Seminar In Optimal Transport

The Entropic Optimal Transport

Technical University Of Berlin
Department Of Mathematics

Master in Scientific Computing

Author:
Moussa Atwi
ID: 464241

Teacher:
Prof. Dr. Gabriele Steidl

Contents

Introduction

We analyze discrete Optimal transport problems in the so-called Kantorovich form, where we seek a transport plan between two marginals that are probability measures on compact subsets of Euclidean space such that a certain cost functional is minimal. Consider the Kantorovich formulation of Optimal transport problem

$$L(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \langle c\ ,\ \gamma\rangle = \min_{\gamma\in\Pi(\mu,\nu)} \sum_{x\in X}\sum_{y\in Y} \gamma_{xy} c(x,y), \tag{1}$$

where,

- γ describes how to move some measure onto another one of the same mass
- the set of admissible couplings is defined as

$$\Pi(\mu,\nu) := \left\{ \gamma \in X\times Y \,|\, \gamma_{xy} \geq 0, \sum_{y\in Y} \gamma_{xy} = \mu_x \,\forall x \in X, \sum_{x\in X} \gamma_{xy} = \nu_y \,\forall y \in Y \right\}.$$

- the constraint $\sum_{y\in Y} \gamma_{xy} = \mu_x$ encodes the fact that all mass from x is transported somewhere in Y, while the constraint $\sum_{x\in X} \gamma_{xy} = \nu_y$ tells us that the mass at y is transported from somewhere in X.

Computing OT is costly: Solving discrete OT i.e., computing transport between sums of Diracs, reduces to solving a large-scale finite dimensional linear program, but it turns to be computationally difficult. When the mass of each Dirac is constant and the two measures have the same number of points (Dirac masses) N. The worst case complexity of computing the optimum with any of the algorithms known so far is at most of order $\mathcal{O}(N^3 \log(N))$ and roughly turns out to be super-cubic $\mathcal{O}(N^3)$, which is still computationally too demanding for most applications.

Singularity of optimal plans: In the case where $c(x,y) = |x-y|^2$ is the squared Euclidean distance, this implies that optimal plans are singular with respect to the Lebesgue measure. Hence, the optimal plan is not a measurable function, and so standard approximation techniques from numerical analysis (e.g. by piecewise constant or piecewise linear functions) are not applicable.

To overcome these issues: Regularization methods have been proposed to approximate the OT problem by adding a penalty and to obtain approximate solutions that are functions instead of measures, which in turn can be treated by classical discretization techniques in order to solve the regularized problem.

The underlying idea of an entropic regularization is to regularize OT by adding a multiple of the entropy of the transportation plan

$$E(\gamma) := \sum_{i,j} \gamma_{i,j}(\log(\gamma_{i,j}) - 1),$$

and changing (1) into a strictly convex program with a unique solution γ^ϵ

$$L^\epsilon(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \langle c\ ,\ \gamma\rangle + \epsilon E(\gamma). \tag{2}$$

The entropy forces the solution to have a spread support, thus deviating from the fact that optimal couplings are sparse. The minimization of this criterion is achieved numerically by the very simple Sinkhorn algorithm. The introdution of the Sinkhorn divergence enables to obtain an approximation of the OT distance which can be computed with a complexity of algorithm of order $\mathcal{O}(N^2\epsilon^{-3})$, hence in a much faster way than the original OT problem.

Other entropic regularization penalties: Note also that others regularizing penalty have been proposed, for instance the **Kullback-Leibler divergence** or **relative entropy** with respect to the product of marginals defined as

$$KL(\gamma|\mu\otimes\nu) := \sum_{i,j} \gamma_{i,j}\left[\log\left(\frac{\gamma_{i,j}}{\mu_i\,\nu_j}\right) - 1\right] \quad \text{or} \quad \sum_{i,j} \gamma_{i,j}\log\left(\frac{\gamma_{i,j}}{\mu_i\,\nu_j}\right)$$

The unique solution γ^ϵ to (2) converges to the optimal solution with minimal entropy within the set of all optimal solutions of OT problem. Furthermore, (2) can be refactored as a projection problem

$$\gamma^\epsilon = \mathrm{Proj}^{KL}_{\Pi(\mu,\nu)}(\mathcal{K}) := \arg\min KL(\gamma \mid \mathcal{K}).$$

γ^ϵ is a projection onto $\Pi(\mu,\nu)$ of the Gibbs kernel associated to the cost matrix

$$\mathcal{K}_{i,j} := e^{-\frac{c_{i,j}}{\epsilon}}.$$

On the other hand, for **continuous** OT, the regularized problem reads

$$(P) = \min_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\,d\gamma(x,y)\ +\ \epsilon\,KL(\gamma \mid \mu\otimes\nu), \tag{3}$$

where for $\gamma \ll \mu\otimes\nu$

$$KL(\gamma \mid \mu\otimes\nu) := \int_{X\times Y} \log\left[\frac{d\gamma(x,y)}{d\mu(x)d\nu(y)}\right] d\gamma(x,y),$$

the dual formulation of the regularized problem (3)

$$\begin{aligned}(D) = \max_{(\phi,\psi)\in C(X)\times C(Y)} &\int_X \phi(x)\,d\mu(x) + \int_Y \psi(y)\,d\nu(y) \\ &- \epsilon\int_{X\times Y} e^{\frac{\phi(x)+\psi(y)-c(x,y)}{\epsilon}} d\mu(x)\,d\nu(y) + \epsilon\end{aligned}$$

1 Discrete Optimal Transport

Let X and Y be two compact subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively and consider two finite discrete probability measures

$$\mu = \sum_{i=1}^{n} \mu_i \delta_{x_i} \qquad \text{and} \qquad \nu = \sum_{j=1}^{m} \nu_j \delta_{y_j}$$

representing as. a finite sum of nonnegative weighted Dirac deltas that sum to one and with respective locations $(x_1, \ldots, x_n) \in \mathbb{R}^n$ and $(y_1, \ldots, y_m) \in \mathbb{R}^m$ where δ_x is the Dirac measure at position x i.e.,

$$\delta_a(A) := \begin{cases} 1 & \text{if} \quad a \in \text{A} \\ 0 & \text{otherwise} \end{cases}$$

