dataset has a different underlying distribution ρ_j. To homogenize the data, one can normalize the features mapping each ρ_j into a single target μ, chosen for instance as the distribution corresponding to the lab with the most data.

- **Effect of a medical treatment:** Clinical variables have been measured in two populations, where one has received the treatment and the other a placebo. The effect of the treatment on those clinical variables can be

conceptualized as a map that transforms the distribution underlying one dataset into the other.

- **Flow inference from tracers:** In order to determine fluid flows, such as ocean currents, atmospheric winds and blood flow through the cardiovascular system, artificial or natural tracers are located at two different times. The flow then maps one underlying density into the other.

Broader applications that involve generalizations of this data-driven transport problem include:

- **Density estimation:** Here the goal is to determine $\rho(x)$ from the samples $\{x_i\}$. The target $\mu(y)$ is a known density proposed by the modeler, typically an isotropic Gaussian. Once the map $y(x)$ is determined, $\rho(x)$ is estimated through
$$\rho(x) = J^y(x)\mu(y(x)),$$
where $J^y(x)$ is the Jacobian determinant of the map.

- **Regression:** A function $y(x)$ relating the features x and labels y is sought –the dimensions of x and y here are typically different– from samples drawn independently from the underlying distributions $\rho(x)$ and $\mu(y)$, and fewer or none jointly drawn samples (x, y). The cost function $c(x, y)$ must be chosen to favor those maps that have desired properties, such as mapping the known pairs (x, y) correctly or being closest to a conjectured prior.

- **Importance Sampling:** In order to estimate integrals of the form
$$I = \int_\Omega f(y)\ \gamma(y)\ dy,$$
importance sampling rewrites it as
$$I = \int_\Omega \frac{f(y)\ \gamma(y)}{\mu(y)}\mu(y)\ dy.$$
Here $\mu(x)$ is a probability distribution comparatively easy to sample and evaluate and such that $\frac{f(x)\ \gamma(x)}{\mu(x)}$ has small variance, so that the Monte Carlo estimate
$$I \approx \frac{1}{m}\sum_{i=1}^{m} \frac{f(y_i)\ \gamma(y_i)}{\mu(y_i)}$$
is accurate, where the $\{y_i\}$ are independent samples drawn from μ. If an ideal target distribution μ is known but hard to sample, one can produce samples from a known nearby distribution by sampling another distribution $\rho(x)$ –easy to sample and evaluate– and mapping it onto $\mu(y)$ (this idea was first developed in [23] using maximum likelihood instead of optimal transport.)

This article proposes both a formulation of the data-driven transport problem and a methodology to solve it, based on a fluid-like flow in phase space. We study here the case in which the data consists only of samples and the number of components of x and y are the same, leaving the extensions necessary for the broader applications mentioned above to further work.

Fist we summarize briefly the formulations of Monge and Kantorovich when the distributions $\rho(x)$ and $\mu(y)$ are known.

1.1 Monge formulation

The optimal transport problem posed by Monge is to find a plan to transport material from one location to another that minimizes the total transportation cost. By normalizing the total amount of material to transport to one, the initial and final distributions can be modeled as two probability densities, thus stating the transport problem in the following terms:

Given two probability densities $\rho(x)$ and $\mu(y)$ with $x, y \in \mathbb{R}^n$, find the map $y = y(x) : \mathbb{R}^n \to \mathbb{R}^n$ transporting the density $\rho(x)$ to $\mu(y)$ that minimizes the cost functional

$$M(y) = \int_{\mathbb{R}^n} c(x, y(x))\rho(x)dx. \tag{1}$$

In order to transport $\rho(x)$ to $\mu(y)$ the map is required to satisfy the relation

$$\int_{y^{-1}(\Omega)} \rho(x)dx = \int_{\Omega} \mu(y)dy \tag{2}$$

for all measurable sets Ω. If $y(x)$ is smooth and one to one, this is equivalent to the point-wise relation

$$\rho(x) = J^y(x)\mu(y(x)), \tag{3}$$

where $J^y(x)$ is the Jacobian of the map: $J^y(x) = \det(\nabla y(x))$. A function $y(x)$ satisfying (3) and minimizing $M(y)$ is called an *optimal map*.

1.2 Kantorovich formulation

Monge's formulation of the transport problem requires a one-to-one map between the points at the origin and destination. Kantorovich proposed a relaxation where the mass reaching each point y may come from various points x and, conversely, the mass from each point x may be split into various destinations y. In this formulation, the way in which ρ must be rearranged to obtain μ is described by a joint probability density $\pi(x, y) \geqslant 0$ satisfying

$$\rho(x) = \int \pi(x, y)dy, \qquad \mu(y) = \int \pi(x, y)dx \tag{4}$$

and the problem reduces to minimizing the functional

$$K(\pi) = \int c(x, y)\pi(x, y)dxdy. \tag{5}$$

A minimizer of (5) subject to (4) is called an *optimal plan*. It can be shown that under relatively mild hypotheses, such as $\rho(x)$ and $\mu(y)$ having finite second moments and $c(x, y)$ being a strictly convex function of $y - x$ ([20]), the relaxed problem and the original Monge problem have the same unique solution, in the sense that

$$\min M(y) = \min K(\pi) \tag{6}$$

and the minimizers satisfy

$$\pi(S) = \rho(\{x : (x, y(x)) \in S\}). \tag{7}$$

Notice that unlike Monge's formulation, in which the cost function and the constraints are nonlinear in the unknown $y(x)$, in Kantorovich's both are linear in $\pi(x, y)$, yielding a continuous version of the *assignment problem*, a category of linear programming problems with wide applications in economics. Then it has a dual formulation, which adopts the form:

$$D(u, v) = \max_{u,v} \int u(x)\rho(x)dx + \int v(y)\mu(y)dy \tag{8}$$

over all integrable functions u and v satisfying

$$u(x) + v(y) \leq c(x, y). \tag{9}$$

To illustrate the relation between the primal and dual variables, consider the classical case in which the cost function adopts the particular form

$$c(x, y) = \frac{\|y - x\|^2}{2}. \tag{10}$$

Redefining $u(x) \to \frac{\|x\|^2}{2} - u(x)$ and $v(y) \to \frac{\|y\|^2}{2} - v(y)$ turns the dual problem into:

$$D(u, v) = \min_{u,v} \int u(x)\rho(x)dx + \int v(y)\mu(y)dy \tag{11}$$

where u and v are continuous functions satisfying

$$u(x) + v(y) \geq x \cdot y, \tag{12}$$

and the following proposition applies [3]:

The functional $D(u, v)$ admits a unique minimizer $(\bar{u}, \bar{v})$; $\bar{u}$ and $\bar{v}$ are convex conjugates:

$$\begin{cases} \bar{u}(x) = \max_y (x \cdot y - \bar{v}(y)) \equiv v^*(x) \\ \bar{v}(y) = \max_x (x \cdot y - \bar{u}(x)) \equiv u^*(y) \end{cases}$$

and the optimal plan for Monge's problem is given by

$$y(x) = \nabla \bar{u}(x).$$

The function $u^*(y) \equiv \max_x(x \cdot y - \bar{u}(x))$ is the Legendre transform of $\bar{u}$. Since $\bar{u}$ is obtained as the envelope of its supporting planes, it is a convex function. It follows from (3) that the optimal map between $\rho(x)$ and $\mu(y)$ is given by the gradient ∇u of a convex function $u : \mathbb{R}^n \to \mathbb{R}$ satisfying the *Monge-Ampere* equation ([26], [1], [8],[27]):

$$\rho(x) = \mu(\nabla u(x)) det(D^2 u)(x). \tag{13}$$

1.3 A data-based formulation

Optimal transport has a broad range of applications in mathematics and beyond, in fields as diverse as image processing and economics. As mentioned above, many more applications arise if one extends the problem to situations in which the two densities $\rho(x)$ and $\mu(y)$ are not known pointwise but through samples: m points $\{x_i\}$ drawn from $\rho(x)$ and n points $\{y_j\}$ drawn from $\mu(y)$.

In order to develop a formulation of the optimal transport problem suited to such data-driven generalization, we notice that in Kantorovich dual formulation (8) the probability densities $\rho(x)$ and $\mu(y)$ appear only in the objective function D, and in a form that can be interpreted in terms of expected values:

$$D(u, v) = E[u(x)]_{\rho(x)} + E[v(y)]_{\mu(y)}. \tag{14}$$

Thus, if $\rho(x)$ and $\mu(y)$ are known only through samples, it is natural to pose the problem in terms of sample averages:

Maximize

$$D(u, v) = \frac{1}{m} \sum_{i=1}^{m} u(x_i) + \frac{1}{n} \sum_{j=1}^{n} v(y_j)$$

over functions u and v satisfying

$$u(x) + v(y) \leq c(x, y).$$

In order to complete the statement of this problem, one needs to specify the space of functions to which u and v belong. As we shall see below, various possibilities arise. Moreover, since this data-based formulation is presented in terms of the dual variables u and v, the meaning of its solution and its relation to the original transport problem must be described. These two kinds of considerations are closely interlinked; we address them in section 2.

1.4 Grid-based versus mesh-free computations

Even when addressing the classical transport problem in which the distributions $\rho(x)$ and $\mu(y)$ are known, it may be convenient to base its numerical solution on samples rather than on the distributions themselves: to resolve distributions, one needs to resort to grids, which become prohibitively expensive in high dimensions (To our knowledge, none of the numerical methodologies proposed to date have been tried in dimensions higher than three, and most

have been restricted to dimensions two or even one. By contrast, a procedure based on samples scales well with dimensionality; for instance we report below the results of computations in up to ten dimensions.)

Thus the methodology proposed in this article has two main goals: to pose and solve the data-driven optimal transport problem, and to solve numerically through a Monte Carlo, sample-based approach the classical, continuous problem of Monge and Kantorovich. The numerical procedure is based on a gradient flow in feature-space, whereby the points x "move" toward their targets $y(x)$ continuously in a fictitious time. In the language of fluid dynamics, the sample-points play the role of active Lagrange multipliers: markers moving with the flow while guiding its evolution. The corresponding "forces" result from the sample's appearance in the "potential" given by the objective function $D(u, v)$, whose maximization drives the flow.

1.5 Prior work

Prior work on optimal transport, both analytical and numerical, has considered only the situation where the distributions $\rho(x)$ and $\mu(y)$, continuous or discrete, are known. Solving the discrete version then becomes a standard linear programming problem. Much effort has been devoted to describing when the optimal solution of Kantorovich's relaxation yields in fact a solution to Monge's problem, i.e. a map ([1] [24], [15], [14] [16] [4]). By contrast, a numerical procedure for the discrete case is developed in [6] that seeks the opposite of a map: a joint distribution with large entropy.

While there exists a substantial body of literature on optimal transport from the viewpoint of mathematical analysis, the numerical aspects of the problem have been much less studied. Almost all the algorithms available are based on methods requiring spatial discretization, as in [1], [10], [24], [15], [14] and [4]. In particular, Benamou and Brenier ([1]) develop a fluid-flow based formulation of the optimal transport problem. It assumes that $\rho(x)$ and $\mu(y)$ are known pointwise, and the corresponding flow is resolved on a grid in phase space. Instead our flow is particle-based and hence mesh-free, with the sample-points playing the role of particles or active Lagrangian markers.

