Frontiers of Science Awards
for Math
pp. 1–13

A numerical method for regularized transportation problems

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

Abstract

In this paper, we discuss several results we have obtained in our Frontiers of Science Award (FSA) paper [4]. In particular we present a numerical method based on iterative Bregman projections and its impact on computational optimal transport.

Contents

1 Historical background **1**

2 Optimal Transport and Sinkhorn **2**

3 Generalised Sinkhorn via Bregman iterations **6**

4 Related research **10**

1 Historical background

Before describing in details the approach of our paper [4] to approximate solutions to linear programs related to optimal transport (OT), we believe it is instructive to give some historical background on entropic optimal transport. The general idea is to introduce an entropic regularization of the initial linear program. For general LP problems, the detailed analysis of this approximation (and its dual) can be found in [19]. In the context of optimal transport, entropic regularization leads to entropy minimization with marginal constraints. It is particularly appealing both theoretically and for computational purposes and, as such, has received a lot of attention in the last decade, following the influential paper of Cuturi [20] who demonstrated the power of Sinkhorn's algorithm in this context.

It is in our opinion fascinating that entropy minimization subject to marginal constraints, which is the core of entropic optimal transport has emerged at different times and in different, seemingly unrelated, contexts. These problems have roots in statistical physics and can be traced back to the seminal work of Schrödinger [41] in 1931, since then, the Schrödinger bridge problem stimulated a lot of interest in connection with large deviations [21], [24], stochastic control [35] and classical OT [34]. Moreover, Kullback-Leibler divergence minimization plays a distinguished role in the information-theoretic approach to many statistical inference problems. In this context, the so-called iterative proportional fitting procedure (IPFP) (equivalent to Sinkhorn), has found numerous applications in the probability and statistics literature, [22], [39],[40], [8]. Entropic regularization is also well-motivated in economics and econometrics, where OT is used to predict flows of commodities in a market. In this context, regularizing the OT problem can ensure the smoothness of such flows [45, 23] or facilitate inference in matching models [27], we refer the reader to the textbook [28] for more and in particular the derivation of entropic OT by random utility choice model (with Gumbel noise). Lastly, it is worth noting that the Sinkhorn algorithm [42, 44, 43] is a fixed-point iteration algorithm for matrix scaling and the connection to entropic OT is not totally obvious at first glance. Given an $n \times m$ matrix A with positive entries, one looks for a an $n \times n$ matrix D and an $m \times m$ matrix Δ (both with positive entries) such that $DA\Delta$ is bi-stochastic (see [10] for an extension to kernels). The solution to this diagonal scaling problem can be found efficiently through the Sinkhorn algorithm. The convergence rate of Sinkhorn's algorithm is linear [25]; it can be implemented in a few lines of code that only require matrix vector products and elementary operations, which can all be easily parallelized on modern hardware. The relevance to entropic OT lies in the fact that the optimality condition for minimizing the Kullback-Leibler with marginal constraints exactly consists in solving a matrix scaling problem and therefore can be solved by Sinkhorn/IPFP.

2 Optimal Transport and Sinkhorn

Discrete Optimal Transport

We now consider the optimal transport problems between probability measures on two finite sets X and Y with, for simplicity, both of cardinality N and we set

$$\mu = \sum_{x \in X} \mu_x \delta_x \qquad \nu = \sum_{y \in Y} \nu_y \delta_y.$$

Remark 2.1 (Notation)**.** *With a slightly abuse of notation we will often identify the measure μ with a vector $\mu \in \mathbb{R}^N$ (which belongs to the N dimensional simplex) containing the weights associated to each point $x \in X$.*

Definition 2.2 (Discrete OT)**.** *The discrete Optimal transport problem between two given measures μ and ν and a given cost function $c : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ is the following minimization problem*

$$\mathcal{MK}_c(\mu, \nu) := \inf \left\{ \sum_{x \in X} \sum_{y \in Y} \gamma_{xy} c(x, y) \mid \gamma \in \Pi(\mu, \nu) \right\}, \tag{2.1}$$

where the set of admissible couplings is now defined as

$$\Pi(\mu, \nu) := \{\gamma \in \mathcal{P}(X \times Y) \mid \gamma \in \mathcal{C}_1 \bigcap \mathcal{C}_2\},$$

where

$$\mathcal{C}_1 := \{\gamma \mid \sum_{y \in Y} \gamma_{xy} = \mu_x \ \forall x \in X\},$$

and

$$\mathcal{C}_2 := \{\gamma \mid , \sum_{x \in X} \gamma_{xy} = \nu_y \ \forall y \in Y\}.$$