To sum up, we have points $(x_i)_i$ and $(y_j)_j$ and weights $\mu_i \geq 0$ and $\nu_j \geq 0$ such that $\sum_{i=1}^{n} \mu_i = 1 = \sum_{j=1}^{m} \nu_j$ that live in a simplex Σ_n and Σ_m with respective dimensions n and m where $\Sigma_n := \{\mu_i \in \mathbb{R}_+^n : \sum_{i=1}^{n} \mu_i = 1\}$

1.1 The Kantorovich Formulation

The discrete OT problem between two discrete distributions $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$ and a given continuous cost function $c : X \times Y \longrightarrow [0, +\infty]$ is the following minimization problem

$$\begin{aligned} &\min \sum_{i=1}^{n} \sum_{j=1}^{m} c_{i,j} \pi_{i,j} \\ &\text{subject to:} \\ &\sum_{i=1}^{n} \pi_{i,j} = \nu_j, \quad \sum_{j=1}^{m} \pi_{i,j} = \mu_i \ \text{ and } \ \pi_{i,j} \geq 0. \end{aligned}$$

To sum up, $c_{i,j}$ is the cost of moving a unit of mass from x_i to y_j and the coupling $\pi_{i,j}$ representing how much mass moves from x_i to y_j where the conservation of mass is satisfied, i.e., the sum on rows (columns) must exactly matched the μ (ν) values, i.e., $\sum_{i=1}^{n} \pi_{i,j} = \nu_j$ and $\sum_{j=1}^{m} \pi_{i,j} = \mu_i$.

Among all the possible coupling we are looking for the one that minimize the total cost we spend in order to transport a mass from distribution μ to distribution ν.

Equivalently, we can rewrite the original OT problem as a large-scale finite dimensional linear programming

$$L(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \langle c\ ,\ \pi\rangle, \tag{4}$$

where $\Pi(\mu,\nu)$ is the set of matrices or coupling with marginals μ and ν which is compact and convex polytope

$$\Pi(\mu,\nu) := \Big\{\pi \in \mathbb{R}_+^{n\times m} \mid \pi\mathbb{1}_m = \mu \text{ and } \pi^\top\mathbb{1}_n = \nu\Big\},$$

with $\mathbb{1}_n = (1,\ldots,1)^\top$. For the OT problem between two discrete distributions with equal size $N = \max\{m,n\}$ the linear programming problem (4) has complexity of order $\mathcal{O}(N^3)$ 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 solution of Kantorovich formulation of Optimal transport with lower computational complexity and easy implementation.

2 Entropic Regularization Of Discrete OT Problems

In this approach, one does not solve the original optimal transport problem (4) exactly, but instead replaces it with a regularized problem involving the entropy of the transport plan. The discrete entropy of a coupling matrix is defined as

$$E(\pi) := \sum_{i,j} \pi_{i,j}\Big(\log(\pi_{i,j}) - 1\Big) < 0,$$

where the function E is strongly convex. The idea of the entropic regularization of Optimal transport is to use the negative entropy E as a regularizing function to obtain approximate solutions to (4). The regularized problem is

$$L^\epsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \langle c\ ,\ \pi\rangle + \epsilon E(\pi) \tag{5}$$

First Observation: Adding the entropy of the transport plan makes the problem (5) strongly convex and smooth. Therefore, the regularized problem has a unique solution π^ϵ.

As $\epsilon \to 0$, the solution of $L^\epsilon(\mu,\nu)$ converges to a transport plan that actually minimize the original problem (4). We know (4) may have no unique solution, thus from all solutions (5) picks the one with minimal entropy, that is actually logical as we want to minimize $\langle c\ ,\ \pi\rangle + \epsilon E(\pi)$, so we are going to make entropy as small as possible.

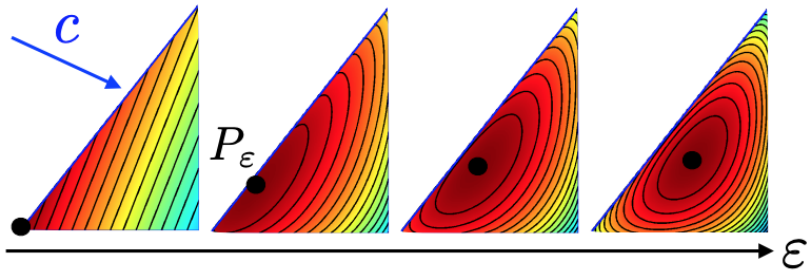

Figure 1: Impact of ϵ on the optimization of a linear function on the simplex, solving $\pi^\epsilon = \arg\min_{\pi\in\Sigma_3}\langle c\ ,\ \pi\rangle + \epsilon E(\pi)$ for varying ϵ. From Peyré and Cuturi: Computational Optimal Transport

Figure 1 illustrates the effect of the entropy to regularize a linear program over the simplex Σ_3. We have triangular feasible region, the cost where red is better and where blue is worst. We are looking for the sweetest spot in this permissible region to solve a linear programming problem that has cost vector c. The sweetest spot given the first feasible region and the given cost vector is the left corner spot, which is the original LP optimum solution. The solutions are always at a corner of polytope or at most on the face of a polytope.

Now one can add a negative entropy as a regularizer to the original objective function and we get a mixed objective function that will have an optimum which no longer lies on the corner of the polytope, the entropy pushes the original solution away from the boundary and the optimal π^ϵ moves progressively toward the entropic center of the triangle. This is further detailed in the theorem below.

2.1 The Main Theorem: Convergence in ϵ

Theorem 2.1. The unique solution π^ϵ to (5) converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is $\pi^\epsilon = \arg\min_{\pi\in\Pi(\mu,\nu)}\langle c\ ,\ \pi\rangle + \epsilon E(\pi)$ is uniquely defined and we claim

$$\lim_{\epsilon\to 0}\pi^\epsilon = \arg\min_{\pi\in\Pi(\mu,\nu)}\left\{E(\pi)\mid \langle c\ ,\ \pi\rangle = L(\mu,\nu)\right\} \tag{6}$$

In particular

$$\lim_{\epsilon\to 0} L^\epsilon(\mu,\nu) = L(\mu,\nu).$$

Proof. Consider a sequence $(\epsilon_n)_n$ such that $\epsilon_n \to 0$ with $\epsilon_n > 0$. The regularized problem (5) is a strictly convex problem and has a unique solution as the objective function is strictly convex and the constraint set is compact.