The literature on mesh-free methods is limited to our knowledge to the algorithm described in [16], where optimal transport is implemented along straight lines enforcing mass conservation of the discretized starting and target densities.

1.6 Plan of the paper

After this introduction, section 2 introduces a data-driven formulation of the optimal transport problem, relaxing Kantorovich's dual formulation in two complementary ways: replacing the expected values in the objective function by empirical means over the available samples and restricting the dual variables to a space of test functions designed not to over-resolve the probability densities underlying the samples. Section 3 develops a fluid-like methodology to solve this problem, introducing a gradient flow in feature-space. Section 4 describes a

mesh-free implementation of this gradient flow and an additional step designed to enforce the optimality of the map. Section 5 extends the methodology to convex costs other than the squared distance. Section 6 illustrates the algorithm with a number of numerical tests. Finally section 7 summarizes the article and discusses future work.

2 Discrete and continuous data-based formulations of optimal transport

Motivated by the discussion above, we consider the problem of maximizing

$$D(u,v) = \frac{1}{m}\sum_{i=1}^{m} u(x_i) + \frac{1}{n}\sum_{j=1}^{n} v(y_j) \tag{15}$$

over functions u and v satisfying

$$u(x) + v(y) \leq c(x,y). \tag{16}$$

The objective function involves u and v evaluated only at the sample points, but the constraints apply at all values of x and y. One could consider instead the simpler, purely discrete problem:

Maximize

$$D(u,v) = \frac{1}{m}\sum_{i=1}^{m} u_i + \frac{1}{n}\sum_{j=1}^{n} v_j \tag{17}$$

over vectors u and v satisfying

$$u_i + v_j \leq c_{ij}. \tag{18}$$

This is dual to the *uncapacitated transportation problem*:

Minimize

$$C(\pi) = \sum_{i,j} c_{ij}\pi_{ij} \tag{19}$$

subject to

$$\sum_{j} \pi_{ij} = n \tag{20}$$

$$\sum_{i} \pi_{ij} = m. \tag{21}$$

However, this is not the problem we are interested in: instead of seeking a map $y(x)$ for all values of x, it seeks an assignment π_{ij} between the two discrete sets of points x_i and y_j.

Therefore we return to the fully constrained formulation in (16). Since this is a Monte Carlo relaxation of the dual to the problem of interest, we first find its own dual. In terms of the Lagrangian

$$L(\pi, u, v) = \frac{1}{m}\sum_{i=1}^{m} u(x_i) + \frac{1}{n}\sum_{j=1}^{n} v(y_j) - \int \pi(x,y)\left[u(x) + v(y) - c(x,y)\right] dxdy,$$

the problem in (15, 16) can be formulated as

$$d: \max_{u(x),v(y)} \min_{\pi(x,y)\geqslant 0} L(\pi, u, v), \tag{22}$$

and its dual as

$$p: \min_{\pi(x,y)\geqslant 0} \max_{u(x),v(y)} L(\pi, u, v). \tag{23}$$

(We adopt the reverse notation d and p for primal and dual since these relate to the dual and primal of the original transportation problem). To study p, it is convenient to rewrite the Lagrangian in the more revealing form:

$$\begin{aligned} L(\pi, u, v) \quad &= \quad \int c(x,y)\ \pi(x,y)\ dxdy & (24)\\ &- \int \left[\int \pi(x,y)dy - \frac{1}{m}\sum_{i=1}^{m} \delta(x - x_i)\right] u(x)dx & (25)\\ &- \int \left[\int \pi(x,y)dx - \frac{1}{n}\sum_{j=1}^{n} \delta(y - y_j)\right] v(y)dy. & (26) \end{aligned}$$

Then, if the functions $u(x)$ and $v(y)$ are only constrained to be integrable, the problem p becomes

$$p: \min_{\pi(x,y)\geqslant 0} \int c(x,y)\ \pi(x,y)\ dxdy \tag{27}$$

subject to

$$\int \pi(x,y)dy = \frac{1}{m}\sum_{i=1}^{m} \delta(x - x_i),\ \int \pi(x,y)dx = \frac{1}{n}\sum_{j=1}^{n} \delta(y - y_j), \tag{28}$$

namely the Kantorovich formulation of the (primal) optimal transport problem with discrete densities $\rho(x)$ and $\mu(y)$ as in (28). This is not a surprising result, since the problem d is precisely the dual to this problem. Yet this result poses a dilemma: our motivation was not to consider a discrete transport problem that assigns points from y_j to x_i, but one where the x_i and y_j are samples from hidden, presumably smooth distributions $\rho(x)$ and $\mu(y)$. How can one formulate a problem with this flavor?

A natural answer involves restricting the space of functions F from where u and v can be selected. If the space F is invariant under multiplication by scalars:

$$u \in F, \lambda \in \mathbb{R} \to \lambda u \in F,$$

then the problem p becomes

$$p: \min_{\pi(x,y)\geqslant 0} \int c(x,y)\ \pi(x,y)\ dxdy \tag{29}$$

subject to

$$\int \left[\int \pi(x,y)dy - \frac{1}{m}\sum_{i=1}^{m} \delta(x-x_i) \right] u(x)dx = 0, \tag{30}$$

$$\int \left[\int \pi(x,y)dx - \frac{1}{n}\sum_{j=1}^{n} \delta(y-y_j) \right] v(y)dy = 0 \tag{31}$$

for all $u(x), v(y) \in F$. This weak formulation of the strong constraints in (28) is quite natural: we constrain the marginals $\rho(x)$ and $\mu(y)$ of $\pi(x,y)$ so that the corresponding expected values of all functions in F agree with their averages over the samples x_i and y_j respectively.

As a simple example that is exactly solvable, adopt F as the set of quadratic functions and $c(x,y)$ as the squared distance $c = \frac{1}{2}\|x-y\|^2$. Redefining u and v as above, $u(x) \to \frac{\|x\|^2}{2} - u(x)$ and $v(y) \to \frac{\|y\|^2}{2} - v(y)$, problem d becomes

$$\min_{A,b} \frac{1}{m}\sum_{i=1}^{m} u(x_i) + \frac{1}{n}\sum_{j=1}^{n} v(y_j), \tag{32}$$

$$u(x) = \frac{1}{2}x^T Ax + b^T x, \quad v(y) = u^*(y) = \frac{1}{2}y^T A^{-1} y - b^T A^{-1} y + \frac{1}{2}b^T A^{-1} b,$$

where b is a vector and A a symmetric matrix (here we have set in the objective function the explicit form of the Legendre transform $v(y) = u^*(y)$ for quadratic functions.)

The corresponding optimal map is given by

$$y(x) = \nabla u = Ax + b.$$

A straightforward calculation shows that the A and b minimizing (32) are precisely the ones that make $y(x)$ transform the empirical mean and covariance of x into those of y.

This example is, of course, just a proof of concept. As the calculation above verifies, restricting F to quadratic functions is equivalent to just considering the empirical mean and covariance of the data, so it misses any other information that the data may convey, such as higher moments and detailed structure. More generally, one may propose for F the space of polynomials of degree q or any suitable finite-dimensional linear space. An alternative to the direct specification of F is to leave it unrestricted, adding instead to the objective function a term that penalizes the non-smoothness of u and v.

The choices above appear sensible yet none is completely natural, since the specification of F or the qualification of smoothness that one seeks should depend on the samples $\{x_i\}$, $\{y_j\}$ themselves: one can allow a richer set of functions in areas where there are enough sample points to fully determine them

than in those where the samples are sparse. In other words, one must restrict F, in order not to over-resolve the densities $\rho(x)$ and $\mu(y)$, to smooth functions with scales long compared to the distance between neighboring sample points, a concept analogous to the bandwidth of kernel density estimation [17]. Again, this can be done either through a careful choice of a basis for F or through an appropriate penalization term that considers the underlying densities. In either case, the choice of F must be adaptive (i.e. data dependent.)

Exploring these possibilities systematically opens a wide avenue of research that will be pursued elsewhere. This article presents instead an effective algorithm for finding an optimal plan consistent with such adaptive weak formulation: the algorithm enforces a local bandwidth for F consistent with the underlying densities by building u through the composition of many local maps, where each has an appropriately chosen length scale.

3 A flow-based, primal-dual approach

We propose a methodology that, rather than solving the primal or dual problems in isolation, combines them through a flow in feature-space (see [1] for a different fluid-based formulation of the optimal transport problem.) Even though our methodology applies to a broad class of cost functions $c(x, y)$, we restrict our attention here for concreteness to the classical cost $c = \frac{1}{2}\|x - y\|^2$, and redefine u and v accordingly as above (The extension to a broader class of convex cost functions is discussed briefly in section 5). In this case the map $y(x)$ that solves Monge's problem can be written in terms of the solution to the dual Kantorovich formulation as

$$y(x) = \nabla u(x).$$

We shall consider this map $y(x)$ as the endpoint of a time-dependent flow $z(x, t)$, such that

$$z(x, 0) = x, \quad z(x, \infty) = y(x).$$

Then, rather than minimizing the objective function directly, we compute the map from ρ to μ by means of the following gradient flow:

$$\begin{cases} \dot{z} = -\nabla_z \left[\dfrac{\delta \tilde{D}_t}{\delta u} \right]_{u=\frac{1}{2}\|x\|^2} \\ z(x, 0) = x \end{cases} \tag{33}$$

where

$$\tilde{D}_t = \int u(z)\rho_t(z)dz + \int u^*(y)\mu(y)dy \tag{34}$$

and ρ_t is the evolving probability distribution underlying the points $z(x, t)$ that are obtained through the flow in (33); in particular, $\rho_0(z) = \rho(z)$ (After considering here the regular transportation problem where $\rho(x)$ and $\mu(y)$ are known,

we shall translate these ideas to our case of interest, where only samples from both are given.)

The variational derivative of $\tilde{D}_t$ adopts the form

$$\frac{\delta \tilde{D}_t}{\delta u} = \rho_t - \mu(\nabla u^{**}) det(D^2 u^{**}). \tag{35}$$

An argument justifying (35) goes as follows (see for instance [4] for a complete proof). One can write

$$u^*(y) = \max_x [x \cdot y - u(x)] = X(y) \cdot (y) - u(X(y)),$$

where

$$X(y) = \arg\max [x \cdot y - u(x)] .$$

Under a small perturbation of $u(x)$,

$$u(x) + \epsilon\eta(x),$$

one has

$$X(y) \to X(y) + \epsilon f(y)$$

for some $f(y)$, and hence to leading order in ϵ,

$$\begin{aligned} u^*(y) &\to X(y) \cdot (y) - u(X(y)) + \epsilon \left\{ f(y) \cdot \left[y - \nabla u|_{X(y)} \right] - \eta(X(y)) \right\} \\ &= u^*(y) - \epsilon\eta(X(y)), \end{aligned}$$

since the expression within brackets vanishes due to the definition of $X(y)$. Then (34) yields

$$\tilde{D}_t \to \tilde{D}_t + \epsilon \left[\int \eta(x)\rho_t(x)dx - \int \eta(x)\mu(Y(x))J_Y(x)dx \right]$$

where $Y(x)$ is the inverse of $X(y)$. This is given by $Y(x) = \nabla u(x)$ if $u(x)$ is convex and more generally by $Y(x) = \nabla u^{**}(x)$, thus justifying (35).

