

Iterative Bregman Projections for Regularized Transportation Problems

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, Gabriel Peyré

▶ **To cite this version:**

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, Gabriel Peyré. Iterative Bregman Projections for Regularized Transportation Problems. SIAM Journal on Scientific Computing, 2015, 2 (37), pp.A1111-A1138. 10.1137/141000439 . hal-01096124

HAL Id: hal-01096124
https://hal.science/hal-01096124

Submitted on 16 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Iterative Bregman Projections for Regularized Transportation Problems

Jean-David Benamou* Guillaume Carlier† Marco Cuturi‡
Luca Nenna* Gabriel Peyré†

December 16, 2014

Abstract

This article details a general numerical framework to approximate solutions to linear programs related to optimal transport. The general idea is to introduce an entropic regularization of the initial linear program. This regularized problem corresponds to a Kullback-Leibler Bregman divergence projection of a vector (representing some initial joint distribution) on the polytope of constraints. We show that for many problems related to optimal transport, the set of linear constraints can be split in an intersection of a few simple constraints, for which the projections can be computed in closed form. This allows us to make use of iterative Bregman projections (when there are only equality constraints) or more generally Bregman-Dykstra iterations (when inequality constraints are involved). We illustrate the usefulness of this approach to several variational problems related to optimal transport: barycenters for the optimal transport metric, tomographic reconstruction, multi-marginal optimal transport and in particular its application to Brenier's relaxed solutions of incompressible Euler equations, partial un-balanced optimal transport and optimal transport with capacity constraints.

1 Introduction

1.1 Previous Works

The theory of Optimal Transport (OT) [58] defines a natural and useful geometry to compare measures supported on metric probability spaces. Its modern formulation as a linear program is due to Kantorovich [40]. OT has recently found a flurry of applications in various fields such as computer vision [50], economy [22] [19], computer graphics [11], image processing [60], astrophysics [34].

*INRIA, MOKAPLAN, jean-david.benamou@inria.fr, luca.nenna@inria.fr
†Ceremade, Université Paris-Dauphine, {carlier,peyre}@ceremade.dauphine.fr
‡Kyoto University, mcuturi@i.kyoto-u.ac.jp

Computational optimal transport. A major bottleneck that prevents the widespread of OT and its various generalizations is the lack of fast (possibly approximate) algorithms. Discrete optimal transport (*i.e.* computing transport between sums of Diracs) reduces to a finite dimensional linear program. When the mass of each Dirac is constant and the two measures have the same number N of Dirac masses, this problem reduces to an optimal matching problem, for which dedicated discrete optimization methods exist [17], that roughly have $O(N^3)$ complexity, which is still computationally too demanding for most applications. Another line of research, initiated by [7] relies on dynamic formulations, which corresponds to computing the transport as a geodesic, and can be re-casted as a convex optimization problem. We refer to [46] for an overview of several proximal optimization methods to tackle this problem. This requires adding an extra dimension (time variable along the geodesic) and is thus also computationally expensive. Semi-discrete optimal transport, i.e. optimal transport from a density to a weighted sum of dirac masses is a classical strategy for a generalised version of the Monge-Ampère equation, see [44] for recent improvements of this approach. Finally and for the quadratic ground cost optimal transport, a direct Newton solver approach to the non-linear Monge Ampère equation can be used to compute density to density optimal transport. This holds under some regularity assumption on the densities and domain and hence on the transport map itself, see [43] and [8] for instance.

Entropic regularization. A different approach consists in computing a regularized version of the OT problem. An interesting choice consists in penalizing the entropy of the joint coupling. This idea can be traced back to Schrodinger [53] and can be related to the the so-called iterative proportional fitting procedure (IPFP) [28] which has found numerous applications in the probability and statistics literature. We refer to [52, 42] for modern perspectives on this problem. Such a regularization also appears in the economy literature, where OT theory can be useful to predict flows of commodities or people in a market. In that context, regularizing the OT problem can also ensure the smoothness of such flows [59, 31] or facilitate inference in matching models [35]. Such a regularization was also recently introduced in [26] where it is shown that, in addition to favorable computational properties (parallelization, quadratic complexity) detailed below, such a regularization also yields a distance between histograms that can perform better in classification tasks than the usual OT distances.

The underlying idea of an entropic regularization is that entropy forces the solution to have a spread support, thus deviating from the fact that optimal couplings are sparse (*i.e.* supported on a graph of a transport plan solving Monge's problem). A first impact of this regularization is that this non-sparsity of the solution helps to stabilize the computation. This can be related to the fact that entropic penalization defines a strongly convex program (as opposed to the initial OT problem) with a unique solution.

Another (even more important) advantage of this entropic regularized OT

problem is that its solution is a diagonal scaling of e^{-C}, the element-wise exponential matrix of $-C$, where C is the ground cost defining the transport (see Section 3.1 for more details). The solution to this diagonal scaling problem can be found efficiently through the IPFP iterative scheme [28]. This algorithm was later studied in detail by Sinkhorn in [54, 56, 55] and its convergence proof was extended to continuous measures in [51].

Entropic regularization of linear programs also shares some connection with interior point methods. These approaches make use of a log-barrier function, which should be self-concordant to ensure a polynomial complexity for a given accuracy [45]. Such a property does not hold for the entropic barrier, so it is not a competitive approach when it comes to approximating solutions of the original linear program by lowering the amount of regularization. As we advocate in this present paper, entropic regularization has however several other computational advantages (in particular because of its close connection with Kullback-Leibler projections), which makes it attractive when a slight amount of smoothing is acceptable in the computed approximation.

Optimization using the Kullback-Leibler Divergence. When considering optimization over the simplex or the cone of positive vectors, it makes sense to replace the usual Euclidean metric by a divergence that quantifies with more relevance the difference between two vectors. Of particular interest for our work is the Kullback-Leibler (KL) divergence, since it is intimately related to entropic regularization. The simplest algorithmic block that can exploit such a divergence is the iterative projection on affine subsets of such cones under the KL divergence, which was introduced by Bregman [12]. Computing the projection on the intersection of generic convex sets requires to replace iterative projections by more complicated algorithms, such as for instance Dykstra's method [29]. This algorithm is extended to Bregman divergences (such as KL) in [21] and a proof of convergence is given in [6]. For references in probability and statistics that also address the case of continuous distributions see [25, 30, 51, 9]. Note that several other proximal algorithms have been extended to this setting [5].

OT Barycenter. The OT metric has been extended in many ways. A natural extension is to consider the barycenter between several distributions (the case of only 2 measures defining the usual transport). Such a barycenter is defined as the solution of a convex variational problem (a weighted sum of OT distances) over the space of measures, which is studied in details in [2]. OT barycenters find applications for instance in statistics to define a mean empirical estimator from a family of observed histograms [10], or in machine learning [27] to provide an extended definition of k-means clustering and compute average histograms-of-features under the OT metric.

Solving this variational problem is challenging. Two recent numerical works have addressed this problem: [27] where a gradient descent on an entropic smoothing of OT distances is used and [20] which is based on a dual formulation and tools from non-smooth optimization and computational geometry.

This barycenter problem can be extended to more complicated variational problem, such as for instance the Wasserstein propagation [57]. Note that similar entropic regularization technics can be applied as well to this problem.

Multi-marginal transport. The OT barycenter problem, as introduced in [2], is essentially equivalent to a multi-marginal optimal transport with quadratic cost as studied in [36]. Multimarginal reformulations are not usually not tractable, since they involve an optimization problem whose size grows exponentially with respect to the number of marginals. Fortunately, the special structure of the OT barycenter problem leads to linear reformulations that are linear in the number of marginals, see [20] and Section 3.2 below. There are however applications where the problem under study intrinsically has a multi-marginal structure, that cannot be factorized as barycenter computations.

Multi-marginal Optimal Transport, [48, 47], is a natural extension of Optimal Transport with many potential fields of applications : Economics [22] [19], Density Functionnal Theory in Quantum Chemistry [24]. The first important instance of multi-marginal transport was probably Brenier's generalised solutions of the Euler equations for incompressible fluids [13, 14, 15] which are clearly described in his review paper [16]. Note that entropic regularization of the multimarginal transport problem leads to a problem of multi-dimensional matrix scaling [33, 3, 49].

1.2 Contributions

In this paper, we present a unified framework to numerically solve entropic approximations of several generalized optimal transport problems. This framework corresponds to defining appropriate entropic penalizations of the initial linear programs. The key idea is then to interpret the corresponding problems as projections of some input Gibbs density on an intersection of convex sets according to the Kullback-Leibler divergence. This problem can then be solved efficiently using either Bregman iterative projection (for intersection of affine spaces) or a more general Dykstra-like algorithm—these being well-known first order non-smooth optimization schemes. We investigate in details the applications of these ideas to several generalized OT problems: barycenter (Section 3.2), tomographic reconstruction (Section 3.3), multi-marginal transport (Section 4), partial transport (Section 5.1) and capacity constrained transport (Section 5.2). The code implementing the methods presented in this article can be found online[1].

1.3 Computational Speed

The goal of this paper is to present a new class of efficient methods to provide approximate solutions to linear program generalizing OT. It is however important to realize that these methods become numerically unstable when the

[1] `https://github.com/gpeyre/2014-SISC-BregmanOT`

regularization parameter (denoted ε in the following) is small for two reasons: *(i)* since some of the quantities manipulated in the proposed algorithms have an order of $e^{-1/\varepsilon}$ (in particular Gibbs distributions denoted ξ in the following) they become smaller than machine precision whenever the regularization ε is small; *(ii)* more importantly, even if the first issue is taken care of by carrying out computations in the log domain, the convergence speed of iterative projection methods degrades significantly as $\varepsilon \to 0$. We observe therefore that these methods are competitive in a range where the regularization term ε cannot be too small, and for which computed solutions exhibit a small amount of smoothing, see for instance Figure 1 for a visual illustration of this phenomenon. It is thus not the purpose of this article to compare these new methods with more traditional ones (such as interior points or simplex), because they do not target the same problem. Let us however single out the work of [27], that solves the regularized barycenter problem described in Section 3.2 using a gradient descent scheme. In all our numerical experiments, we found however that the iterative Bregman projection converge with substantially computational effort than this gradient descent and, because they rely on alternate projections, do not require adjusting gradient step-sizes and are thus easier to deploy.