Notice that with $\gamma \in \mathcal{P}(X \times Y)$ we mean a probability measure of the form $\gamma = \sum_{x \in X, y \in Y} \gamma_{xy} \delta_{(x,y)}$.
Unfortunately, this linear programming problem has complexity $O(N^3)$ which actually means that it is infeasible for large N. A way to overcome this difficulty is by means of the **Entropic Regularization** which provides an approximation of Optimal Transport with lower computational complexity and easy implementation.

The Entropic Optimal Transport

2.1 The discrete case

We start from the primal formulation of the optimal transport problem, but instead of imposing the constraints $\gamma_{xy} \geqslant 0$, we add a term $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$, involving the (opposite of the) entropy

$$e(r) = \begin{cases} r(\log r - 1) & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ +\infty & \text{if } r < 0 \end{cases}$$

More precisely, given a parameter $\varepsilon > 0$ we consider

$$\boxed{P_\varepsilon = \inf \left\{ \langle \gamma | c \rangle + \varepsilon \, \mathrm{Ent}(\gamma) \mid \gamma \in \Pi(\mu, \nu) \right\},} \tag{2.2}$$

where $\langle \gamma | c \rangle = \sum_{x,y} \gamma_{xy} c(x, y)$ and $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$. Notice that the problem can be equivalently formulated as a minimization of a relative entropy (or Kullback-Leibler divergence) $\mathcal{H}(\rho|\mu) := \sum_x \rho_x (\log(\frac{\rho_x}{\mu_x}) - 1)$, that is

$$\boxed{P_\varepsilon = \inf \left\{ \mathcal{H}(\gamma | \pi_\varepsilon) \mid \gamma \in \Pi(\mu, \nu) \right\},} \tag{2.3}$$

where $\pi_\varepsilon = e^{-c/\varepsilon}$.

Theorem 2.3. *The problem P_ε has a unique solution $\gamma^\star$, which belongs to $\Pi(\mu, \nu)$. Moreover, if $\min(\min_{x \in X} \mu_x, \min_{y \in Y} \nu_y) > 0$ then*

$$\gamma_{x,y} > 0 \ \forall (x, y) \in X \times Y.$$

Before introducing the duality, it is important to state the following convergence result in ε.

Theorem 2.4 (Convergence in ε [38])**.** *The unique solution γ_ε to* (2.2) *converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is*

$$\gamma_\varepsilon \xrightarrow[\varepsilon\to 0]{} \operatorname{argmin}\left\{\operatorname{Ent}(\gamma) \mid \gamma \in \Pi(\mu,\nu),\ \langle\gamma|c\rangle = \mathcal{MK}_c(\mu,\nu)\right\}. \tag{2.4}$$

We want now to derive formally the dual problem. For this purpose we introduce the Lagrangian associated to (2.2)

$$\begin{aligned}\mathcal{L}(\gamma,\varphi,\psi) := \sum_{x,y}\gamma_{xy}c(x,y) + \varepsilon e(\gamma_{xy}) + \sum_{x\in X}\varphi(x)\left(\mu_x - \sum_{y\in Y}\gamma_{xy}\right)\\ + \sum_{y\in Y}\psi(y)\left(\nu_y - \sum_{y\in Y}\gamma_{xy}\right),\end{aligned} \tag{2.5}$$

where $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ are the Lagrange multipliers. Then,

$$P_\varepsilon = \inf_\gamma \sup_{\varphi,\psi} \mathcal{L}(\gamma,\varphi,\psi),$$

and the dual problem is obtained by interchanging the infimum and the supremum:

$$\begin{aligned}D_\varepsilon = \sup_{\varphi,\psi}\min_\gamma \sum_{x,y}\gamma_{xy}(c(x,y) - \psi(y) - \varphi(x) + \varepsilon(\log(\gamma_{xy}) - 1)) +\\ \sum_{x\in X}\varphi(x)\mu_x + \sum_{y\in Y}\psi(y)\nu_y.\end{aligned} \tag{2.6}$$