Applying (35) at $u = \frac{1}{2}\|x\|^2$ (the potential corresponding to the identity map) yields the simpler expression

$$\frac{\delta \tilde{D}_t}{\delta u}(z) = \rho_t(z) - \mu(z),$$

since $\left(\frac{1}{2}\|x\|^2\right)^{**} = \frac{1}{2}\|x\|^2$ by convexity.

Then (33) becomes

$$\begin{cases} \dot{z} = -\nabla_z \left[\rho_t(z) - \mu(z) \right] \\ z(x, 0) = x. \end{cases} \tag{36}$$

The probability density ρ_t satisfies the continuity equation

$$\frac{\partial \rho_t(z)}{\partial t} + \nabla_z \cdot [\rho_t(z)\dot{z}] = 0, \tag{37}$$

which in view of (36) becomes a closed evolution equation for ρ_t:

$$\frac{\partial \rho_t(z)}{\partial t} - \nabla_z \cdot [\rho_t(z) \nabla_z (\rho_t(z) - \mu(z))] = 0, \tag{38}$$

with initial condition $\rho_0(x) = \rho(x)$.

Introducing the L^2 norm

$$\|\rho_t - \mu\|^2 = \int (\rho_t(z) - \mu(z))^2 dz, \tag{39}$$

we have that

$$\frac{d}{dt} \frac{\|\rho_t - \mu\|^2}{2} = \int \frac{\partial}{\partial t} \frac{(\rho_t - \mu)^2}{2} dz = - \int \rho_t \|\nabla(\rho_t - \mu)\|^2 dz \leq 0.$$

Since $\|\nabla(\rho_t - \mu)\| \equiv 0$ only when $\rho_t = \mu$, this proves that $\rho_t \xrightarrow{t \to \infty} \mu$ in the L^2 norm, so the function $y(x) = z(x, \infty)$ maps $\rho(x)$ into $\mu(y)$.

One may wonder whether this map is the optimal one, for which it would have to be the gradient of a convex potential. Even though $y(x)$ is built as an infinite composition of gradients, the composition of gradients is not in general a gradient. We show below through a simple numerical experiment that indeed $y(x)$ is not necessarily curl free.

A numerical experiment on a grid

We solve equation (38) numerically in a periodic domain, pseudo-spectrally with a second order Adams-Bashforth time-stepping scheme. The initial distribution (plotted in the upper left panel of figure 1) is given by

$$\rho(x_1, x_2) = \frac{\cos(0.5x_1)^2 + \sin(0.5x_2)^2}{4\pi^2}, \tag{40}$$

while the target distribution (lower right panel of figure (1)) $\mu(x_1, x_2)$ is obtained from ρ through a translation by a distance π along the x_2 coordinate. Even though the starting and the target distributions are related by a simple translation, this is not necessarily the optimal map from ρ to μ: periodic boundary conditions often have the effect of splitting the initial mass into two or more domains separated by branch-like cuts, which reach the target through distinct paths, favoring rearrangements less costly than a rigid translation.

Figure 1 displays the time evolution of the density ρ_t satisfying equation (38). Some observations are in order:

- As expected, the starting density ρ is mapped into the target μ.
- The map is not a rigid translation, in agreement with analogous numerical experiments (see for instance [1] and [15]) investigating optimal transport in periodic domains between probability distributions that are displacements of one another.

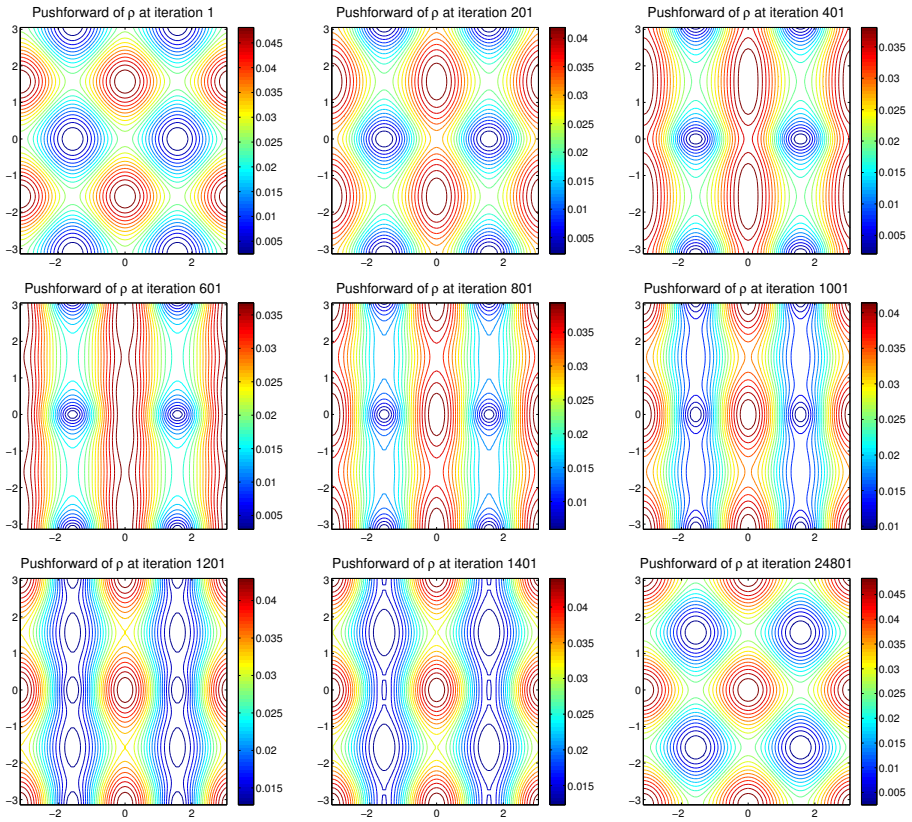

Figure 1: Time evolution of the probability density ρ_t from the $\rho(x_1, x_2)$ defined in (40) (upper left panel) into the target distribution μ obtained from ρ through a translation by $\pi/2$ of the x_2 coordinate (lower right panel).

If the map obtained were optimal, its curl should be zero, except possibly at the location of the branching cuts, where it is undefined. Hence we compute the curl and the divergence of the displacement field, to investigate whether the former is negligible compared to the latter, which provides a natural reference point. Because the displacement field is updated continuously in z-space, particular care has to be taken to compute the curl of the displacement field in the original x-space. This amounts to evaluating a function on an unstructured grid from its Fourier modes. We use for this the fast interpolating method developed by Greengard and Lee in [13].

The average value of the divergence and the curl of the displacement field of the map are documented in figure 2. Even though the L^2 norm of the curl is about two orders of magnitude smaller than that of the divergence, the map is not curl-free (see upper right panel). Other experiments indicate that this

behavior is general: the flow in (33) produces a map transporting ρ to μ with cost low but not strictly optimal. In the sections that follow, we first describe

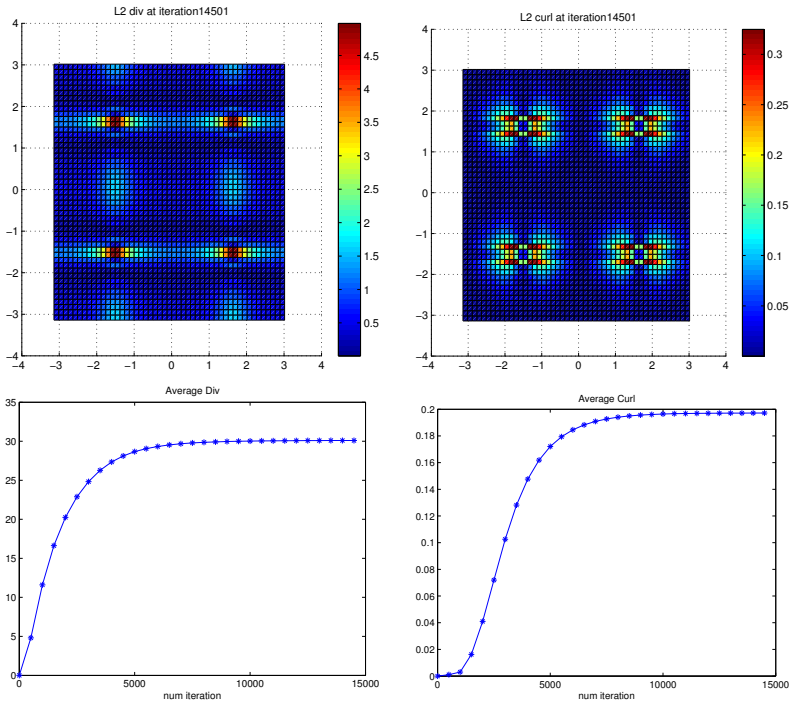

Figure 2: Given the displacement field $M(x) = z(x) - x$ of the map, we compute the average divergence $\int \rho(x) \|\nabla \cdot M(x)\|^2 dx$ and the average curl $\int \rho(x) \|\nabla \times M(x)\|^2 dx$ as the iterations progress (lower left and right panels, respectively). The upper panels report the point-wise absolute value of $\nabla \cdot M(x)$ and $\nabla \times M(x)$ on the periodic domain of the final displacement field.

a data-based, mesh-free version of the flow described above, and then propose a methodology to make the map strictly optimal.

4 Mesh-free implementation

In this section, we adapt the flow above to applications where the densities $\rho(x)$ and $\mu(y)$ are known only through samples $x_i, i \in [1 \dots m]$ and $y_j, j \in [1 \dots n]$ respectively. The following notation will be used:

- $z^k(x)$ is the map at the k-th iteration step, with $z^0(x) = x$ and $z_i^k = z^k(x_i)$.

- $\rho^k(z)$ is the density obtained by mapping $\rho(x)$ via the map $z^k(x)$, and

$$\hat{D}^k \equiv \frac{1}{m}\sum_{i=1}^{m} u(z_i^k) + \frac{1}{n}\sum_{j=1}^{n} u^*(y_j) \tag{41}$$

is the Monte Carlo estimate of the functional

$$\tilde{D}^k(u) = \int_{K_1} u(x)\rho^k(z)dz + \int_{K_2} u^*(y)\mu(y)dy. \tag{42}$$

In the mesh-free version of the algorithm, rather than following the gradient of $\frac{\delta \tilde{D}_t}{\delta u}$, we restrict the maps at each step to a family of perturbations to the identity: $z_i^{k+1} = \nabla u_\beta^k(z_i^k)$, where $u_\beta^k(z) = \frac{z^2}{2} + \sum_l \beta_l F_l^k(z)$, with $F_l^k(z)$ a given set of functions and β a vector of free parameters. At each step k, we perform the following operations:

1. Compute a $\bar{\beta}$ that reduces the value of

$$\hat{D}^k(\beta) = \frac{1}{m}\sum_{i=1}^{m} u_\beta^k(z_i^k) + \frac{1}{n}\sum_{j=1}^{n} \left(u_\beta^k\right)^*(y_j), \tag{43}$$

either through gradient descent:

$$\bar{\beta}_l \propto -\left.\frac{\partial \hat{D}^k(\beta)}{\partial \beta_l}\right|_{\beta=0}, \tag{44}$$

or through Newton's method, as described in section (4.1).

