Numerical Tours - Computational Optimal Transport

Tom ROSSA
ENS Paris-Saclay / ENSAE
Paris, France
tom.rossa@ensae.fr

ABSTRACT

This report briefly presents the results from the practical sessions "Optimal Transport with Linear Programming" and "Entropic Regularization of Optimal Transport," available in this GitHub repository.

KEYWORDS

Discrete Optimal Transport, Entropic Regularization, Sinkhorn Algorithm

1 OPTIMAL TRANSPORT WITH LINEAR PROGRAMMING

This section considers the discrete optimal transport problem between two discrete measures, which can be viewed as weighted point clouds:

$$\alpha = \sum_{i=1}^{n} \mathbf{a}_i \delta_{x_i} \quad \text{and} \quad \beta = \sum_{j=1}^{m} \mathbf{b}_j \delta_{y_j},$$

where $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, and $n = 90, m = 120$.

Discrete Kantorovich Problem. We define a cost matrix $C \in \mathbb{R}^{n\times m}$, where $C_{i,j} = ||x_i - y_j||_2^2$. The discrete Kantorovich problem aims to minimize the total transport cost under mass conservation constraints. Formally, it is expressed as:

$$\min_{P\geq 0} \quad \sum_{i=1}^{n} \sum_{j=1}^{m} C_{i,j} P_{i,j},$$

subject to

$$P \in \mathcal{U}(\mathbf{a}, \mathbf{b}) = \{P \in \mathbb{R}^{n\times m}; P\mathbf{1}_m = \mathbf{a}, P^\top \mathbf{1}_n = \mathbf{b}\}.$$

This problem is a linear program, solved here using the Python package CVXPY for various types of point clouds. The following discrete measures are considered for solving the optimal transport problem:

- **Example 1**: The source sample is drawn from a Gaussian distribution $\mathcal{N}(0, 0.3)$, while the target sample is a three-mode Gaussian mixture.
- **Example 2**: The source sample follows a uniform distribution along the unit circle, while the target sample is drawn from a uniform distribution $\mathcal{U}([-1, 1])$.
- **Example 3**: The source sample is drawn from a standard Gaussian distribution $\mathcal{N}(0, 1)$, while the target sample is drawn from a uniform distribution over a more complex shape, such as a spiral.
- **Example 4**: The source sample follows a uniform distribution within the unit circle, while the target sample is uniformly distributed over a discrete set, such as the vertices of a square $[-1, 1] \times [-1, 1]$, i.e., a transition from a continuous distribution to a discrete one.

After solving the corresponding optimization problems, it is observed that the resulting optimal transport plans are relatively sparse. This means the mass of each point is transported to only a small number of nearby points. These plans are visualized by plotting the point clouds with connections indicating the transported mass between points.

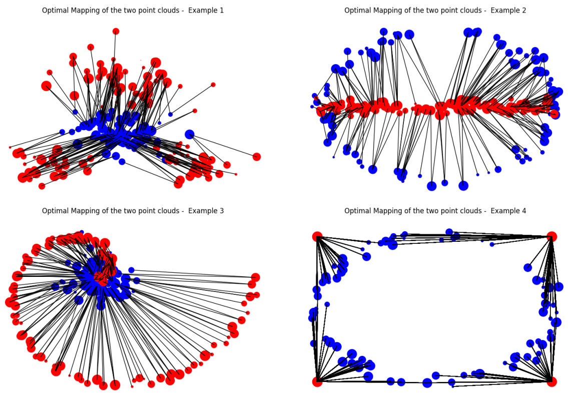


Figure 1: Optimal mappings for different point clouds.

Displacement Interpolation. Displacement interpolation refers to the construction of a geodesic path μ_t in the Wasserstein space, parameterized by $t \in [0, 1]$. This geodesic minimizes the quadratic Wasserstein distance and solves the following variational problem:

$$\mu_t = \underset{\mu}{\text{argmin}}\,(1-t)W_2(\alpha, \mu)^2 + tW_2(\beta, \mu)^2,$$

where α and β are the initial and target distributions, respectively.

This formulation generalizes the concept of Euclidean barycenters to the barycenters of probability distributions. When the optimal transport plan $P^\star$ between α and β is computed, the interpolated distribution at time t is given by:

$$\mu_t = \sum_{i,j} P^{\star}_{i,j} \delta_{(1-t)x_i + t y_j}.$$

This representation shows that the mass transported between the source and target distributions is interpolated linearly in space, while preserving the geodesic structure in the Wasserstein metric.

