Lecture 2
Discrete OT, Entropic regularization and Numerics

Luca Nenna

August 31, 2020

Introduction: Discret Optimal Transport

We now consider the discrete measures

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

Definition 0.1 (Discrete OT)**.** The discrete Optimal transport problem between two given measures μ and ν and a given cost matrix $C_{ij} = c(x_i, y_j)$ is the following minimization problem

$$\inf \left\{ \sum_i \sum_j C_{ij}\gamma_{ij} \mid \gamma \in \Pi(\mu,\nu) \right\}, \tag{0.1}$$

where the set of admissible couplings is now define as

$$\Pi(\mu,\nu) := \{\gamma \in \mathbb{R}^n \times \mathbb{R}^n \mid \gamma_{ij} \geqslant 0,\ \sum_j \gamma_{ij} = \mu_i\ \forall i,\ \sum_i \gamma_{ij} = \nu_j\ \forall j\}.$$

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.

References: Entropic regularisation of Optimal Transport is a very active research field. We refer the interested reader to [1, 4, 8, 9, 5] and the citations therein.

1 The Entropic Optimal Transport

1.1 The discrete case

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

$$P_\varepsilon = \inf \left\{ \sum_{x,y} \gamma_{ij} C_{ij} + \varepsilon \operatorname{Ent}(\gamma) \mid \gamma \in \mathbb{R}^n \times \mathbb{R}^n,\ \sum_j \gamma_{ij} = \mu_i,\ \sum_i \gamma_{ij} = \nu_j \right\} \quad (1.2)$$

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

Theorem 1.1 (Convergence in ε)**.** *The unique solution γ_ε to (1.2) converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is*

$$\gamma_\varepsilon \xrightarrow[\varepsilon \to 0]{} \operatorname{argmin} \left\{ \operatorname{Ent}(\gamma) \mid \gamma \in \Pi(\mu, \nu),\ \sum_{i,j} \gamma_{ij} C_{ij} = \mathcal{MK}_c(\mu, \nu) \right\}. \quad (1.3)$$

Proof. Consider a sequence $(\varepsilon_k)_k$ such that $\varepsilon_k \to 0$ and $\varepsilon_k > 0$ and denote γ_k the solution to(1.2) with $\varepsilon = \varepsilon_k$. Since $\Pi(\mu, \nu)$ is bounded and close we can extract a converging subsequence $\gamma_k \to \gamma^\star \in \Pi(\mu, \nu)$. Take now any optimal γ for the unregularized problem then by optimality of γ_k and γ one has

$$0 \leqslant \langle \gamma_k | c \rangle - \langle \gamma | c \rangle \leqslant \varepsilon_k (\operatorname{Ent}(\gamma) - \operatorname{Ent}(\gamma_k)), \quad (1.4)$$

where $\langle \gamma | c \rangle := \sum_{i,j} \gamma_{ij} C_{ij}$. Since Ent is continuous, by taking the limit $k \to +\infty$ in (1.4) we get $\langle \gamma^\star | c \rangle = \langle \gamma | c \rangle$. Furthermore, dividing by ε_k and taking the limit we obtain that $\operatorname{Ent}(\gamma) \geqslant \operatorname{Ent}(\gamma^\star)$ showing that $\gamma^\star$ is a solution to the minimization problem in (1.3). By strict convexity of Ent the optimization problem (1.3) has a unique solution and the whole sequence is converging to $\gamma^\star$. □

We want now to derive formally the dual problem. For this purpose we introduce the Lagrangian associated to (1.2)

$$\begin{aligned} \mathcal{L}(\gamma, \varphi, \psi) := \sum_{i,j} \gamma_{ij} c(x, y) + \varepsilon e(\gamma_{ij}) + \sum_i \varphi_i \left(\mu_j - \sum_j \gamma_{ij} \right) \\ + \sum_j \psi_i \left(\nu_j - \sum_i \gamma_{ij} \right), \end{aligned} \quad (1.5)$$

where $\varphi \in \mathbb{R}^n$ and $\psi \in \mathbb{R}^n$ 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 :

$$\begin{aligned} D_\varepsilon = \sup_{\varphi, \psi} \min_\gamma \sum_{i,j} \gamma_{ij} (C_{ij} - \psi_j - \varphi_i + \varepsilon(\log(\gamma_{ij}) - 1)) + \\ \sum_i \varphi_i \mu_i + \sum_j \psi_j \nu_j. \end{aligned} \quad (1.6)$$

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

$$\begin{aligned} &C_{ij} - \psi_j - \varphi_i + \varepsilon \log(\gamma_{ij}) = 0, \ \forall (i,j) \\ &\text{i.e. } \gamma_{ij} = \exp(\frac{1}{\varepsilon}(\varphi_i + \psi_j - C_{ij}) \end{aligned} \tag{1.7}$$