2. Update the sample points,

$$z_i^{k+1} = \nabla u_{\bar{\beta}}^k(z_i^k) \tag{45}$$

as well as any other points z_s^k where the map is sought. If the density $\rho^0(x)$ is known or estimated, as is the case in some applications and in synthetic examples, $\rho^{k+1}(z)$ can also be updated using

$$\rho^k(z) = J^{\nabla u_{\bar{\beta}}^k}(z)\rho^{k+1}(\nabla u_{\bar{\beta}}^k(z)), \tag{46}$$

where $J^{\nabla u_{\bar{\beta}}^k}(z)$ is the Jacobian of the map $\nabla u_{\bar{\beta}}^k$.

The algorithm converges when $\hat{D}^k$ reaches a minimum, and so $\beta \to 0$. Since the derivative of $\hat{D}^k$ with respect to β_l is

$$\left.\frac{\partial \hat{D}^k}{\partial \beta_l}\right|_{\beta=0} = \frac{1}{m}\sum_{i=1}^{m} F_l^k\left(z_i^k\right) - \frac{1}{n}\sum_{j=1}^{n} F_l^k\left(y_j\right), \tag{47}$$

then at convergence we have that

$$\frac{1}{m}\sum_{i=1}^{m} F_l^{\infty}\left(z_i^{\infty}\right) = \frac{1}{n}\sum_{j=1}^{n} F_l^{\infty}\left(y_j\right) : \tag{48}$$

the mapped sample points z_i^{∞} are such that they provide the same Monte Carlo estimates as the samples y_j from $\mu(y)$ for the expected values of all candidate functions F_l^{∞}. This is our sample-driven, weak formulation of the transport condition $\rho^{\infty}(y) = \mu(y)$. Its level of resolution depends on the richness of the family F from where the F_l^k are chosen, with too rich or too small a family yielding over and under resolution respectively. Candidate families F are discussed below.

Notice that this approach does not approximate directly the densities ρ and μ through a linear combination of radial basis functions, as done for instance in [16]: the composition of a finite set of nonlinear functions spans a much richer manifold than their linear combination. Hence approximating the flow through map-composition rather than the densities though linear combinations permits a more accurate characterization of the optimal map. We refer to [25], where a numerical experiment is performed to compare the two approaches in the context of density estimation. A flow similar to (36) is discretized to build up a map from a density ρ -only known through samples- to a normal distribution. The resulting estimation of ρ is then compared with the one obtained through kernel density estimation [28] with radial basis functions. Function composition leads to an estimated density closer -in terms of the KL divergence- to the true one underlying the samples.

4.1 Computing β

Because of the simplicity of each elementary map, one can choose β at each time step so as to minimize the local quadratic approximation to $\hat{D}$

$$\hat{D} \approx \hat{D}_0 + G\beta + \frac{1}{2}\beta^T H \beta \tag{49}$$

where

$$\hat{D}_0 = \hat{D}|_{\beta=0}, \qquad G_i = \left.\frac{\partial \hat{D}}{\partial \beta_i}\right|_{\beta=0}, \qquad H_{ij} = \left.\frac{\partial^2 \hat{D}}{\partial \beta_i \partial \beta_j}\right|_{\beta=0}. \tag{50}$$

Minimizing (49) with respect to β yields

$$\beta = -H^{-1}G, \tag{51}$$

which one may want to cap by a *learning rate* ϵ to avoid big jumps based only on local information:

$$\beta \to \alpha\beta, \quad \text{with } \alpha = \min(1, \frac{\epsilon}{\|\beta\|}).$$

The gradient G of $\hat{D}(\beta)$ has already been computed in (47). Since the Hessian of the first sum in (43) is zero, we only need to compute the Hessian of the second sum, involving the Legendre transform of F. Introducing

$$\bar{z}(y,\beta) = \arg\max_z \left[z \cdot y - \frac{\|z\|^2}{2} - \sum_l \beta_l F_l(z) \right] = y - \sum_l \beta_l \nabla F_l(\bar{z}), \tag{52}$$

we have that

$$\begin{aligned} \left(\frac{\|z\|^2}{2} + \sum_l \beta_l F_l(z) \right)^* (y) &= \max_z \left[z \cdot y - \frac{\|z\|^2}{2} - \sum_l \beta_l F_l(z) \right] \\ &= \bar{z} \cdot y - \frac{\|\bar{z}\|^2}{2} - \sum_l \beta_l F_l(\bar{z}) = \end{aligned} \tag{53}$$

$$= \frac{\|y\|^2}{2} - \sum_l \beta_l F_l(y) + \frac{1}{2} \sum_{i,j} \beta_i \beta_j \nabla F_i(y) \cdot \nabla F_j(y) + O\left(\|\beta\|^3\right),$$

so

$$G_l = \frac{1}{m} \sum_{i=1}^m F_l(z_i) - \frac{1}{n} \sum_{j=1}^n F_l(y_j) \tag{54}$$

and

$$H_{ls} = \frac{1}{n} \sum_{j=1}^n \nabla F_l(y_j) \cdot \nabla F_s(y_j). \tag{55}$$

4.2 The family of potentials F

In order to complete the description of the procedure above, we need to specify the form and number of the functions $F_l(z)$ to use at each time-step. The following considerations apply:

1. In order neither to over-resolve nor to under-resolve the probability densities $\rho^k(z)$ and $\mu(y)$ underlying the samples $\{z_i^k\}$ and $\{y_j\}$, the functions F_l must have length scales consistent with these densities.

2. The function F_l should be tailored so as to work well in high-dimensional settings.

An example of elementary potential satisfying the characteristic stated above is the radially symmetric

$$F_l(z) = r \ \texttt{erf}\left(\frac{r}{\alpha}\right) + \frac{\alpha}{\sqrt{\pi}} e^{-\left(\frac{r}{\alpha}\right)^2}, \tag{56}$$

where $r = |z - \tilde{z}_l|$, and α is a bandwidth dependent on the center $\tilde{z}_l$:

$$\alpha \propto \left(\frac{n_p}{n+m} \left(\frac{1}{\tilde{\rho}(\tilde{z}_l)} + \frac{1}{\tilde{\mu}(\tilde{z}_l)} \right) \right)^{\frac{1}{d}}.$$

Here n_p is the desired number of points to be "captured" by F_l, and $\tilde{\rho}$ and $\tilde{\mu}$ are rough approximations to ρ^k and μ. The elementary map associated with this proposal is given by

$$z^{k+1} = z^k + \sum_l \beta_l f_l(r)(z^k - \tilde{z}_l) \tag{57}$$

where

$$f(r) = \frac{1}{r}\frac{dF}{dr} = \frac{\texttt{erf}(r/\alpha)}{r}. \tag{58}$$

The number of functions $F_l(z)$ to use per time-step is a matter of convenience. One function per step suffices for the algorithm described so far, while two functions is the minimum in the variation below that enforces the map's optimality. Yet more free parameters per step might be desired to accelerate the algorithm's convergence. As for the centers z_l, the simplest choice is to pick them at random from among the $\{z_i^k\}$ and $\{y_j\}$, just making sure that any two centers are not too close to each other. An alternative is to sample more frequently points in areas with low probability density, so as to better resolve the tails of the two distributions.

The rough estimates $\tilde{\rho}$ and $\tilde{\mu}$ used to determine the band-width are computed at time zero, for instance through a simple kernel density estimator. The corresponding $\tilde{\rho}^k(z)$ is then updated multiplying by the Jacobian of the map at each time-step.

4.2.1 A note on the effective supports of ρ and μ

It would be virtually impossible for the functions $F_l(z)$ to resolve simultaneously the $\{z_i\}$ and $\{y_j\}$ if the two densities $\rho^k(z)$ and $\mu(y)$ did not overlap, i.e. if there were regions where one density was significant while the other was negligible or zero. At convergence, $\rho^k \approx \mu$, but there is no reason why the initial $\rho(x)$ could not have an effective support distinct from that of $\mu(y)$.

In order to address this issue, one can pre-process the data-points $\{x_i\}$ so that they lie closer to the $\{y_j\}$, deforming them through an isotropic stretching and rigid translation, without compromising the computation of the optimal map of the original problem. The reason is that the composition of these transformations with optimal maps yields optimal maps. To see this, observe that the composition $h(x) = g \circ s(x)$ of the gradient $g(x) = \nabla_x \phi(x)$ of a convex function ϕ and the linear map $s(x) = ax + b$ can be rewritten as $h(x) = \nabla_x \frac{1}{a}\phi(ax + b)$, where $\frac{1}{a}\phi(ax + b)$ is still a convex function.

More general pre-processing maps not necessarily isotropic, linear or curl-free can be performed, if one then implements the steps described below to enforce the optimality of the final map. A natural choice is to use a first map of the form $y = Ax + b$, with the symmetric matrix A and the vector b chosen so as to transform the empirical mean and covariance of the $\{x_i\}$ into those of the $\{y_j\}$.

We have not implemented any pre-conditioning in the numerical examples of section 6 though, so as to let the core of the procedure speak for itself (the first

example, for instance, would otherwise be fully solved by the pre-conditioning step alone!)

4.3 Enforcing the optimality of the final map

As shown in section 3, even though the flow given by (33) maps ρ to μ, it is generally not curl-free, and therefore not optimal. This is accentuated if one pre-conditions the computation with an arbitrary smooth map that brings the $\{x_i\}$ and $\{y_j\}$ closer to each other. Moreover, the numerics are only an approximation to the smooth flow in (33), so even if this should converge to the optimal map, the mesh-free algorithm may not be so precise (not so for the convergence of ρ_t to μ, since this is enforced via the moving samples themselves. The optimality of the map on the other hand depends on their original position, that the algorithm forgets.) In this section, we modify the discrete version of the flow from section 4 so that the resulting map converges to the optimal one.

One possibility, developed in [15] within a grid-based methodology, is based on the results discussed in [2] and [11] stating that any map $w(x)$ from ρ to μ admits a unique decomposition of the form

$$w(x) = \nabla\phi \circ s(x), \tag{59}$$

where ϕ is a scalar convex function and s maps ρ into itself. Therefore, given $w(x)$ one can seek a *rearrangement* map s^{-1} satisfying $w \circ s^{-1} = \nabla\phi$. In our context, one would attempt to rearrange the points $\{w(x_i)\}$ in order to minimize the cost

$$\frac{1}{m}\sum_{i=1}^{m} |w(x_i) - x_i|^2 \tag{60}$$

while keeping the underlying distribution of $\{w(x_i)\}$ fixed to μ.