1.4 Notations

We denote the simplex in $\mathbb{R}^N$

$$\Sigma_N \stackrel{\text{def.}}{=} \left\{ p \in \mathbb{R}_+^N \;;\; \sum_i p_i = 1 \right\}.$$

The polytope of couplings between $(p,q) \in \Sigma_N^2$ is defined as

$$\Pi(p,q) \stackrel{\text{def.}}{=} \left\{ \gamma \in \mathbb{R}_+^{N \times N} \;;\; \gamma \mathbb{1} = p, \gamma^T \mathbb{1} = q \right\},$$

where γ^T is the transpose of γ and $\mathbb{1} \stackrel{\text{def.}}{=} (1, \ldots, 1)^T \in \mathbb{R}^N$.

For a set $\mathcal{C}$, we denote $\iota_{\mathcal{C}}$ its indicator, that is

$$\forall\, x, \quad \iota_{\mathcal{C}}(x) = \begin{cases} 0 & \text{if} \quad x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

For $\gamma \in \mathbb{R}^{N \times N}$ for some $N > 0$, we define its entropy as

$$E(\gamma) \stackrel{\text{def.}}{=} -\sum_{i,j=1}^N \gamma_{i,j}(\log(\gamma_{i,j}) - 1) + \iota_{\mathbb{R}^+}(\gamma_{i,j}),$$

which is a concave function, where we used the convention $0 \log(0) = 0$.

The Kullback-Leibler divergence between $\gamma \in \mathbb{R}_+^{N \times N}$ and $\xi \in \mathbb{R}_{++}^{N \times N}$ (i.e. $\xi_{i,j} > 0$ for all (i,j)) is

$$\mathrm{KL}\,(\gamma|\xi) \stackrel{\text{def.}}{=} \sum_{i,j=1}^N \gamma_{i,j} \left(\log\left(\frac{\gamma_{i,j}}{\xi_{i,j}}\right) - 1 \right).$$

With a slight abuse of notation, we extend these definitions for higher d-dimensional tensor arrays by replacing the sum over indices (i,j) by sums over higher dimension indices.

Given a convex set $\mathcal{C} \subset \mathbb{R}^{N\times N}$, the projection according to the Kullback-Leibler divergence is defined as

$$P_{\mathcal{C}}^{\mathrm{KL}}(\xi) \stackrel{\text{def.}}{=} \operatorname*{argmin}_{\gamma\in\mathcal{C}} \mathrm{KL}\left(\gamma|\xi\right).$$

For vectors $(a,b) \in \mathbb{R}^N \times \mathbb{R}^N$, we denote entry-wise multiplication and division

$$a \odot b \stackrel{\text{def.}}{=} (a_i b_i)_i \in \mathbb{R}^N \quad \text{and} \quad \frac{a}{b} \stackrel{\text{def.}}{=} (a_i/b_i)_i \in \mathbb{R}^N. \tag{1}$$

2 Iterative Bregman Projections and Dykstra Algorithm

In this paper, we focus on regularized generalized OT problems that can be re-cast in the form

$$\min_{\gamma\in\mathcal{C}} \mathrm{KL}\left(\gamma|\xi\right) \tag{2}$$

where ξ is a given point in $\mathbb{R}_+^{N\times N}$, and $\mathcal{C}$ is an intersection of closed convex sets

$$\mathcal{C} = \bigcap_{\ell=1}^{L} \mathcal{C}_\ell$$

such that $\mathcal{C}$ has nonempty intersection with $\mathbb{R}_+^{N\times N}$.

In the following, we extend the indexing of the sets by L-periodicity, so that they satisfy

$$\forall\, n \in \mathbb{N}, \quad \mathcal{C}_{n+L} = \mathcal{C}_n.$$

2.1 Iterative Bregman Projections

In the special case where the convex sets $\mathcal{C}_\ell$ are affine subspaces (note that nonnegativity constraints are already in the definition of the entropy), it is possible to solve (2) by simply using iterative KL projections. Starting from $\gamma^{(0)} = \xi$, one computes

$$\forall\, n > 0, \quad \gamma^{(n)} \stackrel{\text{def.}}{=} P_{\mathcal{C}_n}^{\mathrm{KL}}(\gamma^{(n-1)}). \tag{3}$$

One can then show that $\gamma^{(n)}$ converges towards the unique solution of (2),

$$\gamma^{(n)} \to P_{\mathcal{C}}^{\mathrm{KL}}(\xi) \quad \text{as } n \to \infty.$$

see [12].

2.2 Dykstra's Algorithm

When the convex sets $\mathcal{C}_\ell$ are not affine subspaces, iterative Bregman projections do not converge in general to the KL projection on the intersection. In contrast, Dykstra's algorithm [29], extended to the KL setting, does converge to the projection, see [6].

Dykstra's algorithm starts by initializing

$$\gamma^{(0)} \stackrel{\text{def.}}{=} \xi \quad \text{and} \quad q^{(0)} = q^{(-1)} = \cdots = q^{(-L+1)} \stackrel{\text{def.}}{=} \mathbb{1}.$$

One then iteratively defines

$$\gamma^{(n)} \stackrel{\text{def.}}{=} P^{\mathrm{KL}}_{\mathcal{C}_n}(\gamma_{n-1} \odot q_{n-L}), \quad \text{and} \quad q^{(n)} \stackrel{\text{def.}}{=} q^{(n-L)} \odot \frac{\gamma^{(n-1)}}{\gamma^{(n)}}. \tag{4}$$

Recall here that $\odot$ and $\frac{\cdot}{\cdot}$ denotes entry-wise operations, see (1).

Dykstra algorithm converges to the solution of (2)

$$\gamma^{(n)} \to P^{\mathrm{KL}}_{\mathcal{C}}(\xi) \quad \text{as } n \to \infty,$$

see [6].

3 Entropic Regularization of Transport-like Problems

3.1 Entropic Optimal Transport Regularization

To illustrate the class of methods developed in this paper, we first review a classical approach to optimal transport approximation, that we recast in the language of Kullback-Leibler projections. This allows us to recover well known results, but in a framework that is easily generalizable.

Following many previous works (see Section 1.1 for details) we consider the following discrete regularized transport

$$W_\varepsilon(p,q) \stackrel{\text{def.}}{=} \min_{\gamma \in \Pi(p,q)} \langle C, \gamma \rangle - \varepsilon E(\gamma). \tag{5}$$

The intuition underlying this regularization is that it enforces the optimal coupling $\gamma^\star_\varepsilon$ solution of (5) to be smoother as ε increases. This regularization also yields favorable computational properties since problem (5) is ε-strongly convex. Its unique solution $\gamma^\star_\varepsilon$ can be obtained through elementary operations (matrix products, elementwise operations on matrices and vectors) as detailed below. If the optimal solution $\gamma^\star$ of the (original, non-regularized, i.e. $\varepsilon = 0$) optimal transport problem is unique, then the optimal solution $\gamma^\star_\varepsilon$ of (5) converges to $\gamma^\star$ as $\varepsilon \to 0$. When other optimal solutions exist, $\gamma^\star_\varepsilon$ converges as $\varepsilon \to 0$ to that with the largest entropy among those, again denoted $\gamma^\star$. The convergence of minimizers of the regularized problem as $\varepsilon \to 0$ is actually exponential

$$\|\gamma^\star_\varepsilon - \gamma^\star\|_{\mathbb{R}^{N \times N}} \leq M e^{-\lambda/\varepsilon}$$

where λ and M depend on C, p, q and N as shown by Cominetti and San Martin [23].

Problem (5) can be re-written as a projection

$$W_\varepsilon(p,q) = \varepsilon \min_{\gamma \in \Pi(p,q)} \mathrm{KL}\left(\gamma | \xi\right) \quad \text{where} \quad \xi = e^{-\frac{C}{\varepsilon}} \tag{6}$$

of ξ according to the Kullback-Leibler divergence (here the exponential is computed component-wise).

Problem (6) can in turn be formulated as (2) with $L = 2$ affine subsets of $\mathbb{R}_+^{N\times N}$

$$\mathcal{C}_1 \overset{\text{def.}}{=} \left\{\gamma \in \mathbb{R}_+^{N\times N} \;;\; \gamma \mathbb{1} = p\right\} \quad \text{and} \quad \mathcal{C}_2 \overset{\text{def.}}{=} \left\{\gamma \in \mathbb{R}_+^{N\times N} \;;\; \gamma^T \mathbb{1} = q\right\}.$$

The application of Bregman iterative projection (detailed in Section 2.1) to this splitting corresponds to the so-called IPFP/Sinkhorn algorithm (see Section 1.1 for bibliographical details).

The following well-known proposition details how to compute the relevant projections.

Proposition 1. *One has, for $\bar{\gamma} \in (\mathbb{R}_+)^{N\times N}$,*

$$P_{\mathcal{C}_1}^{\mathrm{KL}}(\bar{\gamma}) = \mathrm{diag}\left(\frac{p}{\bar{\gamma}\mathbb{1}}\right)\bar{\gamma} \quad \textit{and} \quad P_{\mathcal{C}_2}^{\mathrm{KL}}(\bar{\gamma}) = \bar{\gamma}\,\mathrm{diag}\left(\frac{q}{\bar{\gamma}^T\mathbb{1}}\right) \tag{7}$$

In plain words, the two projections in equation (7) normalize (with a multiplicative update) either the rows or columns of $\bar{\gamma}$ so that they have the desired row-marginal p or column-marginal q.

Remark 1 (Fast implementation). An important feature of iterations (3), when combined with projections (7) is that the iterates $\gamma^{(n)}$ satisfy

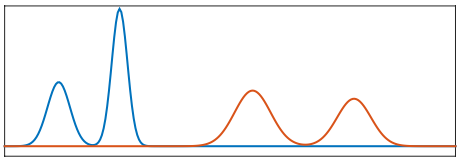
$$\gamma^{(n)} = \mathrm{diag}(u^{(n)})\,\xi\,\mathrm{diag}(v^{(n)})$$

where the vectors $(u^{(n)}, v^{(n)}) \in \mathbb{R}^N \times \mathbb{R}^N$ satisfy $v^{(0)} = \mathbb{1}$ and obey the recursion formula

$$u^{(n)} = \frac{p}{\xi v^{(n)}} \quad \text{and} \quad v^{(n+1)} = \frac{q}{\xi^T u^{(n)}}.$$

This allows to implement this algorithm by only performing matrix-vector multiplications using a fixed matrix ξ, possibly in parallel if several OT are to be computed for several marginals sharing the same ground cost C as shown in [26].

