Optimal transport

IV Entropic Optimal Transport and Numerics

Seongho, Joo

Seoul National University

Discrete Optimal Transport

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

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

Definition 1 (Discrete OT)

The discrete optimal transport problem between two probability measures μ and ν and give cost function $c : X \times Y \to \mathsf{r}_+ \cup \{+\infty\}$ is the following minimization problem

$$\inf \left\{ \sum_{x \in X} \sum_{y \in Y} \gamma_{xy} c(x, y) \,|\, \gamma \in \Pi(\mu, \nu) \right\},$$

where the set of admissible couplings is now defined as

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

Unfortunately, this linear programming problem has complexity $\mathcal{O}(N^3)$ that 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 discrete case

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

$$P_\varepsilon = \inf \left\{ \langle \gamma, c \rangle + \varepsilon \mathsf{Ent}(\gamma) \,|\, \gamma \in X \times Y, \sum_{y \in Y} \gamma_{xy} = \mu_x, \sum_{x \in X} \gamma_{xy} = \nu_y \right\} \qquad (0.1)$$

where $\langle \gamma, c \rangle = \sum_{x,y} \gamma_{xy} c(x,y)$ and $\mathsf{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$.

The discrete case

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

$$P_\varepsilon = \inf \left\{ \langle \gamma, c \rangle + \varepsilon \mathsf{Ent}(\gamma) \,|\, \gamma \in X \times Y, \sum_{y \in Y} \gamma_{xy} = \mu_x, \sum_{x \in X} \gamma_{xy} = \nu_y \right\} \tag{0.1}$$

where $\langle \gamma, c \rangle = \sum_{x,y} \gamma_{xy} c(x,y)$ and $\mathsf{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$.

Theorem 1

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

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

The discrete case

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

Theorem 2 (Convergence in ε)

The unique solution γ to 0.1 converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is

$$\gamma_\varepsilon \xrightarrow[\varepsilon \to 0]{} \arg\min \{\mathsf{Ent}(\gamma) \,|\, \gamma \in \Pi(\mu, \nu), \langle \gamma, c \rangle = \mathcal{MK}_c(\mu, \nu)\}$$

Proof.

The discrete case

• We want now to derive formally the dual problem. For this purpose, we introduce the Lagrangian associated to eq. (0.1).

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

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

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

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

$$D_\varepsilon = \sup_{\varphi,\psi}\min_{\gamma}\sum_{x,y}\gamma_{xy}\left(c(x,y) - \psi(y) - \varphi(x) + \varepsilon(\log(\gamma_{xy}) - 1)\right) + \sum_{x\in X}\varphi(x)\mu_x + \sum_{y\in Y}\psi(y)\nu_y$$

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

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

The discrete case

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

Putting these values in the definition of D_ε gives

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

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

Note that thanks to the relation, one can recover a solution to the primal problem from the dual one. This is true because, unlike the original linear programming formulation, the regularized problem is smooth and strictly convex. The following duality results holds

The discrete case

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

Putting these values in the definition of D_ε gives

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

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

Note that thanks to the relation, one can recover a solution to the primal problem from the dual one. This is true because, unlike the original linear programming formulation, the regularized problem is smooth and strictly convex. The following duality results holds

Theorem 3 (Strong duality)

Strong duality holds and the maximum in the dual problem is attained, that is $\exists \varphi, \psi$ such that

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

The discrete case

Corollary 1

If (φ, ψ) is the solution to the dual problem, then the solution γ^* to the primal is given by

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

Note that the optimal coupling γ can be written as

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

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.

The discrete case

Corollary 1

If (φ, ψ) is the solution to the dual problem, then the solution γ^* to the primal is given by

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

Note that the optimal coupling γ can be written as

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

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 (Matrix scaling problem)

Let $K \in \mathrm{r}^{N\times N}$ be a matrix with positive coefficients Find D_φ and D_ψ positive diagonal matrices in $K \in \mathrm{r}^{N\times N}$ such that $D_\varphi K D_\psi$ is doubly stochastic, that is sum along each row and each column is equal to 1.

Remark. The uniqueness fails since if (D_φ, D_ψ) is a solution then so is $(cD_\varphi, \frac{1}{c}D_\psi)$ for every $c \in \mathrm{r}_+$.

Matrix scaling problem

• 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 ./ stands for the elements-wise division. Denoting by $(K_\varepsilon)_{x,y} = e^{\frac{-c(x,y)}{\varepsilon}}$ the algorithm takes the following form.