Denote π^{ϵ_n} the solution of (5) for $\epsilon = \epsilon_n$. Using the fact that $\Pi(\mu,\nu)$ the set of probability measures is compact, then we can extract a converging subsequence

(using sub-sequence principle and for the sake of simplicity we do not relabel, but we can take $\pi^{\epsilon_{n_k}}$ instead) such that $\pi^{\epsilon_n} \to \pi^* \in \Pi(\mu, \nu)$.

Now take π^* and prove it is optimal for (4), then choosing any (optimal and feasible) $\pi \in \Pi(\mu, \nu)$ such that $\langle c\ ,\ \pi \rangle = L(\mu, \nu)$ and compare it to π^*.

By optimality we should get that

$$0 \leq \langle c\ ,\ \pi^{\epsilon_n} \rangle - \langle c\ ,\ \pi \rangle \leq \epsilon_n \Big(E(\pi) - E(\pi^{\epsilon_n}) \Big).$$

Indeed, as π minimize (4) we get $0 \leq \langle c\ ,\ \pi^{\epsilon_n} \rangle - \langle c\ ,\ \pi \rangle$ and as π^{ϵ_n} minimize (5) we get $\langle c\ ,\ \pi^{\epsilon_n} \rangle + \epsilon_n E(\pi^{\epsilon_n}) \leq \langle c\ ,\ \pi \rangle + \epsilon_n E(\pi))$. Since E is continuous, taking the limit as $n \to \infty$ we get $E(\pi^{\epsilon_n}) \to E(\pi^*)$ hence, $E(\pi^{\epsilon_n})$ is bounded, but $E(\pi)$ is constant, then we obtain

$$0 = \lim_{n \to \infty} \langle c\ ,\ \pi^{\epsilon_n} \rangle - \langle c\ ,\ \pi \rangle = \langle c\ ,\ \pi^* \rangle - \langle c\ ,\ \pi \rangle.$$

Thus, $\langle c\ ,\ \pi^* \rangle = \langle c\ ,\ \pi \rangle$ and $\langle c\ ,\ \pi^* \rangle$ is the optimal value and $\langle c\ ,\ \pi^* \rangle = L(\mu, \nu)$. So, π^* is optimal for original problem (4) and is minimizing entropy among all other transport plans, where

$$E(\pi) \geq E(\pi^{\epsilon_n}) \to E(\lim_{n \to \infty} \pi^{\epsilon_n}) = E(\pi^*).$$

This shows that, π^* is a solution of (6). By strict convexity of E, (6) has a unique solution and the whole sequence is converging to π^*. Since, $\forall \epsilon_n > 0, \pi^{\epsilon_n} \to \pi^*$ as $\epsilon_n \to 0$. Then indeed as $\epsilon \to 0$, we obtain $\pi^\epsilon \to \pi^*$. Therefore,

$$\lim_{\epsilon \to 0} L^\epsilon(\mu, \nu) = \lim_{\epsilon \to 0} \langle c\ ,\ \pi^\epsilon \rangle + \epsilon E(\pi^\epsilon) = \langle c\ ,\ \pi^* \rangle = L(\mu, \nu)$$

□

Note that, one also has

$$\lim_{\epsilon \to \infty} \pi^\epsilon = \mu \otimes \nu = \mu\, \nu^\top = (\mu_i \nu_j)_{i,j}. \tag{7}$$

That means for a large regularization, the solution converges to the coupling with minimal entropy between two prescribed marginals μ and ν.

Figures 2, 3 and 4 show visually the effect of the above convergences. A key insight is that, as ϵ increases, the optimal coupling becomes less and less sparse, which in turn has the effect of both accelerating computational algorithms and leading to faster statistical convergence.

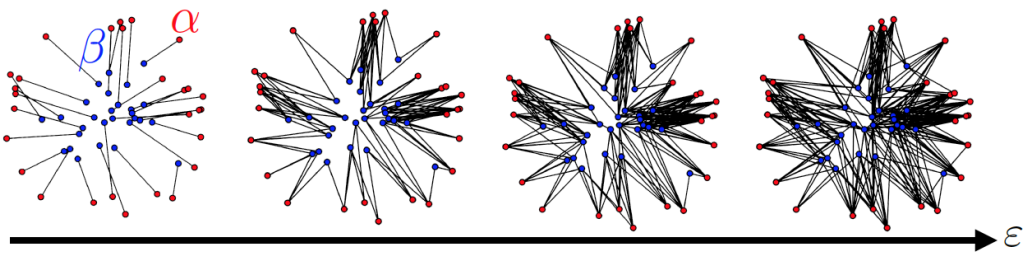

Figure 2: Impact of ϵ on the coupling of between two 2-D discrete densities with the same number $n = m$ of points. From Peyré and Cuturi: Computational Optimal Transport

In figure 2 As $\epsilon \to 0$ we recover the OT. Take two sets of points with the same size, actually as $\epsilon \to \infty$ more segments appear, when temperature increases we get some diffusion and we get more and more points connecting together.

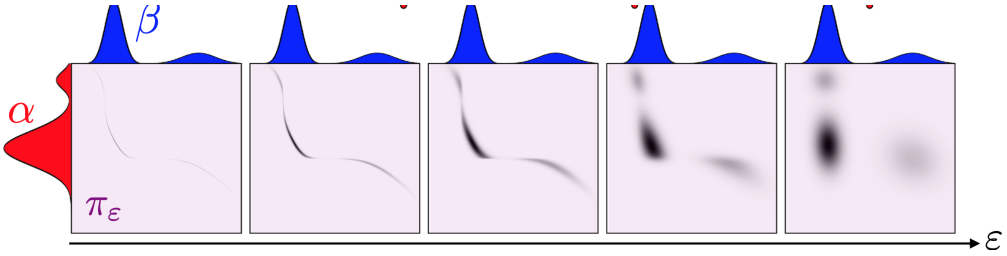

Figure 3: Impact of ϵ on the couplings between two 1-D densities. From a Schrödinger tour of data Sciences. Gabriel Peyré

In figure 3 if we think now about continuous densities, two very nice 1-D densities. The solution of OT is also somehow deterministic map, the coupling we get is very sparse located along the 1-D function. As ϵ increases the solution would be somehow more and more blurry, hard to see and in the limits we get something very blurry where the mass transported from everyone to everyone.