Taking the derivative with respect to γ_{xy}, we find that for a given φ, ψ, the optimal γ must satisfy:

$$\begin{aligned}c(x,y) - \psi(y) - \varphi(x) + \varepsilon\log(\gamma_{xy}) = 0\\ \text{i.e. } \gamma_{xy} = \exp\left(\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}\right)\end{aligned} \tag{2.7}$$

Putting these values in the definition of D_ε gives

$$D_\varepsilon = \sup_{\varphi,\psi}\Phi_\varepsilon(\varphi,\psi) \text{ with} \tag{2.8}$$

$$\Phi_\varepsilon(\varphi,\psi) := \sum_{x\in X}\varphi(x)\mu_x + \sum_{y\in Y}\psi(y)\nu_y - \sum_{x,y}\varepsilon\exp\left(\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}\right)$$

Note that thanks to the relation (2.7), one can recover a solution to the primal problem from the dual one. This is true because, unlike the original linear programming formulation of the optimal transport problem, the regularized problem (2.2) is smooth and strictly convex. The following duality result holds

Theorem 2.5 (Strong duality)**.** *Strong duality holds and the maximum in the dual problem is attained, that is $\exists\varphi,\psi$ such that*

$$P_\varepsilon = D_\varepsilon = \Phi_\varepsilon(\varphi,\psi).$$

Corollary 2.6. *If (φ, ψ) is the solution to* (2.8)*, then the solution $\gamma^\star$ to* (2.2) *is given by*

$$\gamma_{x,y} = \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}\right)$$

Notice now that the optimal coupling γ can be written as

$$\gamma_{x,y} = D_\varphi e^{\frac{-c(x,y)}{\varepsilon}} D_\psi,$$

where D_φ and D_ψ are the diagonal matrices associated to $e^{\varphi/\varepsilon}$ and $e^{\psi/\varepsilon}$, respectively. The problem is now similar to a matrix scaling problem

Definition 2.7 (Matrix scaling problem)**.** *Let $\pi \in \mathbb{R}^{N\times N}$ be a matrix with positive coefficients. Find D_ψ and D_ψ positive diagonal matrices in $K \in \mathbb{R}^{N\times N}$ such that $D_\varphi K D_\psi$ is doubly stochastic, that is sum along each row and each column is equal to 1.*

Remark 2.8. *Uniqueness fails since if (D_φ, D_ψ) is a solution then so is $(cD_\varphi, \frac{1}{c}D_\psi)$ for every $c \in \mathbb{R}_+$.*

The matrix scaling problem can be easily solved by using an iterative algorithm, known as Sinkhorn-Knopp algorithm, which simply alternates updating D_φ and D_ψ in order to match the marginal constraints (a vector $\mathbf{1}_N$ of ones in this simple case).

Algorithm 1 Sinkhorn-Knopp algorithm for the matrix scaling problem

```
function SINKHORN-KNOPP($K$)
    $D_\varphi^0 \leftarrow \mathbf{1}_N,\ D_\psi^0 \leftarrow \mathbf{1}_N$
    for $0 \leqslant k < k_{\max}$ do
        $D_\varphi^{k+1} \leftarrow \mathbf{1}_N./(\pi D_\psi^k)$
        $D_\psi^{k+1} \leftarrow \mathbf{1}_N./(\pi^T D_\varphi^{k+1})$
    end for
end function
```

where ./ stand for the element-wise division. Denoting by $(\pi_\varepsilon)_{x,y} = e^{\frac{-c(x,y)}{\varepsilon}}$ the algorithm takes the form 2 for the regularized optimal transport problem.