Figure 1 displays examples of transport maps $\gamma = \gamma_\varepsilon^\star$ solving (6), for two 1-D marginals $(p,q) \in \mathbb{R}^N \times \mathbb{R}^N$ discretizing continuous densities on a uniform grid $(x_i)_{i=1}^N$ of $[0,1]$, and with a ground cost $C_{i,j} = \|x_i - x_j\|^2$. The computation is performed with $N = 256$. This figure shows how $\gamma_\varepsilon^\star$ converges towards a solution of the original un-regularized transport as $\varepsilon \to 0$. It also shows how the iterates of the algorithm $\gamma^{(n)}$ progressively shift mass away from the diagonal during the iterations.



Marginals p and q

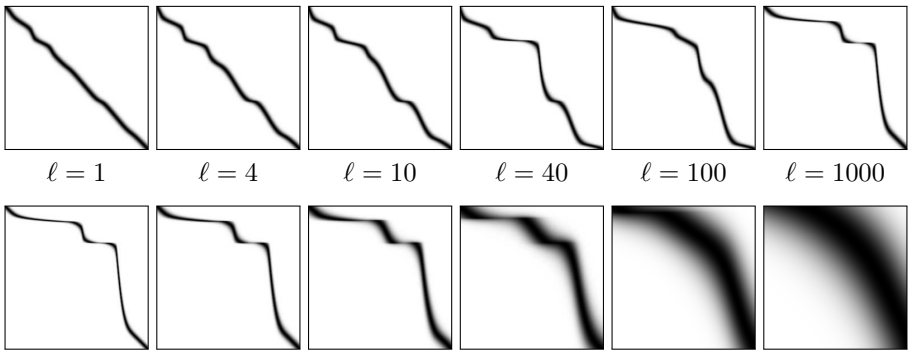


Figure 1: *Top:* the input densities p (blue curve) and q (red curve). *Center:* evolution of the couplings $\gamma^{(\ell)}$ at iteration ℓ of the Sinkhorn algorithm. *Bottom:* solution $\gamma = \gamma_\varepsilon^\star$ of (6) for several values of ε.

3.2 Optimal Transport Barycenters

We are given a set $(p_k)_{k=1}^K$ of input marginals $p_k \in \Sigma_N$, and we wish to compute a weighted barycenter according to the Wasserstein metric. This problem finds many applications, as highlighted in Section 1.1.

Following [2], the general idea is to define the barycenter as a solution of a variational problem mimicking the definition of barycenters in Euclidean spaces. Given a set of normalized weights $\lambda \in \Sigma_K$, we consider the problem

$$\min_{p \in \Sigma_N} \left\{ \sum_{k=1}^K \lambda_k W_\varepsilon(p_k, p) \;;\; p \in \Sigma_N \right\} \tag{8}$$

which as in [20] is re-written as

$$\forall\, k = 1, \ldots, K, \quad p = \gamma_k \mathbb{1}$$

where the set of optimal couplings $\boldsymbol{\gamma} = (\gamma_k)_{k=1}^K \in (\mathbb{R}_+^{N \times N})^K$ solves

$$\min \left\{ \mathrm{KL}_\lambda\left(\boldsymbol{\gamma} | \boldsymbol{\xi}\right) \overset{\text{def.}}{=} \sum_{k=1}^K \lambda_k \,\mathrm{KL}\left(\gamma_k | \xi_k\right) \;;\; \boldsymbol{\gamma} \in \mathcal{C}_1 \cap \mathcal{C}_2 \right\} \tag{9}$$

$$\text{where} \quad \forall k, \quad \xi_k \overset{\text{def.}}{=} \xi \overset{\text{def.}}{=} e^{-\frac{C}{\varepsilon}}$$

and the constraint sets are defined by

$$\mathcal{C}_1 \stackrel{\text{def.}}{=} \left\{ \boldsymbol{\gamma} = (\gamma_k)_k \in (\Sigma_N)^K \;;\; \forall\, k,\; \gamma_k^T \mathbb{1} = p_k \right\}$$
$$\text{and} \quad \mathcal{C}_2 \stackrel{\text{def.}}{=} \left\{ \boldsymbol{\gamma} = (\gamma_k)_k \in (\Sigma_N)^K \;;\; \exists p \in \mathbb{R}^N, \forall\, k,\; \gamma_k \mathbb{1} = p \right\}$$

It is easy to check that the Bregman iterative projection scheme can be applied to this setting by simply replacing KL by KL_λ.

The KL_λ projection on $\mathcal{C}_1$ is computed as detailed in Proposition 1, since it is equal to the KL projection of each $\xi_k = \xi$ on a constraint of fixed marginal p_k. The KL_λ projection on $\mathcal{C}_2$ is computed as detailed in the following proposition.

Proposition 2. *For $\bar{\boldsymbol{\gamma}} \stackrel{\text{def.}}{=} (\bar{\gamma}_k)_k \in (\mathbb{R}_+^{N \times N})^K$, the projection $\boldsymbol{\gamma} \stackrel{\text{def.}}{=} (\gamma_k)_{k=1}^K = P_{\mathcal{C}_2}^{\mathrm{KL}_\lambda}(\bar{\boldsymbol{\gamma}})$ satisfies*

$$\forall\, k, \quad \gamma_k = \operatorname{diag}\left(\frac{p}{\bar{\gamma}_k \mathbb{1}} \right) \bar{\gamma}_k \quad \textit{where} \quad p \stackrel{\text{def.}}{=} \prod_{r=1}^K (\bar{\gamma}_r \mathbb{1})^{\lambda_r} \tag{10}$$

where $\prod$ and $(\cdot)^{\lambda_r}$ should be understood as entry-wise operators.

Proof. Introducing the variable p such that for all k, $\gamma_k \mathbb{1} = p$, the first order conditions of the projection $P_{\mathcal{C}_2}^{\mathrm{KL}_\lambda}(\bar{\boldsymbol{\gamma}})$ states the existence of Lagrange multipliers $(u_k)_k$ such that

$$\forall\, k, \quad \lambda_k \log\left(\frac{\gamma_k}{\bar{\gamma}_k} \right) + u_k \mathbb{1}^T = 0 \quad \text{and} \quad \sum_r u_r = 0.$$

Denoting $a_k = e^{-u_k}$, one has $\prod_k a_k = \mathbb{1}$ and $\gamma_k = \operatorname{diag}(a_k^{1/\lambda_k}) \bar{\gamma}_k$. Condition $\gamma_k \mathbb{1} = p$ thus implies that

$$a_k = \left(\frac{p}{\bar{\gamma}_k \mathbb{1}} \right)^{\lambda_k},$$

and condition $\prod_k a_k = \mathbb{1}$ gives the desired value (10) for p. □

Remark 2 (Special case). Note that when $K = 2$, $(\lambda_1, \lambda_2) = (0, 1)$, one retrieves exactly the IPFP/Sinkhorn algorithm to solve the entropic OT, as detailed in Section 3.1. Our novel scheme to compute barycenters should thus be understood as the natural generalization of this IPFP algorithm to barycenters.

Remark 3 (Memory efficient and parallel implementation). Similarly as for Remark 1, one verifies that iterations (3) in the special case of problem (9) leads to iterates $\boldsymbol{\gamma}^{(n)} = (\gamma_k^{(n)})_k$ which satisfy, for each k

$$\gamma_k^{(n)} = \operatorname{diag}(u_k^{(n)}) \xi \operatorname{diag}(v_k^{(n)})$$

for two vectors $(u_k^{(n)}, v_k^{(n)}) \in \mathbb{R}^N \times \mathbb{R}^N$ initialized as $v_k^{(0)} = \mathbb{1}$ for all k, and computed with the iterations

$$u_k^{(n)} = \frac{p^{(n)}}{\xi v_k^{(n)}} \quad \text{and} \quad v_k^{(n+1)} = \frac{p_k}{\xi^T u_k^{(n)}}$$

where $p^{(n)}$ is the current estimate of the barycenter, computed as

$$p^{(n)} = \prod_{k=1}^{N} \left(u_k^{(n)} \odot (\xi v_k^{(n)}) \right)^{\lambda_k}.$$

A nice feature of these iterations is that they can be computed in parallel for all k using multiplications between the matrix ξ and matrices storing $(u_k^{(n)})_k$ and $(v_k^{(n)})_k$ as columns.

Figure 2 shows an example of barycenters computation for $K = 3$. The three vertices of the triangle show the input densities (p_1, p_2, p_3) which are uniform on binary shapes (diamond, annulus and square). The other points in the triangle display the results for the following values of λ

(0,0,1)
(1, 0, 3)/4 (0, 1, 3)/4
(1,0,1)/2 (1,1,2)/4 (0,1,1)/2
(3,0,1)/4 (2,1,1)/4 (1,2,1)/4 (0,3,1)/4
(1,0,0) (3,1,0)/4 (1,1,0)/2 (1,3,0)/4 (0,1,0)

The computation is performed on an uniform 2-D grid of $N = 256 \times 256$ points in $[0, 1]^2$, and $\varepsilon = 2/N$.

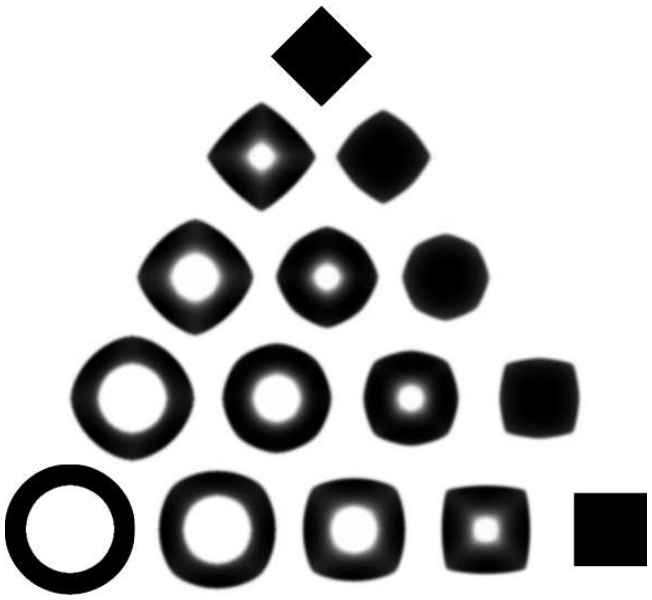


Figure 2: Example of OT barycenters with entropic smoothing.