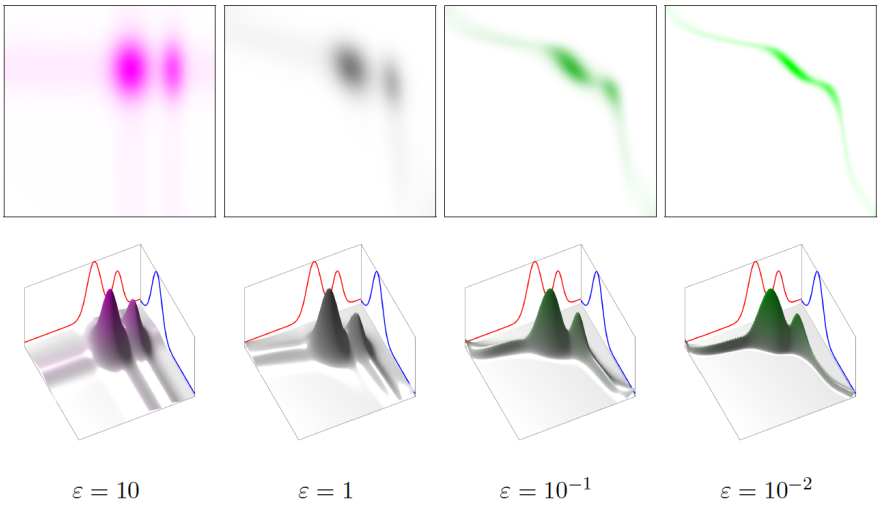

Figure 4: Impact of ϵ on the couplings between two 1-D densities. From Peyré and Cuturi: Computational Optimal Transport

In figure 4 we have two marginals red and blue. We can see that the π^ϵ which here is chosen with the quadratic cost in $\mathbb{R}$, so the solution looks pretty much like the product measure. Now, as $\epsilon \to 0$ we see the solution starts to concentrate more and more along a line. Indeed, in a limit the line we would see here is the graph of OT, in this case there is a unique OT map.

2.2 Projection Problem and KL Divergence

Second Observation: We can interpret π^ϵ as the projection onto $\Pi(\mu,\nu)$ of a very simple coupling. Here the coupling we want to start with. Consider the Gibb's distribution applied to the cost matrix c

$$\mathcal{K}_{i,j} := e^{-\frac{c_{i,j}}{\epsilon}}.$$

If the cost is very small we are happy to move a lot of mass. Otherwise, we don't need to move mass from i to j.

We need to project this onto $\Pi(\mu,\nu)$, we need distance between $\mathcal{K}_{i,j}$ and $\Pi(\mu,\nu)$. The special here is the entropy term, we want to minimize "relative entropy" between $\mathcal{K}$ and $\Pi(\mu,\nu)$. Here is one way of measuring weather two distributions are close to each other.

Definition 2.2. The Kullback-Leibler divergence between couplings or plans π and $\mathcal{K}$ is defined as

$$KL(\pi \mid \mathcal{K}) = \sum_{i,j} \pi_{i,j} \log\left(\frac{\pi_{i,j}}{\mathcal{K}_{i,j}}\right) - \pi_{i,j} + \mathcal{K}_{i,j}. \tag{8}$$

Remark 2.3. $KL(\pi \mid \mathcal{K})$ is minimized if $\pi = \mathcal{K}$. Restrict to one component

$$f(\pi,\mathcal{K}) = \pi \log\left(\frac{\pi}{\mathcal{K}}\right) - \pi + \mathcal{K}.$$

Then, $f_\pi = \log\left(\frac{\pi}{\mathcal{K}}\right), f_\mathcal{K} = 1 - \frac{\pi}{\mathcal{K}}$. $\nabla f = 0$ implies $\pi = \mathcal{K}$. If we want to characterize the minimizer, we have to look for the Hessian:

$$f_{\pi\pi} = \frac{1}{\pi},\ f_{\mathcal{K}\mathcal{K}} = \frac{\pi}{\mathcal{K}^2} \text{ and } f_{\pi\mathcal{K}} = f_{\mathcal{K}\pi} = -\frac{1}{\mathcal{K}}$$

and $|H_f(\pi,\mathcal{K})| = 0$ and $\text{trace}(H_f(\pi,\mathcal{K})) > 0$ then, H_f is positive semidefinite and f is minimized when $\pi = \mathcal{K}$ and the minimum value is zero.

Claim: π^ϵ is indeed the arg min over all feasible transport plans

$$\pi^\epsilon = \underset{\pi\in\Pi(\mu,\nu)}{\arg\min}\, KL(\pi \mid \mathcal{K}).$$

We take our Gibbs distribution which is super simple coupling, we find feasible plan that is closest possible to it, where closest possible is defined by KL divergence

$$\pi^\epsilon = \text{Proj}^{KL}_{\Pi(\mu,\nu)}(\mathcal{K}) := \underset{\pi\in\Pi(\mu,\nu)}{\arg\min}\, KL(\pi \mid \mathcal{K}).$$