Algorithm 2 Sinkhorn-Knopp algorithm for the regularised optimal transport problem

```
function SINKHORN-KNOPP($\pi_\varepsilon, \mu, \nu$)
    $D_\varphi^0 \leftarrow \mathbf{1}_X,\ D_\psi^0 \leftarrow \mathbf{1}_Y$
    for $0 \leqslant k < k_{\max}$ do
        $D_\varphi^{k+1} \leftarrow \mu./(\pi D_\psi^k)$
        $D_\psi^{k+1} \leftarrow \nu./(\pi^T D_\varphi^{k+1})$
    end for
end function
```

Notice that one can recast the regularized OT in the framework of bistochastic matrix scaling by replacing the kernel $e^{\frac{-c(x,y)}{\varepsilon}}$ with $(\pi_\varepsilon)_{x,y} = \text{diag}(\mu)e^{\frac{-c(x,y)}{\varepsilon}}\,\text{diag}(\nu)$, where $\text{diag}(\mu)$ $(\text{diag}(\nu))$ denotes the diagonal matrix with the vector μ (ν) as main diagonal. In this case the problem (2.2) can be re-written as

$$P_\varepsilon(\mu,\nu) = \inf\left\{\langle\gamma|c\rangle + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma\in\Pi(\mu,\nu)\right\}. \tag{2.9}$$

Remark 2.9 (Sinkhorn as coordinate ascent algorithm)**.** *Notice that algorithm 2 can be seen as a coordinate ascent method on the dual functional $\Phi(\varphi,\psi)$, that is*

$$\varphi^{k+1} = \text{argmax}\,\Phi(\varphi,\psi^k) = \varepsilon\log(\mu) - \varepsilon\log\left(\pi D_\psi^k\right), \tag{2.10}$$

$$\psi^{k+1} = \text{argmax}\,\Phi(\varphi^{k+1},\psi) = \varepsilon\log(\nu) - \varepsilon\log\left(\pi^T D_\varphi^{k+1}\right). \tag{2.11}$$

Remark 2.10 (Sinkhorn as an alternate Bregman projection algorithm)**.** *Problem* (2.2) *can be reformulated as follows*

$$\inf_{\mathcal{C}}\mathcal{H}(\gamma|\pi_\varepsilon), \tag{2.12}$$

where $\mathcal{C} = \mathcal{C}_1\bigcap\mathcal{C}_2$. A natural way to solve it (and the original point of view in [4]) consists in projecting alternatively on $\mathcal{C}_1$ and $\mathcal{C}_2$ to obtain a sequence converging to the primal solution, that is

$$\gamma^{2k} = \text{proj}^{\mathcal{H}}_{\mathcal{C}_1}(\gamma^{2k-1}) := \frac{\mu}{\gamma^{2k-1}\mathbf{1}_N}\gamma^{2k-1}, \tag{2.13}$$

$$\gamma^{2k+1} = \text{proj}^{\mathcal{H}}_{\mathcal{C}_2}(\gamma^{2k}) := \gamma^{2k}\frac{\nu}{\gamma^{2k,T}\mathbf{1}_N}. \tag{2.14}$$

Noticing now that γ^{2k} and γ^{2k+1} can be decomposed as $\gamma^{2k} = D_\varphi^k\pi_\varepsilon D_\psi^k$ and $\gamma^{k+1} = D_\varphi^k\pi_\varepsilon D_\psi^{k+1}$, on can recover the Sinkhorn algorithm we detailed above.

3 Generalised Sinkhorn via Bregman iterations

In this section we present the main results we obtained in in our FSA paper [4]. As already pointed out above regularized problem (2.2) corresponds to a relative entropy (Kullback-Leibler Bregman divergence) projection of a vector (representing some initial joint distribution) on the polytope of constraints. We were able to 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), see for instance Remark 2.10 or more generally Bregman-Dykstra iterations (when inequality constraints are involved). In particular this approach let us solve many 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 unbalanced optimal transport, etc. In the following we give a glimpse of our results for some of these problems.