3.3 Partial Radon Inversion with OT Fidelity

The partial Radon transform (i.e. the computation of integrals of the data along parallel rays in a small limited set of directions) is a mathematical model for several scanning medical acquisition devices. This is an ill-posed linear operator, and inverting it while preventing noise and artifacts to blowup is of utmost importance for the targeted imaging applications. It is out of the scope of this paper to review the overwhelming literature on the topic of Radon inversion, and we refer to the book [37] for an overview of classical approaches, and [41] and the references therein for examples of state-of-the art methods.

The goal of this section is not to present a state of the art inversion scheme, but rather to show how the method recently introduced by [1] can be solved using a simple iterative Bregman projection algorithm. We describe here the method in a fully discretized setting, where the Radon transform is implemented using a nearest neighbor interpolation.

We consider a square discretization grid of $N = N_0 \times N_0$ pixels, indexed with

$$s = (s_1, s_2) \in \Omega_N \overset{\text{def.}}{=} \omega_{N_0} \times \omega_{N_0} \quad \text{where} \quad \omega_{N_0} \overset{\text{def.}}{=} \{1, \ldots, N_0\}^2.$$

Given an angle θ, we consider the following discrete lines in Ω_N, $\forall\, (s_1, s_2) \in \Omega_N$,

$$\ell^\theta_{s_1,s_2} \overset{\text{def.}}{=} \begin{cases} (s_2, s_1 + (s_2 - 1)\tan(\theta) \bmod N_0), & \text{if } \mathrm{mod}(\theta, \gamma) \in [0, \gamma/4] \cup [3\gamma/4, \gamma] \\ (s_1 + (s_2 - 1)\tan(\theta) \bmod N_0, s_2), & \text{if } \mathrm{mod}(\theta, \gamma) \in [\gamma/4, 3\gamma/4] \end{cases}$$

where the mod N_0 is a modulo N_0 that maps the indices in the admissible range Ω_N, hereby effectively implementing a convenient cyclic boundary condition.

The discrete Radon integration $R_\theta(f) \in \mathbb{R}^{N_0}$ for such an angle θ of an image $f \in \mathbb{R}^N \sim \mathbb{R}^{N_0 \times N_0}$ computes

$$\forall\, s_1 \in \omega_{N_0}, \quad R_\theta(f)_{s_1} \overset{\text{def.}}{=} \sum_{s_2=1}^{N_0} f_{\ell_{s_1,s_2}} \in \mathbb{R}.$$

Its adjoint R^*_θ, the back-propagation operator along direction θ, reads, for $r \in \mathbb{R}^{N_0}$,

$$f = R^*_\theta(r) \quad \text{where} \quad \forall\, s \in \Omega_N, \quad f_{\ell_{s_1,s_2}} = r_{s_1}.$$

We consider the linear inverse problem of reconstructing an approximation of an unknown image $f_0 \in \mathbb{R}^N$ from (possibly noisy) partial Radon measurements

$$r = (r_k)_{k=1}^K \quad \text{where} \quad r_k = R_{\theta_k}(f_0) + w_k, \tag{11}$$

where $(\theta_k)_{k=1}^K$ is a set of angles, and w_k is the noise perturbing the observation. We denote $R(f) = (R_{\theta_k}(f))_{k=1}^K$ so that $R : \mathbb{R}^N \mapsto \mathbb{R}^{N_0 \times K}$ is a linear operator, and R^* is its adjoint.

The simplest way to perform a reconstructiong is to solve for the least squares estimate

$$R^+(r) = R^*(RR^*)^{-1} r \overset{\text{def.}}{=} \operatorname*{argmin}_f \|f\| \quad \text{s.t.} \quad \forall\, k, \quad R_{\theta_k}(f) = r_k. \tag{12}$$

This linear inverse does a poor job in the case of a small number K of projections, since it exhibits reconstruction artifacts, as shown on Figure 3.

In order to obtain a better reconstruction, and following the idea introduced in [1], we suppose that one has access to a template $g_0 \in \mathbb{R}^N$ which is intended to be some approximation of f_0, possibly with some translation and small deformations. To leverage the strong robustness of the OT distance with respect to translation and small deformations, the reconstruction is obtained by using a sum of Wasserstein distances to the observation and to the template

$$\min_{f \in \Sigma_N} \lambda_1 W_{2,\varepsilon}(f, g_0) + \lambda_2 \sum_{k=1}^{Q} W_{1,\varepsilon}(R_{\theta_k}(f), r_k), \tag{13}$$

where $0 < \lambda_1 \leqslant 1$ is a weight that accounts for the degree of confidence in the template g_0, and $\lambda_2 = 1 - \lambda_1$.

Here, $W_{2,\varepsilon}$ indicates the entropic Wasserstein distance (6) on the 2-D grid Ω_N where we defined the cost matrix $C = C^2$ as

$$\forall\, (s,t) \in (\Omega_N)^2, \quad C^2_{s,t} \overset{\text{def.}}{=} (s_1 - t_1)^2 + (s_2 - t_2)^2$$

and $W_{1,\varepsilon}$ indicate the entropic Wasserstein distance (6) on a 1-D periodic grid for the cost $C = C^1$

$$\forall\, (i,j) \in (\omega_{N_0})^2, \quad C^1_{i,j} \overset{\text{def.}}{=} \min_{k \in \mathbb{Z}} (i - j + kN_0)^2.$$

Similarly to the barycenter problem (9), we compute f solving (13) as $f = \gamma \mathbb{1}_N$ where γ solves

$$\min \left\{ \lambda_1 \operatorname{KL}(\gamma|\xi) + \lambda_2 \sum_{k=1}^{K} \operatorname{KL}(\gamma_k|\xi_k) \;;\; \forall\, k,\ (\gamma, \gamma_k) \in \mathcal{C}_k, \cap \tilde{\mathcal{C}}_k \right\} \tag{14}$$

$$\text{where} \quad \xi = e^{-\frac{C^2}{\varepsilon}} \quad \text{and} \quad \forall\, k, \quad \xi_k = e^{-\frac{C^1}{\varepsilon}}$$

(here the exp should be understood component-wise) and where we introduced

$$\begin{aligned}
&\forall\, k,\ \mathcal{C}_k = \left\{ (\gamma, \gamma_k) \in \Sigma_N \times \Sigma_{N_0} \;;\; \gamma^T \mathbb{1}_N = g_0 \quad \text{and} \quad \gamma_k^* \mathbb{1}_{N_0} = r_k \right\} \\
&\forall\, k,\ \tilde{\mathcal{C}}_k = \left\{ (\gamma, \gamma_k) \in \Sigma_N \times \Sigma_{N_0} \;;\; R_{\theta_k}(\gamma \mathbb{1}_N) = \gamma_k \mathbb{1}_{N_0} \right\}.
\end{aligned}$$

Problem (14) thus corresponds to a (weighted) KL projection on the intersection of $2K$ constraints. Computing the projection on each $\mathcal{C}_k$ is achieved as detailed in Proposition 1. The following proposition, which is a simple extension of Proposition 2, shows how to project on $\tilde{\mathcal{C}}_k$.

Proposition 3. *For any $k = 1, \ldots, K$, the projection $\boldsymbol{\gamma} = (\gamma, \gamma_k) = P^{\mathrm{KL}_\lambda}_{\tilde{\mathcal{C}}_k}(\bar{\boldsymbol{\gamma}})$ of $\bar{\boldsymbol{\gamma}} = (\bar{\gamma}, \bar{\gamma}_k)$ for the KL metric*

$$\mathrm{KL}_\lambda(\boldsymbol{\gamma}|\bar{\boldsymbol{\gamma}}) \overset{\text{def.}}{=} \lambda_1 \operatorname{KL}(\gamma|\bar{\gamma}) + \lambda_2 \operatorname{KL}(\gamma_k|\bar{\gamma}_k)$$

satisfies

$$\gamma = \mathrm{diag}\left(R^*_{\theta_k}\left(\frac{\delta_k}{\alpha_k}\right)\right)\bar{\gamma} \quad \textit{and} \quad \gamma_k = \mathrm{diag}\left(\frac{\delta_k}{\beta_k}\right)\bar{\gamma}_k$$

where we defined

$$\alpha_k \overset{\text{def.}}{=} R_{\theta_k}(\bar{\gamma}\mathbb{1}_N), \quad \beta_k \overset{\text{def.}}{=} \bar{\gamma}_k\mathbb{1}_{N_0}, \quad \textit{and} \quad \delta_k = \alpha_k^{\lambda_1} \odot \beta_k^{\lambda_2}$$

where the exponentiations are component-wise.

Figure 3 shows an example of application of the method for an image f_0 which is discretized on a grid of $N = 80 \times 80$ points, using $\varepsilon = 2/N$ and $K = 12$ Radon directions. There is no additional noise in the measurements, i.e. $w_k = 0$ in (11). The template g_0 is a binary disk. These results show how using a large λ_1 recovers a result that is close to g_0, while using a smaller λ introduces the geometric features of f_0 but also reconstruction artifacts. Note that the linear reconstruction $R^+(r)$ contains reconstruction artifacts.

4 Multi-marginal Optimal Transport

Multi-marginal optimal transport is a natural extension of optimal transport with many potential fields of applications, see Section 1.1.