Proof. We know that $\pi^\epsilon \in \Pi(\mu,\nu)$. Take any other feasible transport plan π and substitute in KL and get larger value. Let π be a feasible transport plan then,

$$
\begin{aligned}
0 &\leq \frac{1}{\epsilon}\Bigg[\Big(\langle c\ ,\ \pi\rangle + \epsilon E(\pi)\Big) - \Big(\langle c\ ,\ \pi^\epsilon\rangle + \epsilon E(\pi^\epsilon)\Big)\Bigg] \\
&= \sum_{i,j}\left[\frac{c_{i,j}\pi_{i,j}}{\epsilon} + \pi_{i,j}\Big(\log(\pi_{i,j}) - 1\Big) - \frac{c_{i,j}\pi^\epsilon_{i,j}}{\epsilon} - \pi^\epsilon_{i,j}\Big(\log(\pi^\epsilon_{i,j}) - 1\Big)\right] \\
&= \sum_{i,j}\Bigg[\left(- \pi_{i,j}\log\left(e^{-\frac{c_{i,j}}{\epsilon}}\right) + \pi_{i,j}\log(\pi_{i,j}) - \pi_{i,j}\right) \\
&- \left(- \pi^\epsilon_{i,j}\log\left(e^{-\frac{c_{i,j}}{\epsilon}}\right) + \pi^\epsilon_{i,j}\log(\pi^\epsilon_{i,j}) - \pi^\epsilon_{i,j}\right)\Bigg] \\
&= \sum_{i,j}\left[\pi_{i,j}\log\left(\frac{\pi_{i,j}}{e^{-\frac{c_{i,j}}{\epsilon}}}\right) - \pi_{i,j}\ -\ \pi^\epsilon_{i,j}\log\left(\frac{\pi^\epsilon_{i,j}}{e^{-\frac{c_{i,j}}{\epsilon}}}\right) + \pi^\epsilon_{i,j}\right] \\
&= \sum_{i,j}\left[\left(\pi_{i,j}\log\left(\frac{\pi_{i,j}}{e^{-\frac{c_{i,j}}{\epsilon}}}\right) - \pi_{i,j} + \mathcal{K}_{i,j}\right)\ -\ \left(\pi^\epsilon_{i,j}\log\left(\frac{\pi^\epsilon_{i,j}}{e^{-\frac{c_{i,j}}{\epsilon}}}\right) - \pi^\epsilon_{i,j} + \mathcal{K}_{i,j}\right)\right] \\
&= KL(\pi \mid \mathcal{K})\ -\ KL(\pi^\epsilon \mid \mathcal{K})
\end{aligned}
$$

This yields,

$$KL(\pi^\epsilon \mid \mathcal{K})\ \leq\ KL(\pi \mid \mathcal{K}) \quad \forall\ \pi \in \Pi(\mu,\nu).$$

Therefore,

$$\pi^\epsilon = \underset{\pi\in\Pi(\mu,\nu)}{\arg\min}\, KL(\pi \mid \mathcal{K})$$

□

2.3 Sinkhorn's Algorithm

Characterizing π involves $n \times m$ unknowns, which is too expensive. Let us look at our Optimal problem:

$$\min_{\pi\in\Pi(\mu,\nu)} \langle c\ ,\ \pi\rangle + \epsilon E(\pi),$$

where,

$$\Pi(\mu,\nu) := \Big\{\pi \in \mathbb{R}^{n\times m}_+ \mid \pi\mathbb{1}_m = \mu \text{ and } \pi^\top\mathbb{1}_n = \nu\Big\}$$

and

$$E(\pi) := \sum_{i,j}\pi_{i,j}\Big(\log(\pi_{i,j}) - 1\Big).$$

Introducing two dual variables $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$ for each marginal constraints, the Lagrangian of our problem reads

$$\mathcal{L}(\pi, \alpha, \beta) = \langle c\ ,\ \pi \rangle + \epsilon E(\pi) + \alpha^\top(\pi \mathbb{1}_m - \mu) + \beta^\top(\pi^\top \mathbb{1}_n - \nu)$$

Karush-Kuhn-Tucker theorem then yield,

$$\nabla_\pi \mathcal{L}(\pi^\epsilon,\ \alpha,\ \beta) = c_{i,j} + \epsilon \log(\pi^\epsilon_{i,j}) + \alpha_i + \beta_j = 0,$$

which result,

$$\begin{aligned} \pi^\epsilon_{i,j} &= e^{-\frac{c_{i,j}}{\epsilon}}\, e^{-\frac{\alpha_i}{\epsilon}}\, e^{-\frac{\beta_j}{\epsilon}} \\ &= \mathcal{K}_{i,j}\, u_i\, v_j \end{aligned} \tag{9}$$

where, $u \in \mathbb{R}^n_+$ and $v \in \mathbb{R}^m_+$. This characterizes the optimal regularize plan π^ϵ using $n + m$ unknowns. Rewriting (9) in matrix form yields,

$$\pi^\epsilon = \operatorname{diag}(u)\, \mathcal{K}\, \operatorname{diag}(v). \tag{10}$$

We require, $\pi \mathbb{1}_m = \mu$ and $\pi^\top \mathbb{1}_n = \nu$ which implies,

$$\operatorname{diag}(u)\, \mathcal{K} v = \mu \quad \text{and} \quad \operatorname{diag}(v) \mathcal{K}^\top u = \nu,$$

equivalently, we obtain

$$u \odot (\mathcal{K} v) = \mu \qquad \text{and} \qquad v \odot (\mathcal{K}^\top u) = \nu, \tag{11}$$

where $\odot$ corresponds to entrywise multiplication of vectors.

Computing π^ϵ has been reduced to the problem of finding u and v satisfying (11). We need an algorithm for doing this, the Newton method has disadvantage that we have to compute the Jacobian. The easiest fixed point iteration is:

$$u^{(k+1)} = \frac{\mu}{\mathcal{K}\, v^{(k)}} \qquad \text{and} \qquad v^{(k+1)} = \frac{v}{\mathcal{K}^\top\, u^{(k+1)}}, \tag{12}$$

we modify first u so that it satisfies the left-hand-side of (12) and then v to satisfy the right-hand-side. The division operator used in (12) between two vectors is to be understood entrywise. These two updates define **Sinkhorn's algorithm** initialized with $v^{(0)} = \mathbb{1}_m$.