Wasserstein barycenter We are given a set $(\mu_k)_{k=1}^K$ of input marginals, and we wish to compute a weighted barycenter according to the Wasserstein metric as defined in [1]. Following [1], 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 $(\lambda_k)_{k=1}^K$, we consider the regularized Wasserstein barycenter problem

$$\min\left\{\sum_{k=1}^{K}\lambda_k\mathcal{H}(\gamma_k|\pi_\varepsilon)\mid(\gamma_k)_{k=1}^K\in\mathcal{C}_1\cap\mathcal{C}_2\right\} \tag{3.1}$$

and the constraint sets are defined by

$$\begin{aligned}\mathcal{C}_1 &:= \{(\gamma_k)_{k=1}^K \mid \gamma_k^T\mathbf{1}=\mu_k,\ \forall k\}\\ \mathcal{C}_2 &:= \{(\gamma_k)_{k=1}^K \mid \exists\overline{\mu}\in\mathbb{R}^N,\forall k,\ \gamma_k\mathbf{1}=\overline{\mu}\},\end{aligned}$$

where $\overline{\mu}$ denotes the barycenter. It is easy to see that the projection on $\mathcal{C}_1$ is computed as in Remark 2.10 where as the one on $\mathcal{C}_2$ is detailed in the following proposition

Proposition 3.1. *For $(\overline{\gamma}_k)_k\in(\mathbb{R}_+^{N\times N})^K$, the projection $(\gamma_k)_{k=1}^K=\mathrm{proj}_{\mathcal{C}_2}^{\mathcal{H}}((\overline{\gamma}_k)_k)$ satisfies*

$$\forall k,\quad \gamma_k=\mathrm{diag}\left(\frac{\overline{\mu}}{\overline{\gamma}_k\mathbf{1}}\right)\overline{\gamma}_k,\ \textit{where}\ \overline{\mu}:=\prod_{r=1}^{K}(\overline{\gamma}_r\mathbf{1})^{\lambda_r} \tag{3.2}$$

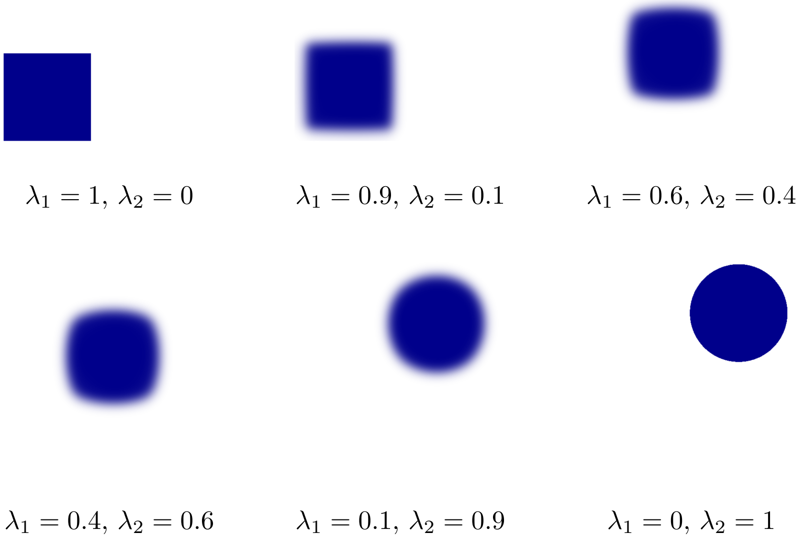
where $\prod$ and $(\cdot)^{\lambda_r}$ should be understood as entry-wise operators.

Notice that Sinkhorn algorithm can be easily generalised to compute the barycenter problem: instead of updating K transport plans one can only consider $2K$ vectors D_{φ_k} and D_{ψ_k} and a vector tracking the update of the barycenter. The iterations take the following form

$$\begin{aligned}D_{\varphi_k}^{(n+1)} &= \frac{\overline{\mu}^{(n)}}{\pi_\varepsilon D_{\psi_k}^{(n)})},\\ D_{\psi_k}^{(n+1)} &= \frac{\mu_k}{\pi_\varepsilon^T D_{\varphi_k}^{(n+1)})},\\ \overline{\mu}^{(n)} &= \prod_{k=1}^{K}\left(D_{\varphi_k}^{(n+1)}\odot(\pi_\varepsilon D_{\psi_k}^{(n+1)})\right)^{\lambda_k}.\end{aligned}$$

Figure 1 shows an example of barycenters computation for $K=2$ (in this case the barycenter is the geodesic between the two measures μ and ν). The computation is performed on an uniform 2D-grid of $N=500\times500$ points in $[0,1]^2$, $\varepsilon=5*10^{-4}$.