Rather than implementing such constrained minimization algorithm to the final map resulting from the flow in (33), we choose to perform the rearrangement gradually along the flow. This results in a much smoother procedure: it takes small changes to alter the flow to make each point move toward its correct target, while fixing the map after the fact involves displacements of order one.

The strategy proposed here, similar to the one developed for constrained density estimation in [19], alternates between two kinds of steps: one that follows the direction of descent of $\hat{D}$ as before, and one that finds a direction along which the original cost decreases while the value of $\hat{D}$ does not deteriorate. Thus, for this second kind of steps, we consider the two objective functions

$$\hat{D}^k(\beta) = \frac{1}{m}\sum_{i=1}^{m} u_\beta^k(z_i^k) + \frac{1}{n}\sum_{j=1}^{n} \left(u_\beta^k\right)^*(y_j) \tag{61}$$

and

$$C^k(\beta) = \frac{1}{m}\sum_{i}^{m} \left|\nabla u_\beta^k(z_i^k) - x_i\right|^2, \tag{62}$$

the Monte Carlo estimates of the functional $\tilde{D}$ and the L^2 cost respectively, with

$$u^k_\beta(z) = \frac{z^2}{2} + \sum_l \beta_l F^k_l(z)$$

as above, but with vectors β of dimension at least two (one is the minimal number of parameters for the first kind of steps.)

The cost-reducing step is described by the following algorithm, illustrated in figure 3:

1. Compute the gradients $\beta_D = -\nabla_\beta \hat{D}^k(\beta)|_{\beta=0}$ and $\beta_C = -\nabla_\beta C^k(\beta)|_{\beta=0}$
2. Find a direction $\bar{\beta}$ such that both $\beta_D \cdot \bar{\beta}$ and $\beta_C \cdot \bar{\beta}$ are positive. One option is to pick $\bar{\beta} = \beta_C$ when $\beta_D \cdot \beta_C > 0$, and otherwise the $\bar{\beta}$ in the line spanned by β_D and β_C such that $\beta_L \cdot \bar{\beta} = \beta_C \cdot \bar{\beta}$ as illustrated in figure (3). An alternative is described below when discussing convergence in the continuous setting.
3. Perform a descent move along the direction $\bar{\beta}$: $z^{k+1} = z^k + \epsilon \nabla_z \sum_l \bar{\beta}_l F^k_l(z)$. Here $\epsilon > 0$ can be a pre-established learning rate, or follow from Newton's method applied to the cost C, capped so that $\bar{D}$ does not increase.

Then, as the density $\tilde{\rho}(z)$ underlying the evolving Lagrangian markers moves toward $\mu(z)$, the $z(x,t)$ are continuously re-arranged so the the cost is minimal.

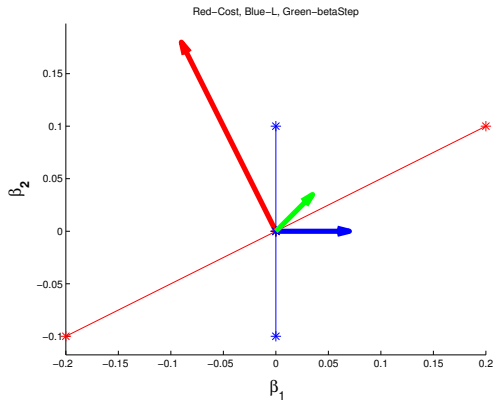

Figure 3: Direction of descent for the cost-decreasing step. The red and blue arrows represent β_C and β_D respectively; the thinner segments delimited by $*$'s are normal to these. The green arrow in the chosen direction of descent $\bar{\beta}$ has positive projection onto both β_C and β_D. This plot depicts the minimal situation with two free parameters β_l, but the general situation is entirely similar when looked at in the plane in β-space spanned by β_C and β_D.

In order to build a continuum flow version of this discrete algorithm, let us consider the problem first in general terms: given two smooth convex cost

functions $C_1(x), C_2(x)$, one defines the set

$$\Omega = \arg\min C_1(x),$$

presumably a manifold with more than one point, and seeks

$$x^* = \arg\min_{x\in\Omega} C_2(x).$$

Our procedure is a variation of gradient descent: we introduce a flow

$$\dot{x} = -\nabla_x C_1 - P\left(\nabla_x C_2\right), \tag{63}$$

where P is the projection onto the cone $\{z : z \cdot \nabla_x C_1 \geqslant 0\}$, defined as follows:

$$P(w) = \begin{cases} w & \text{if } w \cdot \nabla_x C_1 \geqslant 0 \text{ and } x \notin \Omega \\ w - \frac{w\cdot\nabla_x C_1}{\|\nabla_x C_1\|^2}\nabla_x C_1 & \text{if } w \cdot \nabla_x C_1 < 0 \\ \text{proj}_\Omega w & \text{if } x \in \Omega \end{cases}$$

Then we have that

$$\frac{d}{dt}C_1(x(t)) = \dot{x} \cdot \nabla_x C_1 \leqslant 0,$$

with equality only when $\nabla_x C_1 = 0$ (i.e. once x has reached the set Ω), and $\dot{x} = 0$ only once $\nabla_x C_1 = 0$ and $\nabla_x C_2 \cdot \nabla_x C_1 = 0$, the differential conditions characterizing x^*.

In our case, $C_1 = \tilde{D}_t$ and $C_2 = C$, with the additional ingredient that we combine the variational derivative of C with respect to z,

$$\frac{\delta C_t}{\delta z} = 2\rho(x)(z - x) \tag{64}$$

with the z-gradient of the variational derivative of $\tilde{D}_t$ with respect to u:

$$\nabla_z \left(\frac{\delta D_t}{\delta u}\right) = \nabla_z \left[\rho_t(z) - \mu(z)\right].$$

Thus the continuous time formulation yields the following flow:

$$\dot{z} = -\nabla_z \left[\rho_t(z) - \mu(z)\right] - P\left[2\rho(x)(z - x)\right], \tag{65}$$

where for convenience in the proof below we define the projection P as above using as inner product

$$u \cdot v = \int \rho_t \; u(z) \cdot v(z) \; dz.$$

To see that this flow has $\rho_\infty = \mu$, we resort again to the Liouville equation

$$\frac{\partial \rho_t(z)}{\partial t} + \nabla_z \cdot [\rho_t(z)\dot{z}] = 0,$$

with $\dot{z}$ from (65). In terms of the L^2 norm

$$\|\rho_t - \mu\|^2 = \int (\rho_t(z) - \mu(z))^2 dz, \tag{66}$$

we have that

$$\frac{d}{dt}\frac{\|\rho_t - \mu\|^2}{2} = -\int \rho_t \nabla(\rho_t - \mu) \cdot \{\nabla_z [\rho_t(z) - \mu(z)] + P[2\rho(x)(z-x)]\} dz \leq 0,$$

since both summands have positive integral: the first because the integrand is positive, and the second because of the definition of the projection P. Moreover, equality can only hold when $\rho = \mu$, thus proving that the flow transports ρ into μ. To see that the final map is also optimal, we notice that, once $\rho_t = \mu$, we have

$$\dot{z} = -\text{proj}_\Omega \nabla C,$$

so the cost C will decrease until $\text{proj}_\Omega \nabla C = 0$, the condition for constrained optimality.

In the discrete version proposed, one alternates between two time steps, modeling the two terms on the right-hand side of (63). In the first time step, one simply descends C_1 through one step of Newton's method, i.e.

$$x^{n+\frac{1}{2}} = x^n - H_1^{-1} \nabla C_1,$$

capped so as not to take steps that are too big. Here the Hessian H_1 and gradient of C_1 are evaluated at $x = x^n$. In the second step, one performs a simplified line search, also through one step of Netwon's method:

$$x^{n+1} = x^{n+\frac{1}{2}} - \gamma\, d,$$

where

$$\gamma = \frac{d^T \nabla C_2}{d^T H_2 d}$$

It is convenient for this step not to adopt for d simply the line defined by the projection P in the second term on the right hand-side of (63), but rather to combine it with a fraction of the first term:

$$d = -\left[P\left(\nabla_x C_2\right) + \alpha \nabla_x C_1\right],$$

with the extra requirement that

$$\alpha \leqslant \frac{\|\nabla C_2\|^2}{|\nabla C_1 \cdot \nabla C_2|} - \frac{|\nabla C_1 \cdot \nabla C_2|}{\|\nabla C_1\|^2}$$

so as to guaranty that C_2 decreases along this line. The rationale for adding this term is to have C_1 also decrease to leading order rather than being merely stationary, which could lead to an increase for any finite step-size.

5 General cost functions

For strictly convex costs of the form $c(x,y) = c(x-y)$ and sufficiently regular densities ρ and μ, one can show (see for instance [3]) that the solution to the dual Kantorovich problem is uniquely given by the pair

$$\begin{cases} \bar{u}(x) = \min\limits_{y}(c(y-x) - \bar{v}(y)) \equiv v^*(x) \\ \bar{v}(y) = \min\limits_{x}(c(y-x) - \bar{u}(x)) \equiv u^*(y), \end{cases} \tag{67}$$

with the corresponding unique solution to the Monge problem given by

$$y(x) = x - (\nabla c)^{-1}[\nabla \bar{u}(x)]. \tag{68}$$

From the algorithmic viewpoint, the use of a different costs entails a different map associated with the potential optimizing $\tilde{D}_t$. As in the L^2 case, once one has found the value of the parameters β optimizing $\tilde{D}_t$, one moves the sampled points according to the map associated with the corresponding potential (as does (45) for the L^2 case.) This can be implemented easily if ∇c is a one to one function and its inverse has a known closed form. A family of convex cost functions with this property are the L^p costs $c(x-y) = \frac{\|x-y\|^p}{p}$ with $p > 1$: when $c(x) = \frac{\|x\|^p}{p}$, the inverse of ∇c is given by ∇h, where $h(x) = \frac{\|x\|^q}{q}$ and p and q are related by $\frac{1}{p} + \frac{1}{q} = 1$.

The computation of the optimal map associated with the potential u also provides the basis for solving (52) and obtaining an expansion of u^* in powers of β as in (53). In particular, the generalization of (52) to the L^p case is

$$\bar{z} = \arg\min_{z} \left(\frac{\|z-y\|^p}{p} - \beta F(z) \right) \tag{69}$$

(where for clarity we have adopted a scalar β), so

$$\|\bar{z} - y\|^{p-2}(\bar{z} - y) - \beta \nabla F(\bar{z}) = 0. \tag{70}$$

Using the inverse of ∇c described above we obtain from (70)

$$\bar{z} = y - \|\beta \nabla F(\bar{z})\|^{q-2}(\beta \nabla F(\bar{z})), \tag{71}$$

which substituted in the definition of u^* yields

$$\min_{z} \left(\frac{\|z-y\|^p}{p} - \beta \nabla F(z) \right) = -\beta \nabla F(z) + \frac{\|\beta \nabla F(y)\|^q}{q} + O\left(|\beta|^{2q-1}\right), \tag{72}$$

showing that the leading order in β for u^* is the same for any L^p cost function with $p > 1$, while higher order corrections depend on p.