Advantages:

- Number of unknowns reduced from $n \times m$ to $n + m$
- Easy to code and parallelize
- Just have to do matrix-vector products involving Gibbs distribution
- Yields to differentiable approximation of OT
- Complexity of each iteration is $\mathcal{O}(N^2)$
- A large enough regularization breaks the curse of dimension

Disadvantages:

- No transport map
- Convergence rate deteriorates as $\epsilon \to 0$ and has dependence on the dimension
- Restricted to "large" ϵ and low accuracy

Remark 2.4. When we talk about entropy we have to speak with respect to reference measure, but in Optimal transportation naturally we have the product measure $\mu \otimes \nu$ of two input distributions $\mu \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^m$. The entropy term E modifies the linear term in classical optimal transportation to produce a strictly convex functional. This is not the only possible choice. Alternatively, we would fix two reference probability measures and consider

$$\begin{aligned} L_r^{\epsilon}(\mu,\nu) &= \min_{\pi\in\Pi(\mu,\nu)} \langle c\ ,\ \pi\rangle + \epsilon KL(\pi \mid \mu\otimes\nu) \\ &= \min_{\pi\in\Pi(\mu,\nu)} \sum_{i,j} c_{i,j}\pi_{i,j} \ +\ \epsilon\left[\sum_{i,j}\pi_{i,j}\log\left(\frac{\pi_{i,j}}{\mu_i\,\nu_j}\right) - \pi_{i,j} + \mu_i\,\nu_j\right] \end{aligned} \tag{13}$$

where, $\mu\otimes\nu = \mu\,\nu^{\top} = (\mu_i\,\nu_j)_{i,j}$ *and* $KL(\pi \mid \mu\otimes\nu)$ *is the* ***relative entropy*** *or the* ***Kullback-Leibler divergence****.*

3 Entropic Regularization Of Continuous OT Problems

Let X and Y be two compact subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let $\mu \in \mathcal{P}(X)$, $\nu \in \mathcal{P}(Y)$ and $c(x,y)$ be a non-negative, symmetric and Lipschitz continuous function. Consider the Kantorovich problem of optimal transport

$$\min_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\,d\gamma(x,y) \tag{14}$$

where $\Pi(\mu,\nu)$ denotes the set of joint probability measures γ on $X\times Y$ with marginals μ and ν. Thus, in the continuous case, the transport plan γ is a probability distribution on $X\times Y$ while in the discrete case it is a matrix. Furthermore, the cost function c represents the cost to move a unit of mass from x to y, and (14) represents the total cost of moving all mass from μ to ν.

3.1 General Formulation

The choice of the reference measures is arbitrary. One can consider arbitrary measures by replacing the **discrete entropy** by the **relative entropy** with respect to product measure

$$d\mu\otimes d\nu(x,y) = d\mu(x)\,d\nu(y)$$

and propose a regularized problem to (14) where the problem (5) can be re-written as

$$\min_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\gamma(x,y) \;+\; \epsilon\, KL(\gamma \mid \mu\otimes\nu), \tag{15}$$

where γ is absolutely continuous with respect to $\mu\otimes\nu$ and we write $\gamma \ll \mu\otimes\nu$, i.e., if for every $A\times B \in \mathcal{B}(X\times Y)$ with $\mu\otimes\nu(A\times B) = 0 \Longrightarrow \gamma(A\times B) = 0$, and the **relative entropy** is a generalization of the discrete Kullback-Leibler divergence (8)

$$\begin{aligned} KL(\gamma \mid \xi) := &\int_{X\times Y} \log\left(\frac{d\gamma}{d\xi}(x,y)\right) d\gamma(x,y) \\ &- \int_{X\times Y} d\gamma(x,y) \;+\; \int_{X\times Y} d\xi(x,y), \end{aligned} \tag{16}$$

and by convention $KL(\gamma \mid \xi) < \infty$ if $\gamma \ll \xi$ i.e., when γ has a density $\dfrac{d\gamma}{d\xi}$ with respect to ξ. However, $KL(\gamma \mid \xi) = +\infty$ if both measures do not share the same support, which causes discontinuity issues.

Proposition 3.1. For any $\gamma \in \Pi(\mu,\nu)$ and for any (α,β) so that they have both densities with respect to one another i.e., μ and α being matually absolutely continuous with respect to each other and similarly for ν and β. one has

$$KL(\gamma \mid \mu\otimes\nu) = KL(\gamma \mid \alpha\otimes\beta) - KL(\mu\otimes\nu \mid \alpha\otimes\beta). \tag{17}$$

Rearranging (17) we get

$$KL(\gamma \mid \alpha\otimes\beta) - KL(\gamma \mid \mu\otimes\nu) = KL(\mu\otimes\nu \mid \alpha\otimes\beta)$$

and we see that the difference does not depend on γ. The reference measure $\mu\otimes\nu$ chosen in (15) to define the entropic regularizing term $KL(\gamma \mid \mu\otimes\nu)$ plays no role; only its support matters. In particular the minimizer, if it exists, does not depend on the choice of μ and ν. Therefore, choosing $KL(\gamma \mid \alpha\otimes\beta)$ in place of $KL(\gamma \mid \mu\otimes\nu)$ in (15) results in the same solution. The minimal value, however, does depend on the choice of the regularization term.

Remark 3.2. Formula (15) can be refactored as a projection problem

$$\min_{\gamma\in\Pi(\mu,\nu)} KL(\gamma \mid \mathcal{K}) \tag{18}$$

where $\mathcal{K}$ is the Gibbs distributions

$$d\mathcal{K}(x,y) := e^{-\frac{c(x,y)}{\epsilon}} d\mu(x) d\nu(y).$$

This problem is often referred to as the "static Schrodinger problem." As $\epsilon \to 0$, the unique solution to (18) converges to the minimum entropy solution to (14).