Putting these values in the definition of D_ε gives

$$\begin{aligned} &D_\varepsilon = \sup_{\varphi,\psi} \Phi_\varepsilon(\varphi,\psi) \text{ with} \\ &\Phi_\varepsilon(\varphi,\psi) := \sum_i \varphi_i \mu_i + \sum_j \psi_j \nu_j - \sum_{i,j} \varepsilon \exp(\frac{1}{\varepsilon}(\varphi_i + \psi_j - C_{ij})) \end{aligned} \tag{1.8}$$

Note that thanks to the relation (1.7), 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 (1.2) is smooth and strictly convex. The following duality result holds

Theorem 1.2 (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 1.3. *If (φ,ψ) is the solution to (1.8), then the solution $\gamma^\star$ to (1.2) is given by*

$$\gamma_{ij} = e^{\frac{\varphi_i + \psi_j - C_{ij}}{\varepsilon}}$$

Notice now that the optimal coupling γ can be written as

$$\gamma_{ij} = \text{diag}(D_\varphi) e^{\frac{-C_{ij}}{\varepsilon}} \text{diag}(D_\psi),$$

where $\text{diag}(D_\varphi)$ and $\text{diag}(D_\psi)$ are the diagonal matrices associated to $D_\varphi = e^{\varphi/\varepsilon}$ and $D_\psi = e^{\psi/\varepsilon}$, respectively. The problem is now similar to a matrix scaling problem

Definition 1.4 (Matrix scaling problem)**.** Let $K \in \mathbb{R}^{N \times N}$ be a matrix with positive coefficients. Find $\text{diag}(D_\psi)$ and $\text{diag}(D_\psi)$ positive diagonal matrices in $\mathbb{R}^{N \times N}$ such that $\text{diag}(D_\varphi) K \, \text{diag}(D_\psi)$ is doubly stochastic, that is sum along each row and each column is equal to 1.

Remark 1.5. 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).

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

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 $(K_\varepsilon)_{x,y} = \text{diag}(\mu) e^{\frac{-c(x,y)}{\varepsilon}} \text{diag}(\nu)$, where

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**
    $D_\varphi^{k+1} \leftarrow \mathbf{1}_N./(KD_\psi^k)$
    $D_\psi^{k+1} \leftarrow \mathbf{1}_N./(K^T D_\varphi^{k+1})$
  end for
end function

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

function SINKHORN-KNOPP(K_ε, μ, ν)
  $D_\varphi^0 \leftarrow \mathbf{1}_n,\ D_\psi^0 \leftarrow \mathbf{1}_n$
  for $0 \leqslant k < k_{\max}$ **do**
    $D_\varphi^{k+1} \leftarrow \mu./(KD_\psi^k)$
    $D_\psi^{k+1} \leftarrow \nu./(K^T D_\varphi^{k+1})$
  end for
end function

$\mathrm{diag}(\mu)$ ($\mathrm{diag}(\nu)$) denotes the diagonal matrix with the vector μ (ν) as main diagonal. In this case the problem (1.2) can be re-written as

$$P_\varepsilon(\mu,\nu) = \inf\left\{\sum_{i,j}\gamma_{ij}C_{ij} + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma\in\mathbb{R}^n\times\mathbb{R}^n,\ \sum_j\gamma_{ij}=\mu_i,\ \sum_i\gamma_{ij}=\nu_j\right\}, \tag{1.9}$$

where $\mathcal{H}(\rho|\mu) := \sum_i \rho_i(\log(\frac{\rho_i}{\mu_i})-1)$ is the relative entropy or the Kullback-Leibler divergence.

Good to know: one can easily recast the regularized OT in the continuous framework as follows

$$\mathcal{P}_\varepsilon(\mu,\nu) = \inf\left\{\int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma\in\Pi(\mu,\nu)\right\}, \tag{1.10}$$

where

$$\mathcal{H}(\rho|\pi) = \begin{cases}\displaystyle\int_{X\times Y}\Big(\log\Big(\frac{\mathrm{d}\rho(x,y)}{\mathrm{d}\pi(x,y)}\Big)-1\Big)\mathrm{d}\rho(x,y), & \text{if } \rho\ll\pi\\ +\infty, & \text{otherwise,}\end{cases}$$

and the marginals μ,ν are probability measures on the compact metric spaces X and Y, respectively. This problem is often referred to as the *static Schrödinger problem* [8] since it was initially considered by Schrödinger in statistical physics. Once again, under mild assumptions on the cost functions, one can prove that the regularized problem converges to original one as $\varepsilon\to 0$; see [3, 7].

1.2 The convergence of Sinkhorn for the Hilbert metric

We focus now on the global convergence analysis of the Sinkhorn algorithm by using the *Hilbert* projective metric on $\mathbb{R}^n_{+,\star}$ (positive vectors).