The corresponding continuous flow reads

$$\begin{cases} \dot{z} = \nabla_z \left(\rho_t(z) - \mu(z)\right) \|\nabla_z \left(\rho_t(z) - \mu(z)\right)\|^{q-2} \\ z(x,0) = x \end{cases} \tag{73}$$

a flow defined to ascend rather than descend $\tilde{D}_t$, since the change of variable $u \to \frac{x^2}{2} - u$ specific to the L^2 cost is not used here.

The argument for convergence in the L^2 case can be applied for any $p > 1$: the continuity equation implies that

$$\frac{d}{dt} \int |\rho_t - \mu|^2 = - \int \rho_t \{\dot{z} \cdot \nabla(\rho_t - \mu)\} = \int \rho_t \|\nabla(\rho_t - \mu)\|^q \leq 0$$

proving convergence of ρ_t to μ.

Summarizing, the implementation of the algorithm when using a general L^p cost does not differ conceptually from the L^2 case. The only differences are in the computation of the Hessian (see (55) for the L^2 case) used for Newton's method and in the expression in (64) characterizing the cost-step, whose generalization to the L^p case is straightforward.

In applications to economics, one frequently encounters costs that are non-convex, including strictly concave costs such as the Coulomb cost $c(x, y) = \frac{1}{\|x-y\|}$, which has been recently introduced in the context of density functional theory [5]. In contrast to convex costs, a strictly concave cost favors solutions where, rather than moving two masses a distance l each, only one is moved a distance $2l$, an *economy of scale* typical in microeconomics. The mathematical theory of optimal transport has been extended to strictly concave costs in [12]. The resulting optimal maps are much less smooth than in the convex case, and their global nature makes an approach based on local moves far less natural. Hence even though the formulation for sample-based transport proposed in this paper applies to general costs, the flow-based algorithm is restricted to costs that are strictly convex.

6 Numerical results

We now illustrate the procedure proposed in this article with some numerical experiments. In all cases, we have used synthetic data, in the form of samples drawn from two known distributions $\rho(x)$ and $\mu(y)$. In most experiments, the optimal map $y = \bar{T}(x)$ between the two is also known in closed form. This allows us to check the accuracy of the numerical results in a number of ways:

1. Computing the average error e on the sample points between z^k and the optimal map $\bar{T}$: $e = \frac{1}{m} \sum_{i=1}^{m} |z^k(x_i) - \bar{T}(x_i)|^2$

2. Computing the empirical Kullback-Leibler divergence

$$\frac{1}{m} \sum_j \log \left(\frac{\rho(x_j)}{\rho^k(x_j)} \right)$$

between the estimated density $\rho^k(x) = J^k(x)\mu(z^k(x))$ and the exact initial density $\rho(x)$.

3. Introducing a grid x_g, whose points are treated as passive Lagrangian markers that move with the flow $z^k(x)$ without affecting it, unlike the sample points x_j that guide the flow through the objective function $\tilde{D}$. Then we can compute and plot the discrepancy between the numerical optimal map T and $\bar{T}$ on the grid points x_g.

4. Estimating the density ρ on the grid x_g though

$$\rho^k(x_g) = J_{z^k}(x_g)\mu(z^k(x_g)) \tag{74}$$

and comparing to the exact density $\rho(x)$.

5. Estimating the target density μ through

$$\mu_{est}(z^k(x_g)) = \frac{\rho(x_g)}{J_{z^k}(x_g)} \tag{75}$$

and comparing with the exact μ.

6. Same as 1. and 2., but replacing the empirical KL and cost by their calculation on the grid x_g:

$$KL = \int \log\left(\frac{\rho(x)}{\rho^k(x)}\right)\rho(x)dx \approx \sum \log\left(\frac{\rho(x_g)}{\rho^k(x_g)}\right)\rho(x_g)\Delta x_g$$

$$C = \int |z^k(x) - x|^2\rho(x)dx \approx \sum |z^k(x_g) - x_g|^2\rho(x_g)\Delta x_g,$$

where Δx_g is the volume of each cell in the grid.

Non isotropic Gaussian to non isotropic Gaussian

Figures (4) and (5) display the results of applying the procedure to the optimal transport between two non-isotropic Gaussian distributions ρ and μ, an experiment run frequently in the literature (see for instance [9] and [21]). The optimal map between two Gaussians with zero mean and covariance matrices Σ_1 and Σ_2 is given by [7]

$$y = \bar{T}x, \quad \text{with} \quad \bar{T} = \Sigma_2^{1/2}\Sigma_0^{-1}\Sigma_2^{1/2}, \tag{76}$$

where

$$\Sigma_0 = \left(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2}\right)^{1/2}. \tag{77}$$

The upper left and right panels of figure (4) display the initial ρ and the target distribution μ, with covariance matrices

$$\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.25 \end{pmatrix} \tag{78}$$

respectively, while the lower panels display their estimates through (74) and (75).

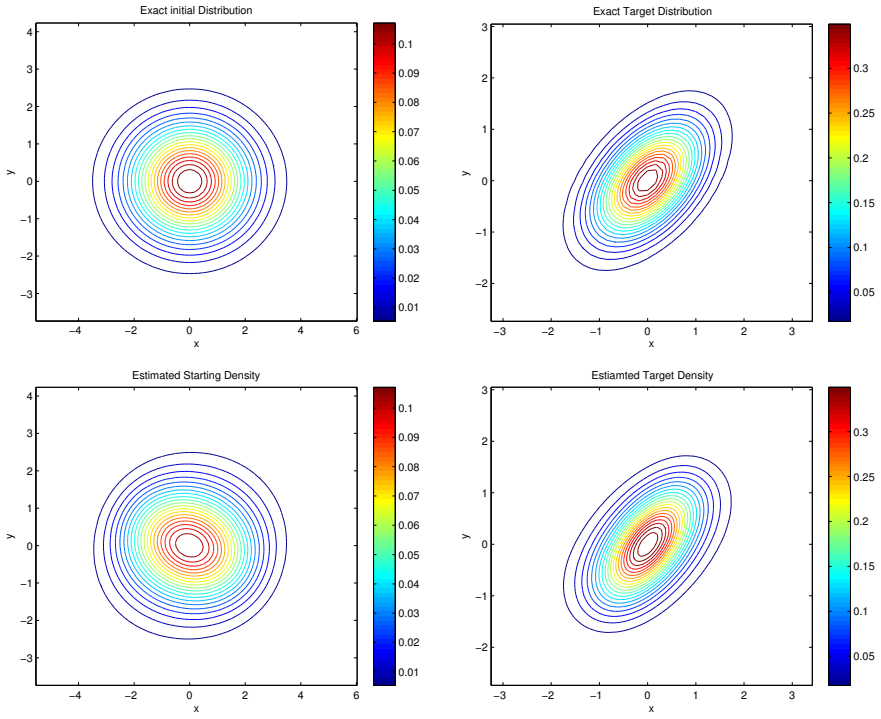

Figure 4: Upper panels: exact starting distribution ρ (left) and exact target distribution μ (right). Lower panels: estimated starting distribution (left) and estimated target distribution (right). The algorithm uses 10^3 points sampled independently from each ρ and μ. The two estimated distributions are computed using the fact that one knows the exact form for the other distribution, the map that the algorithm found and its Jacobian.

The fact that the KL divergence converges to zero (see figure (5)) and that the cost of the map z^k approaches the cost of the the exact map $\bar{T}$ (both evaluated using the grid points x_g and reported in the panel in the middle row of the second column of figure (5)) shows that z^k converges to $\bar{T}$. This is confirmed by the two lower panels reporting the average error (left) and the pointwise absolute value of the difference between the numerical map and the exact one.

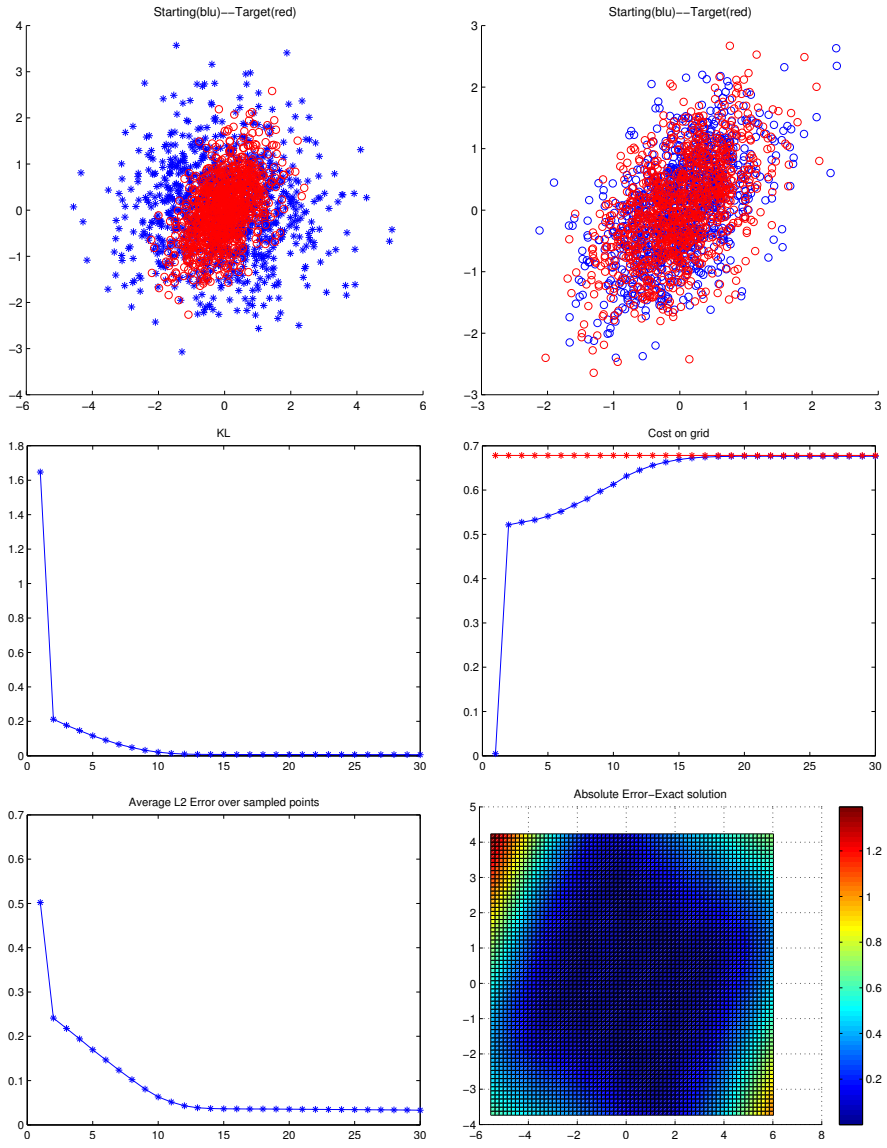

Figure 5: First row, left panel: sampled points independently drawn from ρ (blue) and μ (red); right panel: the blue points have now been mapped by the algorithm and underly the same distribution as the red ones. Second row, left: Leibler-Kullback divergence between $\rho^k(x)$ and $\rho(x)$; right: cost of the map z^k at iteration k in blue and cost of the optimal map $\bar{T}$ in red. Third row: average error (left) and absolute value of the difference between z^k and $\bar{T}$ (right).