Proof. Let $\gamma \in \Pi(\mu,\nu)$. As γ^ϵ solves (15), we get $\gamma^\epsilon \in \Pi(\mu,\nu)$ and

$$\begin{aligned} 0 &\leq \frac{1}{\epsilon}\Big[\Big(\langle c\ ,\ \gamma\rangle + \epsilon KL(\gamma \mid \mu\otimes\nu)\Big) - \Big(\langle c\ ,\ \gamma^\epsilon\rangle + \epsilon KL(\gamma^\epsilon \mid \mu\otimes\nu)\Big)\Big] \\ &= \frac{\langle c\ ,\ \gamma\rangle}{\epsilon} + KL(\gamma \mid \mu\otimes\nu) - \frac{\langle c\ ,\ \gamma^\epsilon\rangle}{\epsilon} - KL(\gamma^\epsilon \mid \mu\otimes\nu) \end{aligned}$$

which is equivalent to

$$\begin{aligned} &= \frac{1}{\epsilon}\int_{X\times Y} c(x,y)\, d\gamma(x,y) + \int_{X\times Y} \log\Big(\frac{d\gamma}{d\mu(x)d\nu(y)}(x,y)\Big)\, d\gamma(x,y) \\ &- \int_{X\times Y} d\gamma(x,y)\ +\ \int_{X\times Y} d\mu(x)d\nu(y) - \frac{1}{\epsilon}\int_{X\times Y} c(x,y)\, d\gamma^\epsilon(x,y) \\ &- \int_{X\times Y} \log\Big(\frac{d\gamma^\epsilon}{d\mu(x)d\nu(y)}(x,y)\Big)\, d\gamma^\epsilon(x,y) + \int_{X\times Y} d\gamma^\epsilon(x,y)\ -\ \int_{X\times Y} d\mu(x)d\nu(y) \end{aligned}$$

writing the cost function in logarithm form, we obtain

$$\begin{aligned} &= -\int_{X\times Y} \log\Big(e^{-\frac{c(x,y)}{\epsilon}}\Big) d\gamma + \int_{X\times Y} \log\Big(\frac{d\gamma}{d\mu(x)d\nu(y)}(x,y)\Big)\, d\gamma(x,y) \\ &- \int_{X\times Y} d\gamma(x,y)\ +\ \int_{X\times Y} d\mu(x)d\nu(y) + \int_{X\times Y} \log\Big(e^{-\frac{c(x,y)}{\epsilon}}\Big) d\gamma^\epsilon \\ &- \int_{X\times Y} \log\Big(\frac{d\gamma^\epsilon}{d\mu(x)d\nu(y)}(x,y)\Big)\, d\gamma^\epsilon(x,y) + \int_{X\times Y} d\gamma^\epsilon(x,y)\ -\ \int_{X\times Y} d\mu(x)d\nu(y) \end{aligned}$$

adding and subtracting the blue term yields

$$\begin{aligned} &= \int_{X\times Y} \log\left[\frac{d\gamma}{d\mu(x)d\nu(y)}(x,y)\ \times\ \frac{1}{e^{-\frac{c(x,y)}{\epsilon}}}\right] d\gamma(x,y) + \int_{X\times Y} e^{-\frac{c(x,y)}{\epsilon}} d\mu(x)d\nu(y) \\ &- \int_{X\times Y} d\gamma(x,y) - \int_{X\times Y} \log\left[\frac{d\gamma^\epsilon}{d\mu(x)d\nu(y)}(x,y)\ \times\ \frac{1}{e^{-\frac{c(x,y)}{\epsilon}}}\right] d\gamma^\epsilon(x,y) \\ &- \int_{X\times Y} e^{-\frac{c(x,y)}{\epsilon}} d\mu(x)d\nu(y) + \int_{X\times Y} d\gamma^\epsilon(x,y) \end{aligned}$$

re-arranging into two similar forms yields,

$$\begin{aligned} &= \int_{X\times Y} \left[\log\left(\frac{d\gamma}{e^{-\frac{c(x,y)}{\epsilon}}\, d\mu(x)d\nu(y)}(x,y)\right) d\gamma(x,y) + \left(e^{-\frac{c(x,y)}{\epsilon}} d\mu(x)d\nu(y) - d\gamma(x,y)\right)\right] \\ &- \int_{X\times Y} \left[\log\left(\frac{d\gamma^\epsilon}{e^{-\frac{c(x,y)}{\epsilon}}\, d\mu(x)d\nu(y)}(x,y)\right) d\gamma^\epsilon(x,y) + \left(e^{-\frac{c(x,y)}{\epsilon}} d\mu(x)d\nu(y) - d\gamma^\epsilon(x,y)\right)\right] \end{aligned}$$

using Gibbs distribution to obtain the simplified form

$$= \int_{X\times Y} \log\left(\frac{d\gamma}{d\mathcal{K}(x,y)}(x,y)\right) d\gamma(x,y) + \int_{X\times Y}\left(d\mathcal{K}(x,y) - d\gamma(x,y)\right)$$
$$- \left[\int_{X\times Y} \log\left(\frac{d\gamma^\epsilon}{d\mathcal{K}(x,y)}(x,y)\right) d\gamma^\epsilon(x,y) + \int_{X\times Y}\left(d\mathcal{K}(x,y) - d\gamma^\epsilon(x,y)\right)\right]$$

which is equal to,

$$KL(\gamma \mid \mathcal{K}) - KL(\gamma^\epsilon \mid \mathcal{K})$$

Therefore,

$$KL(\gamma^\epsilon \mid \mathcal{K}) \leq KL(\gamma \mid \mathcal{K}) \quad \forall\, \gamma \in \Pi(\mu,\nu).$$

□

3.2 Dual Formulation of Continuous EOT

An advantage to consider regularized OT is to get an unconstrained dual problem. The dual of standard OT (14) reads:

$$DP = \max_{(\phi,\psi)\in\mathcal{U}(c)} \int_X \phi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y), \tag{19}$$

where the constraint set $\mathcal{U}(c)$ is defined by

$$\mathcal{U}(c) := \Big\{(\phi,\psi) \in C(X)\times C(Y) \mid \phi(x) + \psi(y) \leq c(x,y), \forall (x,y) \in X\times Y\Big\},$$

where $\phi : X \longrightarrow \mathbb{R}$ and $\psi : Y \longrightarrow \mathbb{R}$ are the Lagrange multipliers.