4.1 Multi-marginal Regularization

As in the barycenter problem, we are given K marginals $(p_k)_{k=1}^K$. We denote $\gamma \in \mathbb{R}_+^{N^K}$ a K-dimensional array, indexed as γ_j for $j = (j_1, \ldots, j_K) \in \{1 \ldots, N\}^K$. The push-forward $S_k(\gamma) \in \mathbb{R}^N$ of such a γ along dimension k is computed as

$$\forall\, i \in \{1, \ldots, N\}, \quad S_k(\gamma)_i \overset{\text{def.}}{=} \sum_{j_1, j_2, \ldots, j_{k-1}, j_{k+1}, \ldots, j_K} \gamma_{j_1, j_2, \ldots, j_{k-1}, i, j_{k+1}, \ldots, j_K}.$$

The set of couplings between the marginals is

$$\Pi(p_1, p_2, \ldots, p_k) \overset{\text{def.}}{=} \left\{\gamma \in \mathbb{R}_+^{N^K} \;;\; \forall k, \quad S_k(\gamma) = p_k\right\}.$$

Given a cost matrix $C \in \mathbb{R}_+^{N^K}$, the regularized OT problem (5) is generalized to this multi-marginal setting as

$$\min_{\gamma \in \Pi(p,q)} \langle C,\, \gamma\rangle - \varepsilon E(\gamma). \tag{15}$$

Similarly as (6), this problem can be re-cast as a KL projection

$$\min_{\gamma} \left\{\mathrm{KL}\left(\gamma | \xi\right) \;;\; \gamma \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \ldots \cap \mathcal{C}_K\right\} \quad \text{where} \quad \xi \overset{\text{def.}}{=} e^{-\frac{C}{\varepsilon}} \tag{16}$$

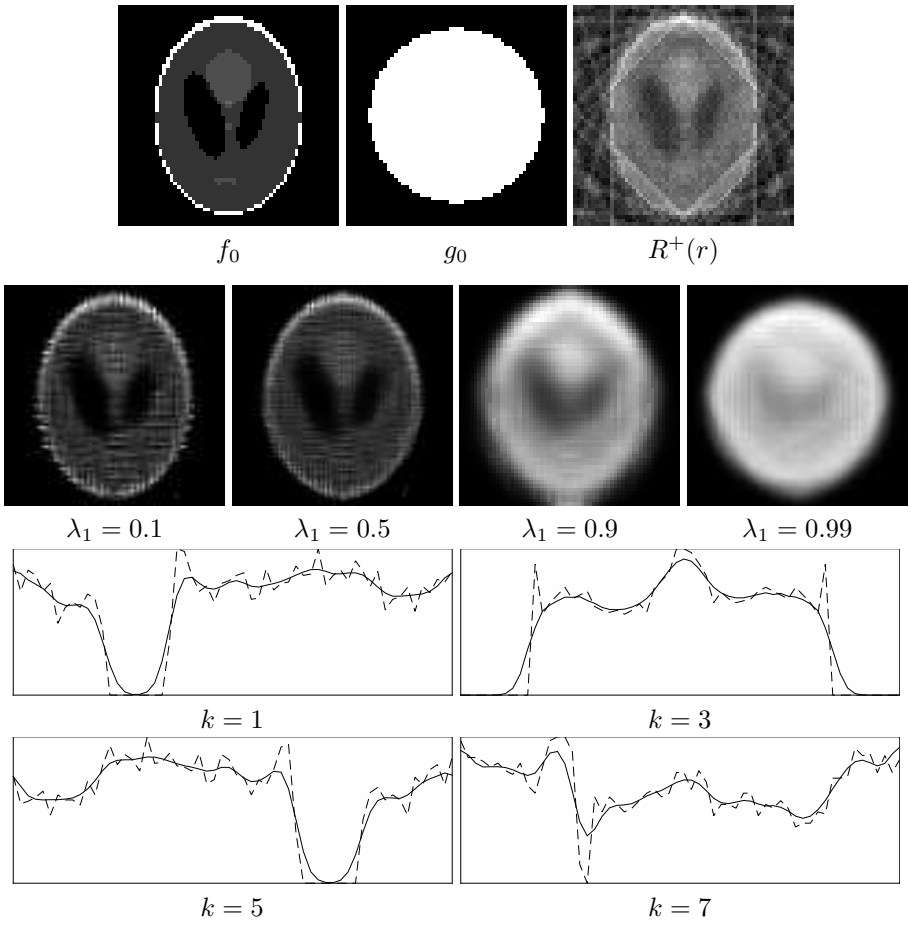


Figure 3: Example of reconstructions from partial Radon measurements. The two bottom rows show the input transforms $r_k = R_{\theta_k}(f_0)$ (dashed curves) and the recovered Radon transforms $R_{\theta_k}(f_0)$ where f solved 13 (plain curves), for a few values of k.

where the exponentiation is exponent-wise, and where

$$\mathcal{C}_k \stackrel{\text{def.}}{=} \left\{ \gamma \in \mathbb{R}_+^{N^K} \;;\; S_k(\gamma) = p_k \right\}.$$

The Bregman projection on each of the convex $\mathcal{C}_k$ are again given by a simple normalisation as detailed in the following proposition.

Proposition 4. *For any k, denoting $\gamma = P^{\mathrm{KL}}_{\mathcal{C}_k}(\bar{\gamma})$, one has*

$$\forall\, j = (j_1, \ldots, j_K), \quad \gamma_j = \frac{(p_k)_{j_k}}{S_k(\bar{\gamma})_{j_k}} \bar{\gamma}_j$$

It is thus possible to use the Bregman iterative projection detailed in Section 2.1 to compute the projection (16).

We now detail in the two following sections two typical cases of application of multi-marginal OT: grid-free barycenter computation (Section 4.2) and resolution of generalized Euler flow (Section 4.3).

4.2 Multi-marginal Barycenters

As shown in [2], the computation of barycenters of measures (thus the continuous analogous of (3.2)) can be computed by solving a multi-marginal transport problem.

Let us suppose that the input measures $(\mu_k)_{k=1}^K$ defined on $\mathbb{R}^d$ are of the form $\mu_k = \sum_{i=1}^N p_{k,i}\delta_{x_i}$, where $p_k = (p_{k,i})_{i=1}^N \in \Sigma_N$, where $\{x_i\}_i \subset \mathbb{R}^d$ and δ_x is the Dirac measure at location $x \in \mathbb{R}^d$. It is shown in [2] that the Wasserstein barycenter of the μ_k with weights $(\lambda_k)_k \in \Sigma_K$ for the quadratic Euclidean distance ground cost is

$$\mu_\lambda \overset{\text{def.}}{=} \sum_{j=(j_1,j_2,\ldots,j_K)} \gamma_j \delta_{A_j(x)} \tag{17}$$

where $A_j(x) \overset{\text{def.}}{=} \sum_k \lambda_k x_{j_k}$ is the Euclidean barycenter and $\gamma \in \mathbb{R}_+^{N^K}$ is an optimal multi-marginal coupling that solves (15) for the following cost

$$C_j = \sum_{1 \leqslant k \leqslant K} \frac{\lambda_k}{2} \|x_{j_k} - A_j(x)\|^2.$$

An important point to note is that the measure barycenter (17) is in general composed of more than N Diracs, and that these Diracs are not constrained to be on the discretization grid $(x_i)_i$. In particular, the obtained result is different from the one obtained by solving (8), which computes a barycenter that lies on the same grid as the input measures. In some sense, formulation (17) is able to compute the "true" barycenter of measures, whereas (8) computes an approximation on a fixed grid, but the price to pay is the resolution of a high-dimensional multi-marginal program.

Figure 4 shows an histogram depiction of the measure μ_λ defined in (17), for the iso-barycenter (i.e. $\lambda_k = 1/K$ for all k). It is computed by first solving (15) with the same three marginals used in Figure 2. The histogram $p \in \Sigma_N$ computed on a grid of $N = 60 \times 60$ points. Each p_i is the total mass of μ_λ in the discretization square S_i of size $1/\sqrt{N} \times 1/\sqrt{N}$, i.e. $p_i = \mu_\lambda(S_i)$.

4.3 Generalized Euler Flows

Brenier proposed in a series of papers [13, 14, 15] a relaxation of the Euler equation of incompressible fluids with constrained initial and final data. These data are conveniently expressed as a volume preserving map Ξ of the domain. This relaxation can be understood as requiring the resolution of a multi-marginal transportation with an infinite number of marginals. Following [16] (equation (21) section VII), when discretizing this problem with K steps in time, one thus faces the resolution of a K marginals OT problem.

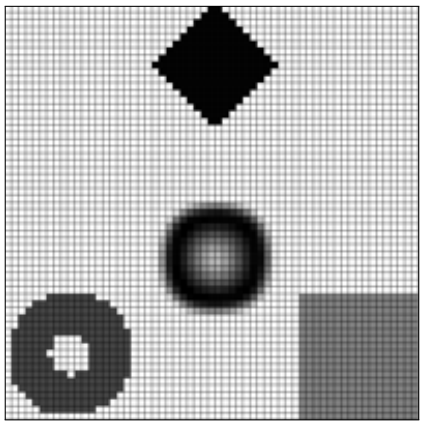


Figure 4: Barycenter computed by solving the multi-marginal problem with three marginals (annulus, diamond and square) discretized on an uniform 2-D grid of $N = 60 \times 60$ points in $[0,1]^2$ and $\varepsilon = 0.005$. See the main text body for details about how the display of the barycenter measure is performed.

We consider a fixed uniform discretization of $[0,1]^d$ with points $(x_i)_{i=1}^N$. The marginals are the uniform measure on this set (as discretization of the Lebesgue measure), i.e. $p_k = \mathbb{1}/N$ for all k. The prescribed volume preserving maps $\Xi : [0,1]^d \to [0,1]^d$ is discretized using a permutation of the grid points, i.e. a discrete bijection $\sigma : \{1,\dots,N\} \to \{1,\dots,N\}$.

The cost function is then

$$C_{j_1,\dots,j_K} = \sum_{k=1,\dots,K-1} \|x_{j_{k+1}} - x_{j_k}\|^2 + \|x_{\sigma(j_1)} - x_{j_K}\|^2$$

and the optimal coupling γ solves (15).

For each $k \in \{1,\dots,K\}$, the transition probability from "time" 1 to "time" k : $T_{1,k} \in \mathbb{R}^{N\times N}$ is defined as

$$\forall\,(s,w) \in \{1,\dots,N\}^2, \quad (T_{1,k})_{s,w} = \sum_{j_i \neq j_1, j_k} \gamma_{s,j_2,\dots,j_{k-1},w,j_{k+1},\dots,j_K}. \tag{18}$$

It represents the evolution of a generalized flow of particles at time $t = \frac{k-1}{K-1}$. Note that in this setting, particles trajectories are non deterministic and their mass may split and spread across the domain.

Brenier's numerical method [16] is based on an approximation of the measure preserving map by a one to one permutation of the domain and the representation of the diffuse coupling therefore needs a large number of particles. Our resolution method is different and computes a space discretization of the coupling matrix and naturally encodes non-diffeomorphic volume preserving maps. The coupling γ is an array of size $(N^d)^K$ where the d-dimensional physical domain is discretized on N^d points and we have K times steps. However, as explained in Remark 4 below, because of the structure of the cost we only need to store and multiply $(N^d)^2$ matrices.