Mass splitting in two dimensions

This subsection illustrates mass splitting. The exact optimal map between a normal $\rho(x_1, x_2)$ and a target distribution

$$\mu(y_1, y_2) = \frac{9 y_1 y_2}{2\pi} e^{-\frac{y_1^6 + y_2^6}{2}} \tag{79}$$

is given by

$$y = \bar{T}(x) = \frac{1}{3} \begin{pmatrix} {x_1}^{1/3} \\ {x_2}^{1/3} \end{pmatrix}. \tag{80}$$

The numerical computation of this map is comparatively intensive due to the unbounded gradient of the map along the Cartesian axes: since the support of the initial density ρ is connected while that one of μ is not, the mass initially supported in the domain of ρ needs to be split into the different components of the domain of μ. The results are displayed in figures 6 and 7.

We close this subsection with a note regarding the optimality of the map. Figure 8 reports results for the same data above but computed without using the cost-reducing step of section (4.3). In particular, we compare the values of the cost and the KL obtained using the cost-reducing step (figure 7) with the values obtained without using it (figure 8). Even though the cost of the map is lower when using the cost-reducing step, the map obtained without it is still a very good approximation to the optimal one. This is quite surprising considering that the flow described in (33) has no memory of the initial distribution of the sampled points $\{x_i\}$.

Compact support

One advantage of computing the optimal map using only sample points from the starting and the target distributions is to avoid dealing with boundary conditions when the distributions have compact support. By contrast, using a PDE-based method involves equipping the Monge-Ampere equation with a set of non trivial boundary conditions (see for instance [9]). Since the elementary map in (57) is also defined outside the support of the distributions, the fact that these have compact support is hidden in the Monte-Carlo estimates of the functional (43).

Figure 9 reports the results of mapping a uniform distribution supported on a square centered at the origin with side $L = 3$ to another uniform square distribution centered at the origin but with side $L = 2$. The points sampled from the starting distribution (in blue in the upper left panel of figure 9) are mapped into points whose distribution overlaps quite accurately with the support of the target distribution spanned by the red points (upper right panel), showing that the boundary of the starting distribution is accurately mapped onto the boundary of the target distribution.

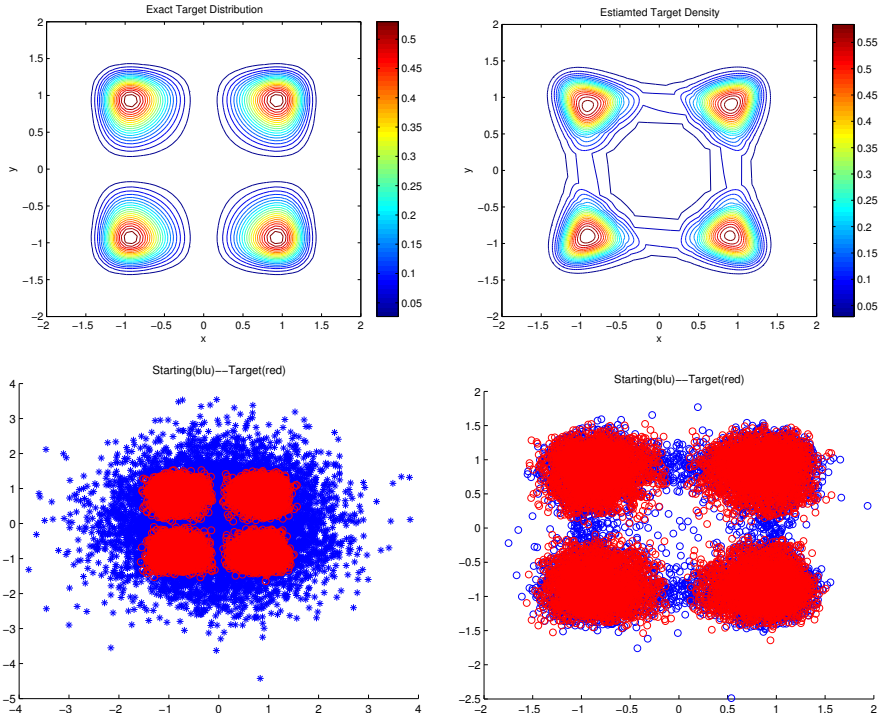

Figure 6: Results from mapping an isotropic normal to the distribution defined by (79). The upper left panel depicts the exact (left) and estimated (right) target distributions. The lower panels depict the data sampled from the starting (blue) and the target (red) distributions on the left and the same starting samples mapped through the numerical optimal map and again the target on the right. The algorithm uses 10^4 points sampled from each ρ and μ.

Higher dimensional tests

An example of optimal transport in dimension higher than three is given in figures 10 and 11,where we map a non isotropic Gaussian with covariance matrix Σ given by a 7×7 diagonal matrix with diagonal entries $\Sigma_{ii} = 1, 1.5, 2, 2.5, \cdots, 4$ to a standard isotropic normal distribution. Figure 10 shows the projection along three different planes of the estimated starting density ρ_{est} (panels on the left column). The projections of ρ_{est} should be compared with the values of the exact starting density ρ projected on the same planes (panels on the right column) . Each plane is specified by the two coordinates labeling the axis of each panel in figure 10 and by setting all the entries relative to the other coordinates to zero. The first row, for instance, displays the contour lines of the projection of ρ_{est} (left panel) and ρ (right panel) on the (x_4, x_5) plane.

Since $\Sigma_{11} = 1$ and $\Sigma_{77} = 4$, the biggest deformation needed in order to

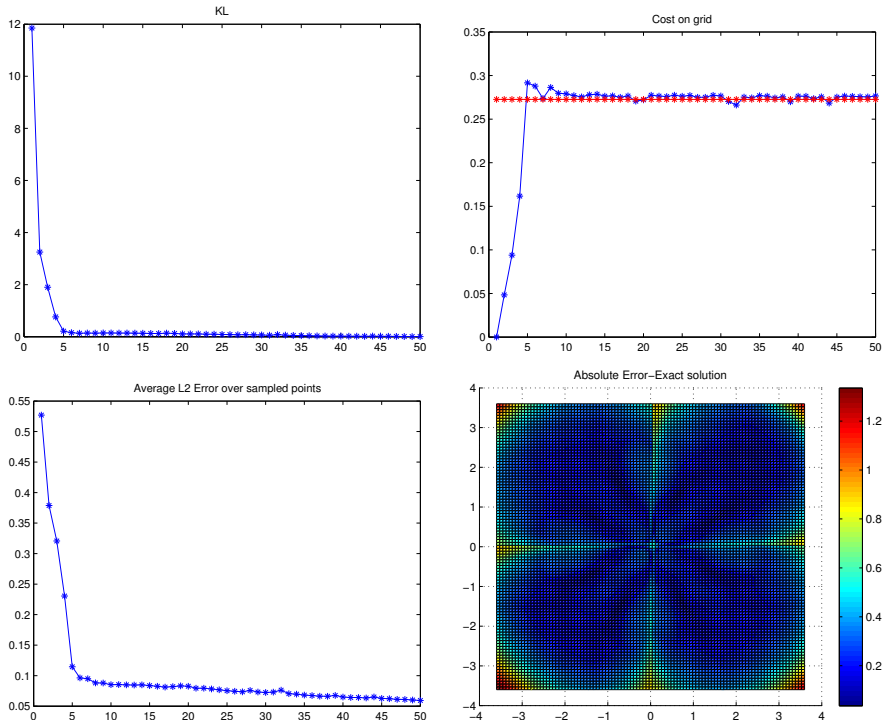

Figure 7: Same quantities described in figure (5) but relative to the splitting map.

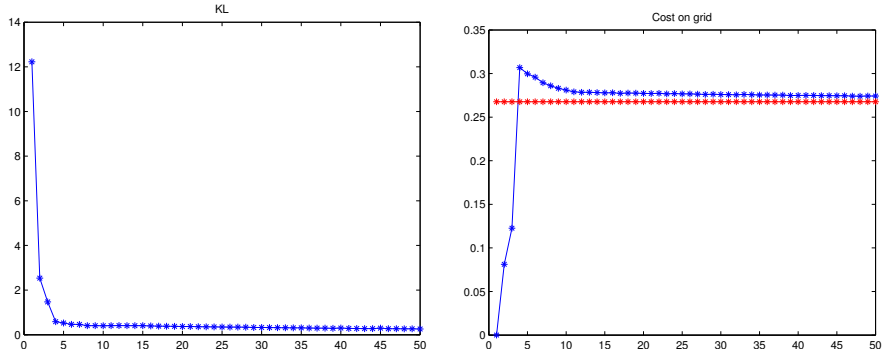

Figure 8: Same as the upper row panels of figure (7) but computed without alternating the cost-reducing step described in (4.3).

map ρ into μ occurs along the (x_1, x_7) plane. Looking at the projection of ρ_{est}

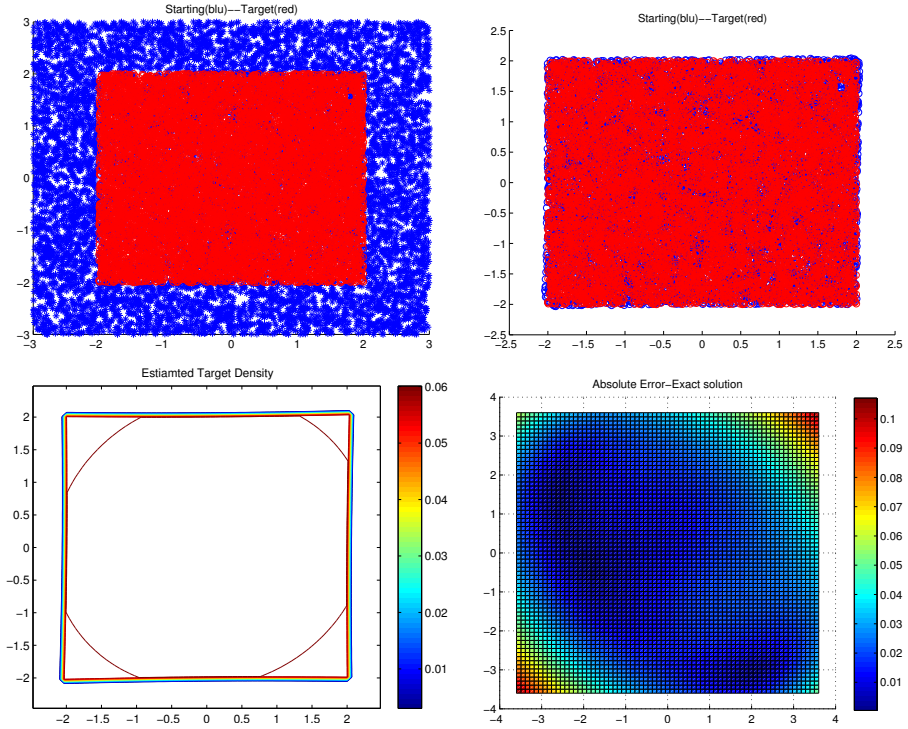

Figure 9: Mapping between two uniform distributions with compact support on two squares of different sizes, using 10^4 points sampled from each ρ and μ. Upper panels: sample points from both distributions before and after the map. Lower panels: estimated target density and pointwise difference between the exact and numerical maps.

on this plane one can indeed notice that the projection is not as accurate in reproducing ρ as those on the (x_4, x_5) and (x_2, x_6) planes.