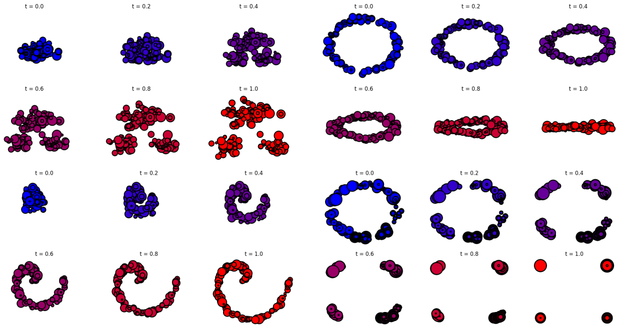


Figure 2: Displacement Interpolation for Different Examples

Special Case: $n = m$ ***and uniform weights.*** In the case where the point clouds are of the same size, $n = m$, and the weights are uniform, $\mathbf{a} = \left(\frac{1}{n}\right)_i$, $\mathbf{b} = \left(\frac{1}{m}\right)_j$, it can be shown that the optimal transport plans are permutation matrices:

$$\exists \sigma \in \text{Perm}(n), \quad P^* = P_\sigma = (\mathbf{1}_{\{j=\sigma(i)\}}(i,j))_{i,j}$$

This results from the fact that, in this case, $\mathcal{U}(\mathbf{1}, \mathbf{1})$ consists precisely of bistochastic matrices, and its extreme points are permutation matrices. Furthermore, it can be shown that the intersection between the set of solutions to the discrete Kantorovich problem and the extreme points of $\mathcal{U}(\mathbf{1}, \mathbf{1})$ is non-empty. For each of the previously mentioned examples, by setting $n = m = 90$ and using uniform weights, permutation matrices are indeed obtained as optimal transport plans.

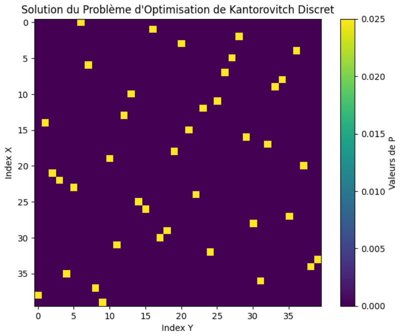


Figure 3: Permutation Matrix for Optimal Transport Plan

2 ENTROPIC REGULARIZATION OF OPTIMAL TRANSPORT

In this section, we consider the regularized optimal transport problem, specifically its discrete version. In this problem, the set of constraints remains unchanged, but the objective function is modified by adding a regularization term:

$$E(P) = -\sum_{i,j} P_{i,j} \left(\log(P_{i,j}) - 1\right)$$

This problem can be reformulated in terms of the Kullback-Leibler Divergence:

$$W_\epsilon(a,b) = \epsilon \min_{P \in U(a,b)} \text{KL}(P \| K)$$

where $K = e^{-\frac{C}{\epsilon}}$ and the KL divergence is:

$$\text{KL}(P\|K) = \sum_{i,j} P_{i,j} \left(\log\left(\frac{P_{i,j}}{K_{i,j}}\right) - 1\right)$$

It can be shown that the solution to this problem is unique since the objective function is ϵ-strongly convex. Therefore, solving this problem falls within the framework of convex optimization. Entropic regularization introduces a penalty based on entropy, where the optimal solution is influenced in part by entropy to avoid "discontinuous" or overly concentrated couplings. This encourages a more balanced and distributed solution, improving computational stability. The solution to this problem is of the form:

$$P_\epsilon = \text{diag}(u) K \text{diag}(v)$$

where:

- $K = e^{-\frac{C}{\epsilon}}$ is the Gibbs kernel,
- u, v are positive vectors ensuring $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$.

This problem can be solved efficiently with good convergence properties using the Sinkhorn algorithm, which iteratively updates the vectors u and v, such that:

$$u \leftarrow \frac{a}{Kv}, \quad v \leftarrow \frac{b}{K^\top u}, \quad \text{entry-wise division operator}$$

Thus, as the regularization parameter ϵ increases, the solution becomes more diffuse, and the convergence of the algorithm becomes faster. However, the trade-off is that with higher ϵ, the regularized solution becomes farther from the true solution of the unregularized problem.