Remark 4. [Reduction to Transitions probabilities] As defined in (16), the resolution of the regularized K-marginal OT problem boils down to the computation of a KL projection of $\xi = e^{-\frac{C}{\varepsilon}}$ We can rewrite the coupling using only 2 smaller matrices $\xi^0, \xi^1 \in \mathbb{R}^{N\times N}$ since

$$\xi_{j_1,\dots,j_K} = \left(\prod_{k=1}^{K-1} \xi^0_{j_k,j_{k+1}}\right) \xi^1_{j_K\sigma(j_1)} \tag{19}$$

$$\text{where} \quad \xi^0_{\alpha,\beta} = e^{-\frac{D_{\alpha\beta}}{\varepsilon}}, \qquad \xi^1_{\beta,\alpha} = e^{-\frac{D_{\beta\sigma(\alpha)}}{\varepsilon}}.$$

and $D_{\alpha\beta} = \|x_\alpha - x_\beta\|^2$. We recall that all the marginals are equal to (a discretization of) the Lebesgue measure and $(x_i)_i$ are discretized on the unit cube $[0,1]^d$. As already noticed in Remark 1, the iterative Bregman projections (3) (always on the same Lebesgue marginal constraint) can be simplified as an IPFP iterative procedure. The optimal coupling γ that solves (15) can be actually written as follows

$$\gamma_{j_1,\dots,j_K} = \xi_{j_1,\dots,j_K} \prod_{k=1}^{K} u^k_{j_k},$$

where $u^k_{j_k}$ is the $(j_k)^{\text{th}}$ component of $u^k \in \mathbb{R}^N$, for $k = 1,\dots,K$ (the k^{th} vector being associated to the k^{th} marginal). Moreover the u^k are uniquely determined by the constraints over the marginals of γ

$$u^k_{j_k} = \frac{1/N}{\sum_{m\neq k}\sum_{j_m}\{\xi_{j_1,\dots,j_K}\prod_{\ell\neq k} u^\ell_{j_\ell}\}}$$

and the IPFP procedure for K marginals is

$$u^{k,(n)}_{j_k} = \frac{1/N}{\sum_{m\neq k}\sum_{j_m}\{\xi_{j_1,\dots,j_K}\prod_{\ell\neq k} u^{\ell,(n-1)}_{j_\ell}\}}$$

However, one can notice that the sum in (4) could be computationally onerous, but thanks to (19) we can rearrange it as

$$u^{k,(n)}_{j_k} = \frac{1/N}{B_{j_k,j_k}}$$

where $u^{k,(n)}$ is the k^{th} vector at step (n) and B is the the product of K smaller $N\times N$ matrices

$$B = \xi^{\text{init}} \bigotimes_{\ell=1}^{K-1} \tilde{\xi}_\ell \quad \text{with} \quad \xi^{\text{init}} = \begin{cases} \xi^0 & \text{for} \quad k \neq K, \\ \xi^1 & \text{otherwise} \end{cases}$$

$$\text{and} \quad \tilde{\xi}_\ell = \begin{cases} \text{diag}(u^{\sigma^k(\ell),(n)}) \otimes \xi^0 & \text{for} \quad \sigma^k(\ell) \neq K \quad \text{and} \quad \sigma^k(\ell) < k, \\ \text{diag}(u^{\sigma^k(\ell),(n-1)}) \otimes \xi^0 & \text{for} \quad \sigma^k(\ell) \neq K \quad \text{and} \quad \sigma^k(\ell) > k, \\ \text{diag}(u^{K,(n-1)}) \otimes \xi^1 & \text{otherwise,} \end{cases}$$

where $\otimes$ is the standard matrix product and we use the Matlab convention that diag of vector is the diagonal matrix with vector values and diag of matrix is the vector of its diagonal. We also highlight that, to use the simplification the u^l must be ordered in the correct way so that the computation of the sum for the $k$ieth update starts at $k+1$ and finishes at $k-1$. We have introduced the circular permutations $\sigma^k(\ell) = (\ell+k-1 \mod K)+1$ which returns the $(\sigma^k(\ell))^{\text{th}}$ term at the ℓ^{th} position of the product.

Each iteration of the IPFP procedure therefore only involves $(2K)$ 2-coupling matrices multiplications and only requires storing K vectors and the 2-coupling cost matrices ξ^0 and ξ^1. The computation of the 2-coupling maps (18) can be simplified with the same remark.

Figures 5, 6 and 7 show $T_{1,k}$ for three test cases in dimension $d = 1$ proposed in [16]. The computation is performed with a uniform discretization $(x_i)_i$ of $[0, 1]$ with $N = 200$ points, $\varepsilon = 10^{-3}$ and $K = 16$. They agree with the solutions produced by Brenier and the mass spreading of the generalized flow is nicely captured by the 2 marginals couplings (18).

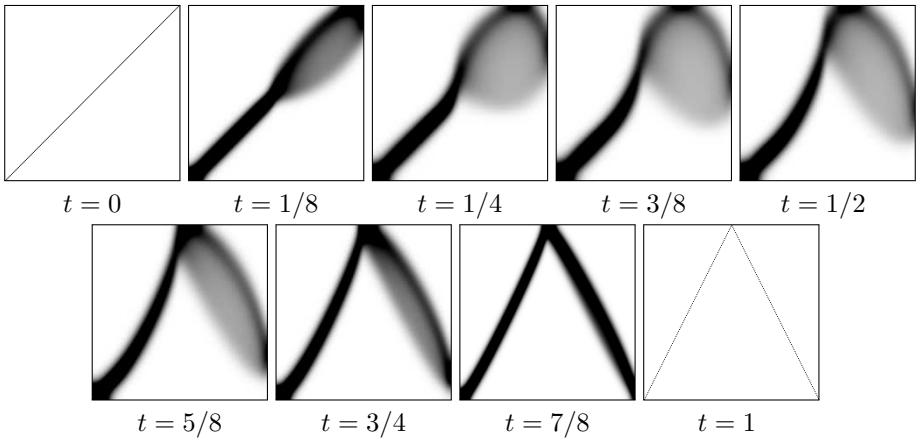


Figure 5: Display of $T_{1,k}$ showing the evolution of the fluid particles from x to $\Xi(x) = \min(2x, 2-2x)$ for $x \in [0, 1]$. The corresponding time is $t = \frac{k-1}{K-1} \in [0, 1]$.

5 Transport Problems with Inequality Constraints

In this section, we consider transport problems with inequality constraints. Again we have to project for the KL divergence on the intersection of convex sets of nonnegative vectors.

5.1 Partial Transport

In the partial transport problem, one is given two marginals $(p, q) \in (\mathbb{R}_+^N)^2$, not necessarily with the same total mass. We wish to transport only a given

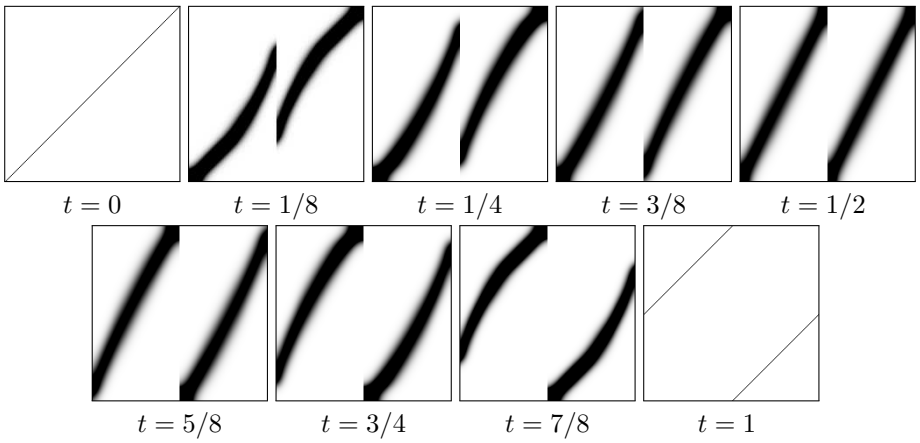


Figure 6: Same as Figure 5 for the map $\Xi(x) = (x + 1/2) \bmod 1$ for $x \in [0, 1]$.

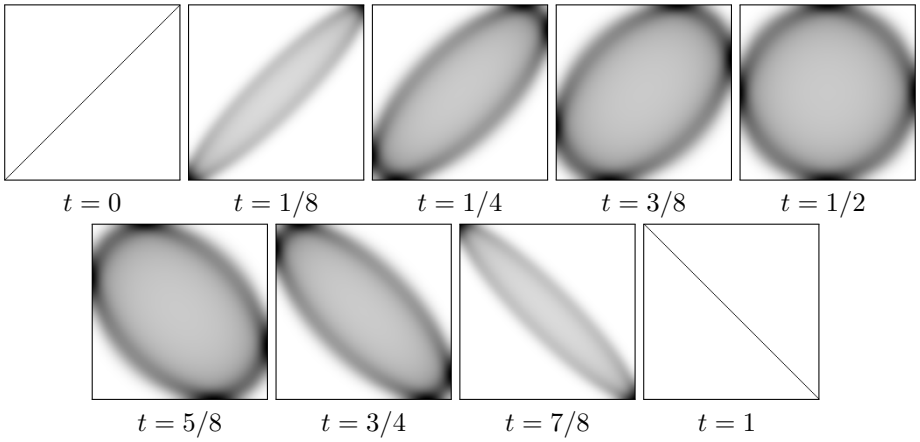


Figure 7: Same as Figure 5 for the map $\Xi(x) = 1 - x$ for $x \in [0, 1]$.

fraction of mass

$$m \in [0, \min(p^T \mathbb{1}, q^T \mathbb{1})],$$

minimizing the transportation cost $\langle C, \gamma \rangle$ where $C \in (\mathbb{R}_+)^{N \times N}$ is the ground cost.

The corresponding regularized problem reads

$$\min_{\gamma \in \mathbb{R}_+^{N \times N}} \left\{ \langle C, \gamma \rangle - \varepsilon E(\gamma) \; ; \; \gamma \mathbb{1} \leqslant p, \; \gamma^T \mathbb{1} \leqslant q, \mathbb{1}^T \gamma \mathbb{1} = m \right\} \tag{20}$$

where the inequalities should be understood component-wise.

Similarly to (6), this is equivalent to computing the projection of $\xi = e^{-\frac{C}{\varepsilon}}$ on the intersection $\mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3$ of $K = 3$ convex sets where

$$\mathcal{C}_1 \stackrel{\text{def.}}{=} \{\gamma \; ; \; \gamma \mathbb{1} \leqslant p\}, \quad \mathcal{C}_2 \stackrel{\text{def.}}{=} \left\{\gamma \; ; \; \gamma^T \mathbb{1} \leqslant q\right\}, \quad \mathcal{C}_3 \stackrel{\text{def.}}{=} \left\{\gamma \; ; \; \mathbb{1}^T \gamma \mathbb{1} = m\right\}. \tag{21}$$

The following proposition shows that the KL projection onto those three sets can be obtained in closed form.