Consider the regularized OT problem

$$Pr_\epsilon = \min_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} c(x,y)\, d\gamma(x,y) \;+\; \epsilon\, KL(\gamma \mid \mu\otimes\nu), \tag{20}$$

where the relative entropy with $\gamma \ll \mu\otimes\nu$ is defined as follows

$$KL(\gamma \mid \mu\otimes\nu) := \int_{X\times Y}\left[\log\Big(\frac{d\gamma(x,y)}{d\mu(x)d\nu(y)}\Big) - 1\right] d\gamma(x,y),$$

The dual of the regularized OT is given by an unconstrained maximization problem where the constraint we had in OT problem becomes smooth here and replaced by exponential penalty

$$DP_\epsilon = \max_{(\phi,\psi)\in C(X)\times C(Y)} \int_X \phi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y)$$
$$- \epsilon \int_{X\times Y} e^{\frac{\phi(x)+\psi(y)-c(x,y)}{\epsilon}} d\mu(x)\, d\nu(y) \tag{21}$$

3.3 Strong Duality

The constraints are relaxed by associating a penalty parameters, which we call the Lagrange multipliers ϕ and ψ with each of the $m+n$ equality-constraints. The Lagrangian dual function reads

$$\mathcal{L}(\gamma,\phi,\psi) = \int_{X\times Y} c(x,y)\,d\gamma(x,y) \;+\; \epsilon \int_{X\times Y}\left[\log\Big(\frac{d\gamma}{d\mu(x)d\nu(y)}\Big) - 1\right] d\gamma(x,y)$$

$$+ \int_X \phi(x)\Big[d\mu(x) - \int_Y d\gamma(x,y)\Big] + \int_Y \psi(y)\Big[d\nu(y) - \int_X d\gamma(x,y)\Big]. \tag{22}$$

Then, we can write primal problem Pr_ϵ as $\min - \max$ of the Lagrangian

$$Pr_\epsilon = \min_{\gamma}\max_{\phi,\psi} \mathcal{L}(\gamma,\phi,\psi)$$

and the unconstrained dual problem

$$\begin{aligned} Pr_\epsilon \geq DP_\epsilon &= \max_{\phi,\psi}\min_{\gamma} \mathcal{L}(\gamma,\phi,\psi) && (23)\\ &= \max_{(\phi,\psi)\in C(X)\times C(Y)} \int_X \phi(x)\,d\mu(x) + \int_Y \psi(y)\,d\nu(y)\\ &\quad - \epsilon\int_{X\times Y} e^{\frac{\phi(x)+\psi(y)-c(x,y)}{\epsilon}} d(\mu\otimes\nu). && (24)\end{aligned}$$

Strong duality holds, thanks to the application of **Fenchel-Rockafellar** theorem to the dual problem, where X and Y are compact spaces, hence $Pr_\epsilon = DP_\epsilon$ and therefore, if optimal dual solutions $\phi*$ and $\psi*$ exist, they are related to the optimal transport plan $\gamma*$ by

$$d\gamma * (x,y) = e^{\frac{\phi*(x)+\psi*(y)-c(x,y)}{\epsilon}} d\mu(x)d\nu(y).$$

Remark 3.3. For the regularized optimal problem (20) if the relative entropy with $\gamma \ll \mu\otimes\nu$ is defined as

$$KL(\gamma \mid \mu\otimes\nu) := \int_{X\times Y} \log\Big(\frac{d\gamma(x,y)}{d\mu(x)d\nu(y)}\Big)\,d\gamma(x,y),$$

then the unconstrained dual formulation of (20) is given by

$$\begin{aligned} DP_\epsilon = &\max_{(\phi,\psi)\in C(X)\times C(Y)} \int_X \phi(x)\,d\mu(x) + \int_Y \psi(y)\,d\nu(y)\\ &- \epsilon\int_{X\times Y} e^{\frac{\phi(x)+\psi(y)-c(x,y)}{\epsilon}} - 1 d(\mu\otimes\nu). \end{aligned} \tag{25}$$

Furthermore, if optimal dual solutions $\hat\phi$ and $\hat\psi$ exist, they are related to the optimal transport plan $\hat\gamma$ by

$$\hat\gamma = \Big(e^{\frac{\hat\phi(x)+\hat\psi(y)-c(x,y)}{\epsilon}}\Big)\mu\otimes\nu.$$

Conclusion

The Linear programming problem may have no unique solution and computing the optimum is costly, where the complexity turns out to be super cubic which actually means that it is infeasible for large number of samples. To overcome this issue, regularization methods have been proposed to approximate the original problem by adding a multiple of the negative entropy of the transport plan that forces the solution to have spread support and makes the problem strongly convex and smooth.

The regularized problem has a unique solution that converge to the optimal solution minimizing the original problem with minimal entropy. The minimization is achieved numerically by a simple Sinkhorn algorithm allowing us to compute the EOT with cost that is quadratic in the number of samples. Furthermore, the solution can always be factored through Gibbs kernel and instead of looking for big matrix we are looking for two scalars, that is the number of unknown reduced from $m \times n$ to $m + n$.

A large enough regularization breaks the curse of dimension, which in turn has the effect of both accelerating the computational algorithm and leads to faster statistical convergence. Besides, the optimal coupling becomes not sparse and will be very stable, very good in high dimensions. On the other hand, the optimal coupling can be interpreted as the projection of Gibbs kernel onto the set of transport plans, that is we take our Gibbs distribution and we find feasible plan that is closest possible to it.

Finally, for EOT dual we don't have constraints like in OT, the constraints we had in OT becomes smooth and replaced by exponential penalty. Therefore, we get an unconstrained and smooth dual formulation.

References

[1] Neumayer S., Steidl G. (2021) From Optimal Transport to Discrepancy

[2] Nenna, L., 2020. Lecture 4: Entropic Optimal Transport and Numerics.

[3] Peyré, Gabriel and Cuturi, Marco. (2018) Computational Optimal Transport

[4] Quentin Merigot, Boris Thibert. (2020) Optimal transport: discretization and algorithms

[5] Christian Clason, Dirk A. Lorenz, Hinrich Mahler and Benedikt Wirth. (2021) Entropic regularization of continuous optimal transport problems

[6] Genevay, A., Gabriel Peyré (2019) Entropy-regularized optimal transport for machine learning