Multi-Marginal OT In this case we have to deal with K marginals μ_k, meaning that the optimal γ is now a coupling on $\times_{k=1}^K X_i$. The solution lies now at the intersection of m convex sets $\bigcap_{i=1}^K\mathcal{C}_i$ associated to each marginal. Sinkhorn algorithm can now be straightforward generalised to this case by iterating over K D_{φ_i} variables for each marginal constraint. For more details on this generalisation as well as for the applications, see [6, 5, 7, 36].



$\lambda_1 = 1,\ \lambda_2 = 0$ $\lambda_1 = 0.9,\ \lambda_2 = 0.1$ $\lambda_1 = 0.6,\ \lambda_2 = 0.4$

$\lambda_1 = 0.4,\ \lambda_2 = 0.6$ $\lambda_1 = 0.1,\ \lambda_2 = 0.9$ $\lambda_1 = 0,\ \lambda_2 = 1$

Figure 1: Left: support of the densities μ and ν and the barycenter for different values of λ.

Partial OT In the partial transport problem, one is given two marginals (μ, ν), not necessarily with the same total mass. We wish to transport only a given fraction of mass

$$m \in [0, \min(\mu^T \mathbf{1}, \nu^T \mathbf{1})],$$

minimizing the transportation cost $\langle \gamma | c \rangle$.

The corresponding regularized problem reads

$$\min \left\{ \mathcal{H}(\gamma|\pi_\varepsilon) \mid \gamma \mathbf{1} \leqslant \mu,\ \gamma^T \mathbf{1} \leqslant \nu, \sum_{x,y} \gamma_{xy} = m \right\} \tag{3.3}$$

where the inequalities should be understood component-wise.

This is equivalent to a relative entropy minimization problem on the intersection $\mathcal{C}_1 \cap \mathcal{C}_2 \cap \mathcal{C}_3$ of $K = 3$ convex sets associated to the marginal and mass constraints

$$\mathcal{C}_1 := \{\gamma \mid \gamma \mathbf{1} \leqslant \mu\},\ \mathcal{C}_2 := \{\gamma \mid \gamma^T \mathbf{1} \leqslant \nu\},\ \mathcal{C}_3 := \{\gamma \mid \sum_{x,y} \gamma_{xy} = m\}.$$

The following proposition shows that the projection, with respect to the relative entropy, onto those three sets can be obtained in closed form.

Proposition 3.2. *Let $\gamma \in \mathbb{R}_+^{N\times N}$. Denoting $\gamma^k := \mathrm{proj}^{\mathcal{H}}_{\mathcal{C}_k}(\gamma)$ for $k \in \{1,2,3\}$ where $\mathcal{C}_k$ is defined as above, one has*

$$\gamma^1 = \mathrm{diag}\left(\min\left(\frac{\mu}{\gamma\mathbf{1}}, \mathbf{1}\right)\right)\gamma,$$
$$\gamma^2 = \gamma\,\mathrm{diag}\left(\min\left(\frac{\nu}{\gamma^T\mathbf{1}}, \mathbf{1}\right)\right),$$
$$\gamma^3 = \gamma\frac{m}{\sum_{xy}\gamma_{xy}},$$

where the minimum is component-wise.

In Figure 2 we plot the support of the densities μ and ν (left) and the support of the marginals of the optimal solution (right). The computation is performed on an uniform 2D-grid of $N = 500 \times 500$ points in $[0,1]^2$, $\varepsilon = 10^{-3}$ and $m = 0.4\min(\mu^T\mathbf{1}, \nu^T\mathbf{1})$.

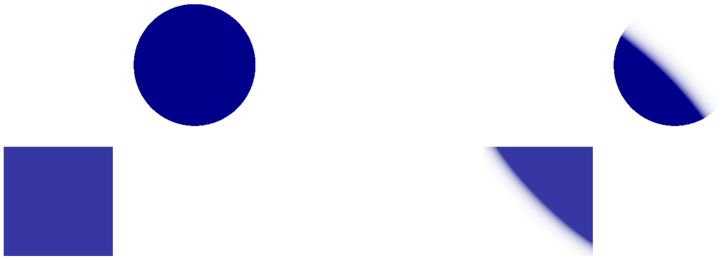


Figure 2: Left: support of the densities μ and ν. Right: support of the marginals of the optimal γ.