Proposition 5. *Let $\gamma \in \mathbb{R}_+^{N\times N}$. Denoting $\gamma^k \overset{\text{def.}}{=} P^{\mathrm{KL}}_{\mathcal{C}_k}(\gamma)$ for $k \in \{1,2,3\}$ where $\mathcal{C}_k$ is defined by* (21)*, one has*

$$\gamma^1 = \operatorname{diag}\left(\min\left(\frac{p}{\gamma\mathbb{1}}, \mathbb{1}\right)\right)\gamma,$$
$$\gamma^2 = \gamma \operatorname{diag}\left(\min\left(\frac{q}{\gamma^T\mathbb{1}}, \mathbb{1}\right)\right),$$
$$\gamma^3 = \gamma\frac{m}{\mathbb{1}^T\gamma\mathbb{1}},$$

where the minimum is component-wise.

Since the considered sets $\mathcal{C}_1$ and $\mathcal{C}_2$ are convex but not affine, one thus needs to use Dykstra iterations (4) which are ensured to converge to the solution of (20).

If $\gamma^\star$ is the optimal solution of (20) and

$$p_m \overset{\text{def.}}{=} \gamma^\star\mathbb{1} \quad \text{and} \quad q_m \overset{\text{def.}}{=} \gamma^{\star,T}\mathbb{1}$$

are its marginals, then we define the active source $\mathcal{S}_m$ and the active target $\mathcal{T}_m$ regions as follow

$$\mathcal{S}_m \overset{\text{def.}}{=} \{x_i \;;\; (p_m)_i/m \geqslant \eta\},$$
$$\mathcal{T}_m \overset{\text{def.}}{=} \{x_i \;;\; (q_m)_i/m \geqslant \eta\},$$

where $\eta > 0$ is a threshold we use to detect the region, namely the active region, where the transported mass is concentrated.

The continuous partial optimal transport problem has been studied in Caffarelli-McCann [18] and Figalli [32]. They show in particular that if there exists an hyperplane separating the support of the two marginals then the "active region" is separated from the "inactive region" by a free boundary which can be parameterized as a semi concave graph over the separating hyperplane. This can be observed on the test case presented in Figure 8. The computation is performed on an uniform 2D-grid of $N = 256 \times 256$ points in $[0,1]^2$, $\varepsilon = 10^{-3}$ and $m = 0.7\min(\langle p,\, \mathbb{1}\rangle, \langle q,\, \mathbb{1}\rangle)$.

5.2 Capacity Constrained Transport

Korman and McCann proposed and studied in [38, 39] a variant of the classical OT problem when there is an upper bound on the coupling weights so as to capture transport capacity constraints.

The capacity is described by $\theta \in (\mathbb{R}_+)^{N\times N}$, where $\theta_{i,j}$ is the maximum possible mass that can be transferred from i to j. The corresponding regularized problem reads, for a ground cost $C \in (\mathbb{R}_+)^{N\times N}$ and marginals $(p,q) \in (\mathbb{R}_+^N)^2$,

$$\min_{\gamma\in\mathbb{R}_+^{N\times N}} \left\{\langle C,\, \gamma\rangle - \varepsilon E(\gamma) \;;\; \gamma\mathbb{1} = p,\; \gamma^T\mathbb{1} = q,\; \gamma \leqslant \theta\right\} \tag{22}$$

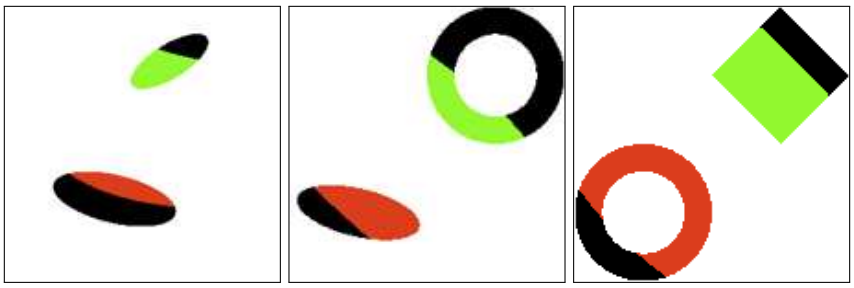


Figure 8: The red region is the active source $\mathcal{S}_m$, the green region is the active target $\mathcal{T}_m$ and the black ones are the inactive regions.

where the inequalities should be understood component-wise.

This problem is equivalent to a KL projection problem of type (2) with $K = 3$ convex sets and

$$\mathcal{C}_1 \stackrel{\text{def.}}{=} \{\gamma \,;\, \gamma \mathbb{1} = p\}, \quad \mathcal{C}_2 \stackrel{\text{def.}}{=} \left\{\gamma \,;\, \gamma^T \mathbb{1} = q\right\}, \quad \mathcal{C}_3 \stackrel{\text{def.}}{=} \{\gamma \,;\, \gamma \leqslant \theta\}. \tag{23}$$

The projection on $\mathcal{C}_1$ and $\mathcal{C}_2$ is given by Proposition 1. The projection on $\mathcal{C}_3$ is simply

$$P^{\mathrm{KL}}_{\mathcal{C}_3}(\gamma) = \min(\gamma, \theta)$$

where the minimum is component-wise.

Korman and McCann [39] established several interesting properties of minimizers in the continuous setting. In particular, they proved (theorem 3.3) that optimal plans must saturate the constraint $\mathcal{C}_3$: the optimal γ is of the form $\theta\, 1_W$, where 1_W is the characteristic function of a subset W of $\mathbb{R}^d \times \mathbb{R}^d$. They also prove in lemma 4.1 [39] symmetries properties between minimizers $\gamma^\star$, $\tilde{\gamma}^\star$ with the same marginals but different capacity constraints θ and $\tilde{\theta}$ which are Hölder conjugate, i.e. $\frac{1}{\theta} + \frac{1}{\tilde{\theta}} = 1$. More precisely, assuming for simplicity that the marginals are symmetric with respect to 0, they show that

$$\gamma^\star = \theta\, 1_W \iff \tilde{\gamma}^\star = \tilde{\theta}\, 1_{R(\Omega \setminus W)} \tag{24}$$

where $R(x, y) = (x, -y)$ is the symmetry with respect to the second marginal axis and W the optimal support of the saturated constraint $\mathcal{C}_3$.

Korman and McCann [39] illustrated their theory with two 1-D numerical test cases (Figures 1 and 2 of [39]) computed by linear programming and a discretization of the problem on a cartesian grid. We tested our method on the same tests. The ground cost is the standard quadratic distance $C_{i,j} = \|x_i - x_j\|^2$, the marginals are discretizations of the uniform distribution on $[-1/2, 1/2]$ using $N = 100$ points $(x_i)_i$. The simulation uses $\varepsilon = 10^{-3}$. We reproduce the expected symmetries (24) in Figure 9 in the 1-D test case for $\theta = \frac{3}{2}$ and $\tilde{\theta} = 3$ and also for the self dual Hölder conjugate $\theta = \tilde{\theta} = 2$.

We also computed the solutions of similar test cases but this time in 2-D, which would be computationally too expensive to solve with linear programming methods. The marginals (p, q) are discretization of the uniform distribution on

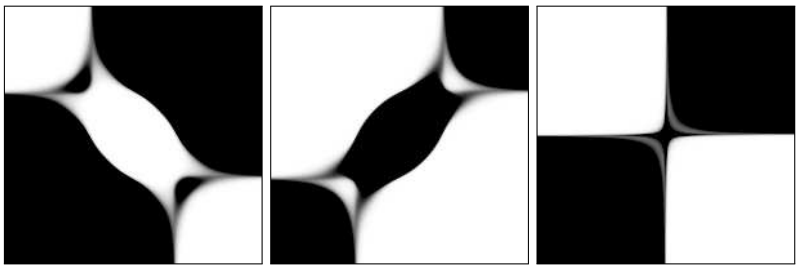


Figure 9: Comparison of optimal coupling $\gamma^\star$ for different values of θ.

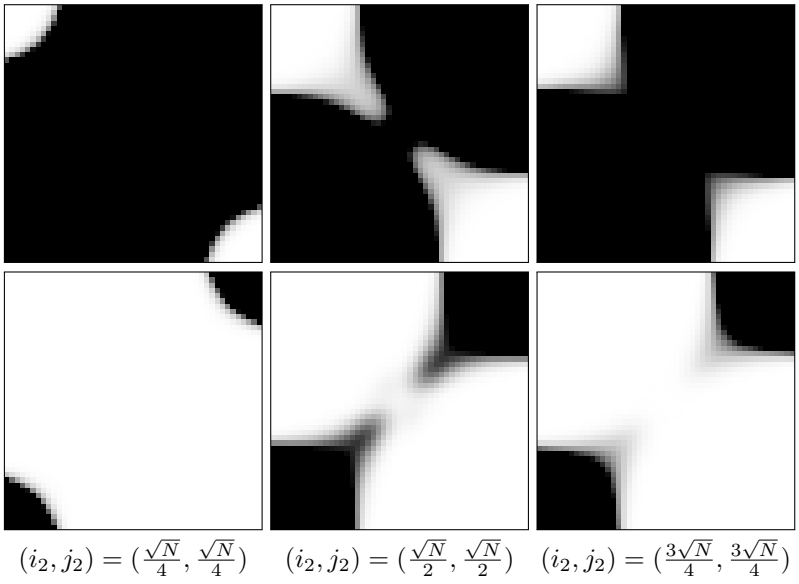


Figure 10: 2-D slices of the optimal coupling $\gamma^\star$ of the form $(\gamma^\star_{(i_1,i_2)(j_1,j_2)})_{i_1,j_1}$, each time for some fixed value of $(i_2, j_2) \in \{1, \dots, \sqrt{N}\}^2$, for $\theta = 3/2$ (top row) and $\theta = 3$ (bottom row).

the square $[-1/2, 1/2]^2$, discretized on a grid of $N = 50 \times 50$ points $(x_i)_i$. The simulation uses $\varepsilon = 10^{-3}$. Figure 10 shows some slices of the 4-D array representing the optimal transport plan $\gamma^\star$, $\tilde{\gamma}^\star$, illustrating the symmetries (24) in this setting.