Transport Between Histograms. In this section, we consider two histograms obtained from observations of 2-dimensional Gaussian mixtures with respective parameters:

$$\Pi_1 = \mathcal{N}(0.3, 0.06) + \mathcal{N}(0.6, 0.1), \quad \Pi_2 = \mathcal{N}(0.1, 0.05) + \mathcal{N}(0.7, 0.05)$$

The trade-off between the regularization parameter ϵ can be illustrated between rapid convergence when ϵ is high and proximity to the true solution when ϵ is low:

$$L^\epsilon_{\mathbf{C}}(\mathbf{a}, \mathbf{b}) \mapsto L_{\mathbf{C}}(\mathbf{a}, \mathbf{b}), \quad \text{as } \epsilon \longrightarrow 0$$

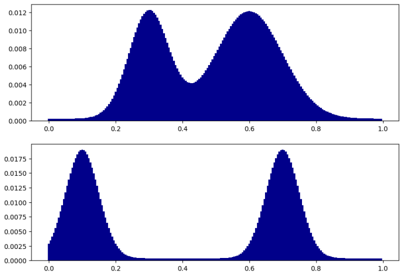


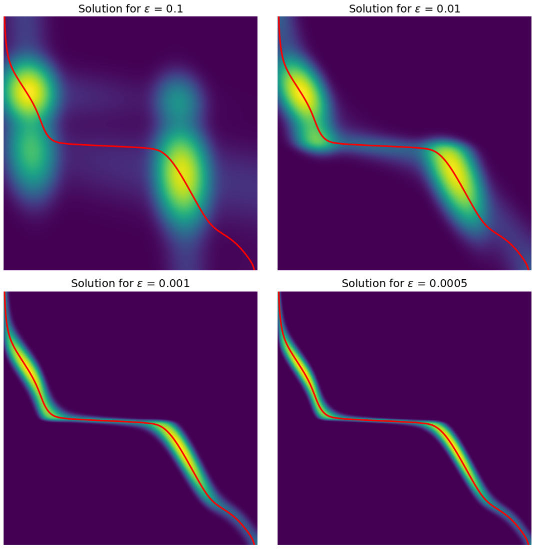
Figure 4: Initial and Target Histograms

Effect of the Regularization Term on the Solution. One can compute an approximation of the transport plan between two measures by calculating the so-called barycentric projection map:

$$t_i \in [0,1] \longmapsto s_j \frac{\sum_j P_{i,j} t_j}{\sum_j P_{i,j}} = \frac{[u \odot K(v \odot t)]_j}{a_i}$$

where $\odot$ and $\div$ are the entry-wise multiplication and division operations, respectively. This computation can thus be performed efficiently using only multiplication with the kernel K.

Thus, to observe the impact of the parameter ϵ on the solution of the problem, we can visualize the solutions of the couplings obtained using the Sinkhorn algorithm with different values of the parameter. We overlay the approximation of the optimal transport plan between the two measures.


Solution for ε = 0.1

Solution for ε = 0.01

Solution for ε = 0.001

Solution for ε = 0.0005

Figure 5: Solutions of the regularized problem for different values of ϵ

It is observed that as the entropy regularization term increases, the solution becomes more diffuse. Conversely, as ϵ approaches 0, the solution approximates a coupling with a deterministic dependency structure between the two measures, close to the optimal Monge Map (which is deterministic since it is the gradient of a convex function according to Brenier's theorem). Furthermore, we track the evolution of constraint errors over iterations, which serve as good indicators of algorithm convergence.

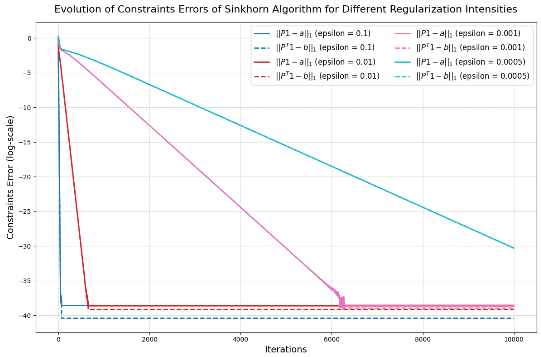

Evolution of Constraints Errors of Sinkhorn Algorithm for Different Regularization Intensities

Figure 6: Evolution over iterations of the Sinkhorn Algorithm of Constraint Errors for different values of ϵ

It is evident that as ϵ increases, the algorithm converges more rapidly, with the error approaching zero quickly. However, it is noted that the constraint error never truly converges to zero due to the entropic bias.

Wasserstein Barycenters. Wasserstein barycenters extend the concept of centroids to probability measures in Wasserstein space. Given a set of probability measures $\{\mu_i\}_{i=1}^n$ on $\mathbb{R}^d$ with associated weights $\{\lambda_i\}_{i=1}^n$, the Wasserstein barycenter μ minimizes the weighted sum of Wasserstein distances:

$$\mu = \arg \min_{\nu \in \mathcal{P}(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(\mu_i, \nu)$$

where W_2 denotes the 2-Wasserstein distance.

Therefore, we compute barycenters between multiple images, corresponding to discrete probability measures for various weight vectors λ, and summarize the interpolations obtained in the following plot.

Received 26 December 2024