Capacity constraint OT Korman and McCann proposed and studied in [31, 32] 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 $\overline{\gamma} \in (\mathbb{R}_+)^{N\times N}$, where $\overline{\gamma}_{xy}$ is the maximum possible mass that can be transferred from x to y. The corresponding regularized problem reads

$$\min\left\{\mathcal{H}(\gamma|\pi_\varepsilon) \mid \gamma\mathbf{1} = \mu,\ \gamma^T\mathbf{1} = \nu, \gamma \leqslant \overline{\gamma}\right\} \tag{3.4}$$

where the inequalities should be understood component-wise. This is equivalent to a relative entropy minimization problem with $K = 3$ convex sets and

$$\mathcal{C}_1 := \{\gamma \mid \gamma\mathbf{1} = \mu\},\ \mathcal{C}_2 := \{\gamma \mid \gamma^T\mathbf{1} = \nu\},\ \mathcal{C}_3 := \{\gamma \mid \gamma \leqslant \overline{\gamma}\}.$$

The projection on $\mathcal{C}_1$ and $\mathcal{C}_2$ is as in Remark 2.10. The projection on $\mathcal{C}_3$ is simply

$$\mathrm{proj}^{\mathcal{H}}_{\mathcal{C}_3}(\gamma) = \min(\gamma, \overline{\gamma}).$$

where the minimum is component-wise.

4 Related research

As explained in the previous paragraphs, in [4], we formulated a variety of entropically regularized generalizations of (discrete) OT in the framework of Bregman iterative projections, and proposed simple scaling algorithms à la Sinkhorn to solve them. The scope and analysis of scaling algorithms for generalized OT problems has been significantly extended in recent years. Of particular importance, in our opinion, is the extension to unbalanced OT, see [17] and gradient flows [37], [13].

In the setting of [4], convergence is in principle ensured by general results from [11] (affine constraints) and [3] (inequality constraints). This being said, [4] does not address at all quantitative convergence issues of the algorithm or accuracy as the penalization parameter ε goes to 0. There are actually several approaches to the convergence analysis of Sinkhorn. Linear convergence of the Sinhkorn algorithm for two marginals and a bounded cost is well-known both in the discrete and continuous cases. A very elegant proof uses a celebrated theorem of Birkhoff [9] to show that the Sinkhorn algorithm consists in iterating a contraction for the Hilbert projective metric, see Franklin and Lorenz [26], Chen, Georgiou and Pavon, [15]. There are convergence proofs rather based on entropic and functional inequality arguments, see Rüschendorf [39] and more recently, Léger [33], Ghosal and Nutz [29] who obtained sublinear rates under much more general conditions. The Hilbert metric contraction proof does to extend to the multi-marginal case, which can be addressed by the coordinate ascent on the dual interpretation of Sinkhorn see [12]. Note that most of the convergence results mentioned above involve constants that scale exponentially badly with ε. Quantitative convergence results to OT as ε to 0 is still a very active area of research, but since it depends very much on the cost, marginals and reference measure, it is far beyond the scope of this short discussion.

Finally, we wish to mention some recent works concerning the specific case of Wasserstein barycenters, which is a representative illustration of the results of [4]. An interesting alternative to Sinkhorn is the stochastic algorithm of [18]. Among improvements, let us mention the doubly regularized approach of Chizat [16] where one does not only add an entropic penalty on the plans but also on the unknown (barycenter) marginal (as in [14]), resulting in an improvement of the entropic smoothing bias, see [30]. The story is certainly not over yet, in particular in view of the inspiring analysis of Altschuler and Boix-Adserà [2] who proved that the problem is NP-hard.

References

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

[2] Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters are np-hard to compute. *SIAM Journal on Mathematics of Data Science*, 4(1):179–203, 2022.

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

[4] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[5] Jean-David Benamou, Guillaume Carlier, Simone Di Marino, and Luca Nenna. An entropy minimization approach to second-order variational mean-field games. *Math. Models Methods Appl. Sci.*, 29:1553–1583, 2019.