5.3 Multi-Marginal Partial Transport

In [4] Pass and Kitagawa studied the multi-marginal partial transport problem and, as a natural extension, the partial barycenter problem. Let us consider K marginals $(p_k)_{k=1}^K$ and a transport plan $\gamma \in \mathbb{R}_+^{N^K}$. We now combine the (regularized) partial optimal transport and the "standard" multi-marginal problem, as described in Sections 5.1 and 4.1 respectively. We obtain the following problem

$$\min_\gamma \{\mathrm{KL}(\gamma|\xi) \;;\; \gamma \in \mathcal{C}_1 \cap \mathcal{C}_2 \cap \ldots \cap \mathcal{C}_{K+1}\} \quad \text{where} \quad \xi \overset{\text{def.}}{=} e^{-\frac{C}{\varepsilon}} \tag{25}$$

where

$$\begin{aligned} \mathcal{C}_k &\overset{\text{def.}}{=} \left\{\gamma \in \mathbb{R}_+^{N^K} \;;\; S_k(\gamma) \leqslant p_k\right\} \quad k = 1, \ldots, K, \\ \mathcal{C}_{K+1} &\overset{\text{def.}}{=} \left\{\gamma \in \mathbb{R}_+^{N^K} \;;\; \textstyle\sum_j \gamma_j = m\right\}, \end{aligned}$$

with $m \in [0, \min_k(\langle p_k, \mathbb{1}\rangle)]$

The KL projections on these convex sets are detailed in the following proposition.

Proposition 6. *For any $k = 1, \ldots, K+1$, denoting $\gamma^k = P_{\mathcal{C}_k}^{\mathrm{KL}}(\bar{\gamma})$, one has*

$$\forall\, k = 1, \ldots, K,\ \forall\, j = (j_1, \ldots, j_K), \quad \gamma_j^k = \min\left(\frac{(p_k)_{j_k}}{S_k(\bar{\gamma})_{j_k}}, 1\right)\bar{\gamma}_j$$

$$\gamma^{K+1} = \frac{m}{\sum_j \bar{\gamma}_j}\bar{\gamma}.$$

Once again the sets $\mathcal{C}_k$ are not affine, so one needs to use Dykstra iterations (4).

Figure 11 shows the results obtained when solving (25) with the same three marginals (p_1, p_2, p_3) used in Figure 2, using the cost

$$C_{j_1, j_2, \ldots, j_K} = \sum_{1 \leqslant s,t \leqslant K} \frac{1}{2}\|x_{j_s} - x_{j_t}\|^2, \tag{26}$$

The computation is performed on an uniform 2D-grid of $N = 60 \times 60$ points in $[0,1]^2$, $\varepsilon = 0.005$ and $m = 0.7 \min_k(\langle p_k, \mathbb{1}\rangle)$.

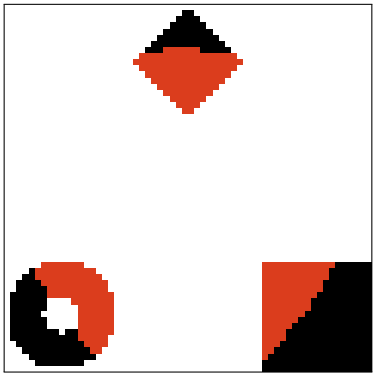


Figure 11: Multi-marginal partial transport. The active regions are displayed in red.

Conclusion

In this paper, we have presented a unifying framework to approximate solutions of various OT-related linear programs through entropic regularization. This regularization enables the use of simple, yet powerful, iterative KL projection methods. While the entropy penalization is not a competitor with interior point methods when it comes to accurately solve the initial linear program, it produces fast approximations at the expense of an extra smoothing. It is thus a method of choice for many applications such as machine learning, image processing or economics.

Aknowledgements

We would like to thank Yann Brenier and Brendan Pass for stimulating discussions. The work of G. Peyré has been supported by the European Research Council (ERC project SIGMA-Vision). JD. Benamou, G. Carlier and L. Nenna gratefully acknowledge the support of the ANR, through the project ISOTACE (ANR-12-MONU-0013) and INRIA through the "action exploratoire" MOKAPLAN. M. Cuturi gratefully acknowledges the support of JSPS young researcher A grant 26700002 and the gift of a K40 card from the NVIDIA corporation.

References

[1] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier. Tomographic reconstruction from a few views: a multi-marginal optimal transport approach. *Preprint Hal-01065981*, 2014.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. on Mathematical Analysis*, 43(2):904–924, 2011.

[3] R. Bapat. d_1ad_2 theorems for multidimensional matrices. *Linear Algebra and its Applications*, 48:437–442, 1982.

[4] B. Bass and J. Kitagawa. The multi-marginal optimal transport problem. *preprint*, 2014.

[5] H. H. Bauschke and P. L. Combettes. A Dykstra-like algorithm for two monotone operators. *Pacific Journal of Optimization*, 4(3):383–391, 2008.

[6] H. H. Bauschke and A. S. Lewis. Dykstra's algorithm with Bregman projections: a convergence proof. *Optimization*, 48(4):409–427, 2000.

[7] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution of the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[8] J-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the Monge–Ampère equation. *Journal of Computational Physics*, 260(1):107–126, 2014.

[9] B. Bhattacharya. An iterative procedure for general probability measures to obtain I-projections onto intersections of convex sets. *Ann. Statist.*, 34(2):878–902, 2006.

[10] J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *Preprint arXiv:1212.2562*, 2012.

[11] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using Lagrangian mass transport. *ACM Transactions on Graphics (SIGGRAPH ASIA'11)*, 30(6), 2011.

[12] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[13] Y. Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *J. of the AMS*, 2:225–255, 1990.

[14] Y. Brenier. The dual least action problem for an ideal, incompressible fluid. *Archive for Rational Mechanics and Analysis*, 122(4):323–351, 1993.

[15] Y. Brenier. Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations. *Communications on Pure and Applied Mathematics*, 52(4):411–452, 1999.

[16] Y. Brenier. Generalized solutions and hydrostatic approximation of the Euler equations. *Phys. D*, 237(14-17):1982–1988, 2008.

[17] R. Burkard, M. Dell'Amico, and S. Martello. *Assignment Problems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[18] L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Ann. of Math. (2)*, 171(2):673–730, 2010.

[19] G. Carlier and I. Ekeland. Matching for teams. *Econom. Theory*, 42(2):397–418, 2010.

[20] G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. Preprint, Ceremade, 2014.

[21] Y. Censor and S. Reich. The Dykstra algorithm with Bregman projections. *Communications in Applied Analysis*, 2:407–419, 1998.

[22] P-A. Chiappori, R. J. McCann, and L. P. Nesheim. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Econom. Theory*, 42(2):317–354, 2010.

[23] R. Cominetti and J. San Martin. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, 1992.

[24] C. Cotar, G. Friesecke, and C. Klzppelberg. Density functional theory and optimal transportation with Coulomb cost. *Communications on Pure and Applied Mathematics*, 66(4):548–599, 2013.

[25] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probability*, 3:146–158, 1975.

[26] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 2292–2300, 2013.

[27] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML), JMLR W&CP*, volume 32, 2014.

[28] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals Mathematical Statistics*, 11(4):427–444, 1940.

[29] R. L. Dykstra. An algorithm for restricted least squares regression. *J. Amer. Stat.*, 78(384):839–842, 1983.

[30] R. L. Dykstra. An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Ann. Probab.*, 13(3):975–984, 1985.

[31] S. Erlander and N.F. Stewart. *The gravity model in transportation analysis: theory and extensions*. Vsp, 1990.

[32] A. Figalli. The optimal partial transport problem. *Arch. Ration. Mech. Anal.*, 195(2):533–560, 2010.

[33] J. Franklin and J. Lorentz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

[34] U. Frisch, S. Matarrese, R. Mohayaee, and A. Sobolevski. Monge-Ampere-Kantorovitch (MAK) reconstruction of the early universe. *Nature*, 417(260), 2002.

[35] A. Galichon and B. Salanié. Matching with trade-offs: Revealed preferences over competing characteristics. Technical report, Preprint SSRN-1487307, 2009.

[36] W. Gangbo and A. Święch. Optimal maps for the multidimensional Monge-Kantorovich problem. *Comm. Pure Appl. Math.*, 51(1):23–45, 1998.

[37] G. Herman. *Image reconstruction from projections: the fundamentals of computerized tomography*. Academic Press, 1980.

[38] K. Jonathan and R. J. McCann. Optimal transportation with capacity constraints. *Preprint arXiv:1201.6404*, 2012.

[39] K. Jonathan and R. J. McCann. Insights into capacity constrained optimal transport. *Proc. Natl. Acad. Sci. USA*, 110:10064–10067, 2013.

[40] L. Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

[41] E. Klann. A Mumford-Shah-like method for limited data tomography with an application to electron tomography. *SIAM J. Imaging Sciences*, 4(4):1029–1048, 2011.

[42] C. Leonard. A survey of the Schrodinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst. A*, 34(4):1533–1574, 2014.

[43] G. Loeper and F. Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *C. R. Acad. Sci. Paris, Ser. I*, 340(4):319–324, 2005.

[44] Q Merigot. A Multiscale Approach to Optimal Transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[45] Y. E. Nesterov and A. S. Nemirovsky. *Interior Point Polynomial Methods in Convex Programming : Theory and Algorithms*. SIAM Publishing, 1993.

[46] N. Papadakis, G. Peyré, and E. Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[47] B. Pass. Uniqueness and Monge solutions in the multimarginal optimal transportation problem. *SIAM Journal on Mathematical Analysis*, 43(6):2758–2775, 2011.

[48] B. Pass. On the local structure of optimal measures in the multi-marginal optimal transportation problem. *Calc. Var. Partial Differential Equations*, 43(3-4):529–536, 2012.

[49] T. E. S. Raghavan. On pairs of multidimensional matrices. *Linear Algebra and its Applications*, 62:263–268, 1984.

[50] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV: International Journal of Computer Vision*, 40(2), 2000.

[51] L. Ruschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, 23(4):1160–1174, 1995.

[52] L. Ruschendorf and W. Thomsen. Closedness of sum spaces and the generalized Schrodinger problem. *Theory of Probability and its Applications*, 42(3):483–494, 1998.

[53] E. Schrodinger. Uber die umkehrung der naturgesetze. *Sitzungsberichte Preuss. Akad. Wiss. Berlin. Phys. Math.*, 144:144–153, 1931.

[54] R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

[55] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly*, 74:402–405, 1967.

[56] R. Sinkhorn and P . Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.

[57] J. Solomon, R.M. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proc. ICML 2014*, 2014.

[58] C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics Series. American Mathematical Society, 2003.

[59] A. G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

[60] G-S. Xia, S. Ferradans, G. Peyré, and J-F. Aujol. Synthesizing and mixing stationary Gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508, 2014.