Figure 12 reports the value of the execution time per iteration, obtained using 10^4 points sampled from each ρ and μ, and dimensions ranging from 2 to 10. The time is computed running the algorithm for a fixed amount of iterations –10^4– and then taking the average over all their execution times. Since the computational cost of the algorithm is essentially the one needed to compute β in equation (44), the time per iteration should scales linearly with the dimension of the space, as confirmed in figure 12. Of course this does not mean that the execution time for the algorithm to converge to an acceptable solution need also to scale linearly with the dimension, since the number of iterations may also depend on the dimension, as well as on the specific form of the starting and target distributions. Yet the linear dependence of the computational cost per iteration on the number of dimensions makes the algorithm a good candidate

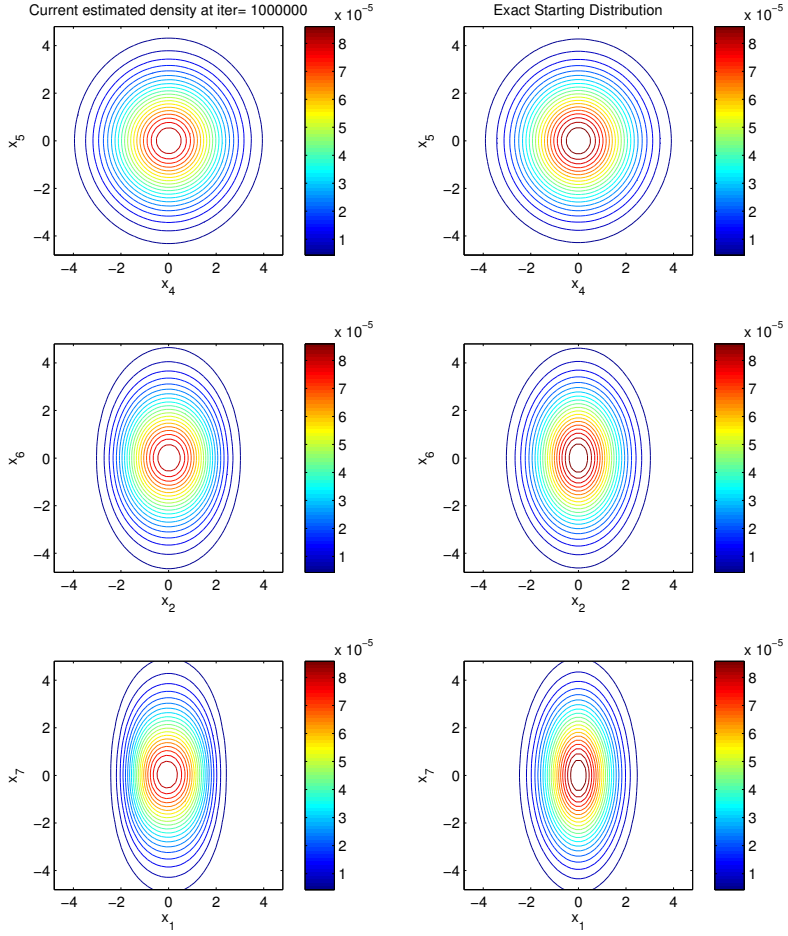

Figure 10: Results from a seven dimensional example, showing the estimated and exact densities along three planes.The algorithm uses 10^4 points sampled from each ρ and μ.

to solve high-dimensional problems. By contrast, an algorithm based on a grid would demand a time per iteration growing exponentially with the dimension of the space.

7 Conclusions and extensions

This paper proposes and explores a new data-driven formulation of the L^2 Monge-Kantorovich optimal transport problem and a mesh-free algorithm for finding its numerical solution. The formulation is based on discrete sample

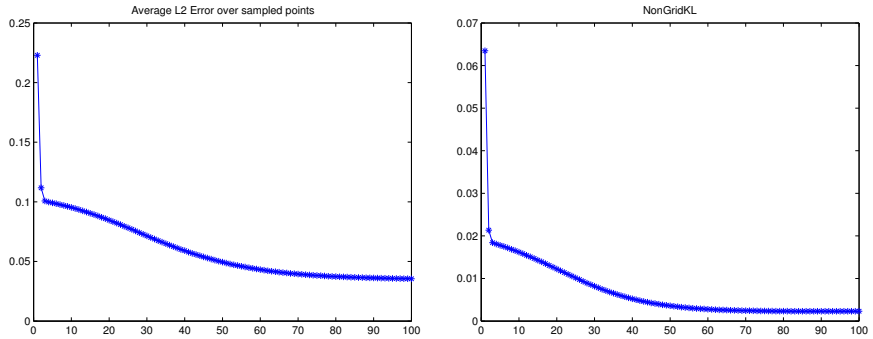

Figure 11: Convergence of the Kullback-Leibler divergence and of the average error, both computed on the sample points.

points from the starting and the target distribution, but under the assumption that these are drawn from continuous probability distributions. This contrasts with the purely combinatorial problem of finding the correspondence between two discrete data sets minimizing a given cost function, and the continuous approach in which the two distributions $\rho(x)$ and $\mu(y)$ are known pointwise. The former reduces to a standard linear programming problem, while the latter has been addressed in most numerical approaches through the introduction of a grid where a partial differential equation is approximately solved. The Monte Carlo-like, particle-based approach proposed here scales much better in high dimensions. More fundamentally, in many applications the two distributions to map onto each other are not known other than through samples.

This new formulation may be extended in various directions. In particular, it is left to further work the discussion of which weak formulation of Kantorovich's dual problem is the most appropriate to data-driven optimal transport. This article instead discusses some general principles, and then proceeds to characterize an appropriate formulation algorithmically. A fully analytical treatment of the kind of adaptive weak solutions proposed here may be out of reach, but even partial steps in this direction could prove very rewarding.

As discussed in the introduction, other extensions left to further work are the mixed scenario where one distribution is known in closed form and the other only through samples, necessary for density estimation, and a generalization of Kantorovich's formulation of the transport problem to situations where the original and target space have different dimensions, as required for regression. Still other extensions down the line include the situation with more than two distributions and the inclusion of constraints additional to the marginals of π, such as the Martingale condition in describing some stochastic processes. Last but not least, the procedure developed here can be extended to costs different from the classical L^2. This is relatively straightforward for convex cost functions, no so for non-convex costs, which might require a substantially different

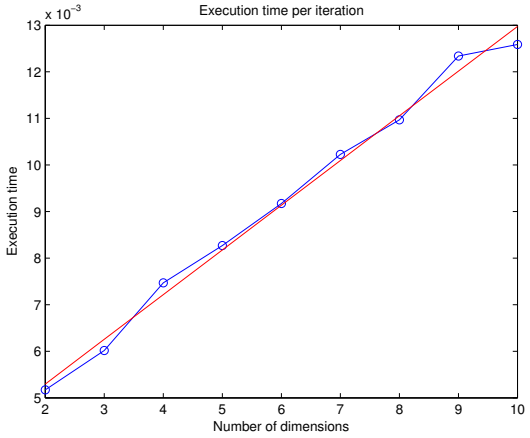

Figure 12: Dependence of the average time per iteration (in seconds) on the dimension of the space, in a computation using 10^4 sample points per distribution.

approach.

Acknowledgments

This work was partially supported by the Division of Mathematical Sciences of the National Science Foundation, under Grant Number DMS-1211298.

References

[1] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[2] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[3] Luis A Caffarelli. The Monge-Ampère equation and optimal transportation, an elementary review. In *Optimal transportation and applications*, pages 1–10. Springer, 2003.

[4] Rick Chartrand, K Vixie, B Wohlberg, and Erik Bollt. A gradient descent solution to the Monge-Kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009.

[5] Codina Cotar, Gero Friesecke, and Claudia Klüppelberg. Density functional theory and optimal transportation with coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013.

[6] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[7] DC Dowson and BV Landau. The Frechet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[8] Lawrence C Evans. Partial differential equations and Monge-Kantorovich mass transfer. pages 65–126. International Press, Cambridge, 1997.

[9] Brittany D Froese. A numerical method for the elliptic Monge-Ampère equation with transport boundary conditions. *SIAM Journal on Scientific Computing*, 34(3):A1432–A1459, 2012.

[10] Brittany D Froese and Adam M Oberman. Fast finite difference solvers for singular solutions of the elliptic Monge-Ampère equation. *Journal of Computational Physics*, 230(3):818–834, 2011.

[11] Wilfrid Gangbo. An elementary proof of the polar factorization of vector-valued functions. *Archive for rational mechanics and analysis*, 128(4):381–399, 1994.

[12] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[13] Leslie Greengard and June-Yub Lee. Accelerating the nonuniform fast Fourier transform. *SIAM review*, 46(3):443–454, 2004.

[14] Eldad Haber, Tauseef Rehman, and Allen Tannenbaum. An efficient numerical method for the solution of the L_2 optimal mass transfer problem. *SIAM Journal on Scientific Computing*, 32(1):197–211, 2010.

[15] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of Computer Vision*, 60(3):225–240, 2004.

[16] Angelo Iollo and Damiano Lombardi. A lagrangian scheme for the solution of the optimal mass transfer problem. *Journal of Computational Physics*, 230(9):3430–3442, 2011.

[17] M Chris Jones, James S Marron, and Simon J Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

[18] Leonid V Kantorovich. On the translocation of masses. *Compt. Rend. Akad. Sei*, 7:199–201, 1942.

[19] Peter Laurence, Ricardo J Pignol, and Esteban G Tabak. Constrained density estimation. *Proceedings of the 2011 Wolfgang Pauli Institute conference on energy and commodity trading, Springer Verlag.*

[20] Robert J McCann. A convexity principle for interacting gases. *advances in mathematics*, 128(1):153–179, 1997.

[21] Robert J McCann and Adam M Oberman. Exact semi-geostrophic flows in an elliptical ocean basin. *Nonlinearity*, 17(5):1891, 2004.

[22] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais.* De l'Imprimerie Royale, 1781.

[23] Nadav Recca. *A new methodology for importance sampling.* Master thesis, Courant Institute, New York University, 2011.

[24] Mohamed M Sulman, JF Williams, and Robert D Russell. An efficient approach for the numerical solution of the Monge-Ampère equation. *Applied Numerical Mathematics*, 61(3):298–307, 2011.

[25] EG Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.

[26] Neil S Trudinger and Xu-Jia Wang. The Monge-Ampère equation and its geometric applications. *Handbook of geometric analysis*, 1:467–524, 2008.

[27] Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2003.

[28] Larry Wasserman. *All of nonparametric statistics.* Springer, 2006.