[6] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. A numerical method to solve multi-marginal optimal transport problems with Coulomb cost. In Roland Glowinski, Stanley J. Osher, and Wotao Yin, editors, *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 577–601. Springer, 2016.

[7] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numer. Math.*, 142:33–54, 2019.

[8] 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.

[9] Garrett Birkhoff. Extensions of Jentzsch's theorem. *Trans. Amer. Math. Soc.*, 85:219–227, 1957.

[10] J. M. Borwein, A. S. Lewis, and R. D. Nussbaum. Entropy minimization, DAD problems, and doubly stochastic kernels. *J. Funct. Anal.*, 123(2):264–307, 1994.

[11] 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.

[12] Guillaume Carlier. On the linear convergence of the multimarginal Sinkhorn algorithm. *SIAM J. Optim.*, 32(2):786–794, 2022.

[13] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

[14] Guillaume Carlier, Katharina Eichinger, and Alexey Kroshnin. Entropic-wasserstein barycenters: Pde characterization, regularity, and clt. *SIAM Journal on Mathematical Analysis*, 53(5):5880–5914, 2021.

[15] Yongxin Chen, Tryphon Georgiou, and Michele Pavon. Entropic and displacement interpolation: a computational approach using the Hilbert metric. *SIAM J. Appl. Math.*, 76(6):2375–2396, 2016.

[16] Lénaïc Chizat. Doubly regularized entropic wasserstein barycenters. *arXiv preprint arXiv:2303.11844*, 2023.

[17] Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, 87(314):2563–2609, 2018.

[18] Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2018.

[19] R. Cominetti and J. San Martín. Asymptotic analysis of the exponential

penalty trajectory in linear programming. *Math. Programming*, 67(2):169–187, 1994.

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

[21] Donald A. Dawson and Jürgen Gärtner. Large deviations from the McKean-Vlasov limit for weakly interacting diffusions. *Stochastics*, 20(4):247–308, 1987.

[22] 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.

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

[24] Hans Föllmer. Random fields and diffusion processes. In *École d'Été de Probabilités de Saint-Flour XV–XVII, 1985–87*, volume 1362 of *Lecture Notes in Math.*, pages 101–203. Springer, Berlin, 1988.

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

[26] Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

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

[28] Alfred Galichon. *Optimal transport methods in economics.* Princeton University Press, Princeton, NJ, 2016.

[29] Promit Ghosal and Marcel Nutz. On the convergence rate of sinkhorn's algorithm, 2022.

[30] Hicham Janati, Marco Cuturi, and Alexandre Gramfort. Debiased sinkhorn barycenters. In *International Conference on Machine Learning*, pages 4692–4701. PMLR, 2020.

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

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

[33] Flavien Léger. A gradient descent perspective on Sinkhorn. *Appl. Math. Optim.*, 84(2):1843–1855, 2021.

[34] 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.

[35] Toshio Mikami and Michèle Thieullen. Optimal transportation problem by stochastic optimal control. *SIAM J. Control Optim.*, 47(3):1127–1139, 2008.

[36] Luca Nenna. *Numerical methods for multi-marginal optimal transportation.* PhD thesis, Université Paris sciences et lettres, 2016.

[37] Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM J. Imaging Sci.*, 8(4):2323–2351, 2015.

[38] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With

applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

[40] 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.

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

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

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

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

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

MOKAPLAN, INRIA Paris
Paris, France.
E-mail address: Jean-David.Benamou@inria.fr

Ceremade, Université Paris-Dauphine, PSL and Inria-Paris Mokaplan
Pl. de Lattre de Tassigny, 75775 Paris cedex 16, France.
E-mail address: carlier@ceremade.dauphine.fr

CREST-ENSAE, Institut Polytechnique de Paris
Paris, France.
E-mail address: marco.cuturi@ensae.fr

Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, ParMA, Inria Saclay
91405, Orsay, France.
E-mail address: luca.nenna@universite-paris-saclay.fr

CNRS, ENS - PSL University
Paris, France.
E-mail address: gabriel.peyre@ens.fr