Definition 1.6 (Hilbert projective metric)**.** The *Hilbert* projective metric on $\mathbb{R}^n_{+,\star}$ is defined as

$$\forall (u,v) \in (\mathbb{R}^n_{+,\star})^2,\ d_H(u,v) := ||\log(u) - \log(v)||_V,$$

Where

$$||x||_V = \max_i x_i - \min_i x_i.$$

Before stating the convergence result we need the following fundamental theorem, which shows that a positive matrix is a strict contraction on the cone of positive vector

Theorem 1.7 ([2, 10])**.** *Let $K \in \mathbb{R}^{n\times n}_{+,\star}$, then for $(u,v) \in (\mathbb{R}^n_{+,\star})^2$*

$$d_H(Ku, Kv) \leqslant \lambda(K) d_H(u,v), \tag{1.11}$$

where

$$\lambda(K) = \frac{\sqrt{\eta(K)} - 1}{\sqrt{\eta(K)} + 1} < 1$$

and

$$\eta(K) = \max_{i,j,kl} \frac{K_{ik}K_{jl}}{K_{jk}K_{il}}.$$

We have then the following convergence result

Theorem 1.8 ([6])**.** *One has $(D^k_\varphi, D^k_\psi) \to (D^\star_\varphi, D^\star_\psi)$ and*

$$d_H(D^k_\varphi, D^\star_\varphi) = O(\lambda(K)^{2k}),\ d_H(D^k_\psi, D^\star_\psi) = O(\lambda(K)^{2k}), \tag{1.12}$$

where $D^\star_\varphi, D^\star_\psi$ are the optimal solutions. Moreover,

$$d_H(D^k_\varphi, D^\star_\varphi) \leqslant \frac{d_H(\gamma^k \mathbf{1}_n, \mu)}{1 - \lambda(K)^2}, \tag{1.13}$$

$$d_H(D^k_\psi, D^\star_\psi) \leqslant \frac{d_H(\gamma^k \mathbf{1}_n, \nu)}{1 - \lambda(K)^2}, \tag{1.14}$$

where $\gamma^k = \operatorname{diag}(D^k_\varphi) K \operatorname{diag}(D^k_\psi)$. Last, one has

$$||\log(\gamma^k) - \log(\gamma^\star)||_\infty \leqslant d_H(D^k_\varphi, D^\star_\varphi) + d_H(D^k_\psi, D^\star_\psi). \tag{1.15}$$

where $\gamma^\star$ is the unique solution to (1.2).

Proof. Notice that for any $(u,v) \in (\mathbb{R}^n_{+,\star})^2$, on has

$$d_H(u,v) = d_H(u/v, \mathbf{1}_n) = d_H(\mathbf{1}_n/u, \mathbf{1}_n/v).$$

This shows that

$$d_H(D^k_\varphi, D^\star_\varphi) = d_H(\frac{\mu}{KD^k_\psi}, \frac{\mu}{KD^\star_\psi}) = d_H(KD^k_\psi, KD^\star_\psi) \leqslant \lambda(K) d_H(D^k_\psi, D^\star_\psi),$$

where we used Theorem 1.7. This shows (1.12). By using triangular inequality we have

$$\begin{aligned}
d_H(D^k_\varphi, D^\star_\varphi) &\leqslant d_H(D^{k+1}_\varphi, D^k_\varphi) + d_H(D^{k+1}_\varphi, D^\star_\varphi) \\
&\leqslant d_H(\frac{\mu}{KD^k_\psi}, D^k_\varphi) + \lambda(K) d_H(D^k_\varphi, D^\star_\varphi) \\
&= d_H(\mu, D^k_\varphi \odot (KD^k_\psi)) + \lambda(K)^2 d_H(D^k_\varphi, D^\star_\varphi) \\
&= d_H(\mu, \gamma^k \mathbf{1}_n) + \lambda(K)^2 d_H(D^k_\varphi, D^\star_\varphi),
\end{aligned}$$

where $\odot$ denotes the element wise multiplication. (1.14) can be proved in an analogous way. (1.15) is trivial. □

References

[1] 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** (2015), no. 2, A1111–A1138.

[2] Garrett Birkhoff, *Extensions of jentzsch's theorem*, Transactions of the American Mathematical Society **85** (1957), no. 1, 219–227.

[3] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM Journal on Mathematical Analysis **49** (2017), no. 2, 1385–1418.

[4] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *Scaling algorithms for unbalanced transport problems*, arXiv preprint arXiv:1607.05816 (2016).

[5] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems, 2013, pp. 2292–2300.

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

[7] Christian Léonard, *From the schr\" odinger problem to the monge-kantorovich problem*, arXiv preprint arXiv:1011.2564 (2010).

[8] ———, *A survey of the schr\" odinger problem and some of its connections with optimal transport*, arXiv preprint arXiv:1308.0215 (2013).

[9] Gabriel Peyré, Marco Cuturi, et al., *Computational optimal transport*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.

[10] Hans Samelson et al., *On the perron-frobenius theorem.*, The Michigan Mathematical Journal **4** (1957), no. 1, 57–59.

[11] François-Xavier Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*, (2019).