Algorithm 1 Sinkhorn-Knopp algorithm for the regularised optimal transport problem

$D_\varphi^0 \leftarrow \mathbf{1}_X$, $D_\psi^0 \leftarrow \mathbf{1}_Y$
for $0 \leq k < k_{max}$ **do**
  $D_\varphi^{k+1} \leftarrow \mu./(KD_\varphi^k)$
  $D_\psi^{k+1} \leftarrow \nu./(K^\top D_\varphi^{k+1})$
end for

Matrix scaling problem

- Notice that one can recast 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 $\text{diag}(\mu)$ denotes the diagonal matrix with the vector μ as main diagonal. In the case, the primal problem can be re-written as

$$P_\varepsilon(\mu,\nu) = \inf \left\{ \langle \gamma, c \rangle + \varepsilon \mathcal{H}(\gamma|\mu\otimes\nu) \,|\, \gamma \in X\times Y, \sum_{y\in Y} \gamma_{x,y} = \mu_x, \sum_{x\in X} = \nu_y \right\},$$

where $\mathcal{H}(\rho|\nu) := \sum_x \rho_x(\log(\frac{\rho_x}{\mu_x}) - 1)$ is the relative entropy or the Kullback-Leibler divergence.

- One can easily recast the regularized OT in the continuous framework as follows

$$\mathcal{P}_\varepsilon(\mu,\nu) = \inf \left\{ \int_{X\times Y} c(x,y)\,\mathrm{d}\gamma(x,y) + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \,|\, \gamma\in\Pi(\mu,\nu) \right\}, \qquad (0.2)$$

where

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

and the marginals μ, ν are probability measures on the compact metric spaces X and Y, resp. This problem is often referred to as the *static Schrödinger problem*.

The convergence of Sinkhorn in the continuous setting

• We show the existence of the regularized dual problem (and convergence of Sinkhorn at the same time) in the continuous framework. The Sinkhorm algorithm is actually a coordinate ascent algorithm: the main idea is indeed to maximize $\Phi_\varepsilon(\varphi, \psi)$ by maximizing alternatively in φ and ψ. From now on we assume for simplicity that $X = Y$.

Proposition 1

The dual problem to 0.2 reads as

$$D_\varepsilon = \sup \left\{ \Phi_\varepsilon(\varphi, \psi) \,|\, \varphi, \psi \in C_0(X) \right\},$$

where

$$\Phi_\varepsilon(\varphi, \psi) := \int_X \varphi(x)\, \mathrm{d}\mu(x) + \int_Y \psi(y)\, \mathrm{d}\nu(y) - \varepsilon \int_{X \times Y} \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon} \right) \mathrm{d}\mu \otimes \mathrm{d}\nu(x,y).$$

It is strictly convex w.r.t each argument φ and ψ and strictly convex w.r.t $\varphi(x)+\psi(y)$. Is is also Fréchet differentiable for the $(C_0, \|\cdot\|)$ topology. Furthermore, if a maximize exists it is unique up to a constant, that is $\varphi_\varepsilon(\varphi, \psi) = \varphi_\varepsilon(\varphi + C, \psi - C)$ for every $C \in \mathsf{r}$.

The convergence of Sinkhorn in the continuous setting

Proposition 2

The maximization of $\Phi_\varepsilon(\varphi, \psi)$ w.r.t each variable can be made explicit, and the Sinkrhom algorithm can be defined as

$$\varphi_{k+1}(x) = -\varepsilon \log \left(\int_X \exp(\frac{1}{\varepsilon}(\psi_k(y) - c(x,y)) \, \mathrm{d}\nu(y) \right) := S_\mu(\psi_k), \tag{0.3}$$

$$\psi_{k+1}(y) = -\varepsilon \log \left(\int_X \exp(\frac{1}{\varepsilon}(\varphi_{k+1}(x) - c(x,y)) \, \mathrm{d}\mu(x) \right) := S_\mu(\varphi_{k+1}). \tag{0.4}$$

Moreover, the following properties hold

1. $\Phi_\varepsilon(\varphi_k, \psi_k) \leq \Phi_\varepsilon(\varphi_{k+1}, \psi_k) \leq \Phi_\varepsilon(\varphi_{k+1}, \psi_{k+1})$;
2. The continuity modulus of $\varphi_{k+1}, \psi_{k+1}$ is bounded by that of $c(x,y)$;
3. If $\psi_k - C(\varphi_{k+1} - C)$ is bounded by M on the support of $\mu(\mu)$, then so is $\varphi_{k+1}(\psi_{k+1})$.

Proof.

The convergence of Sinkhorm in the continuous setting

Proposition 3

The sequence (φ_k, ψ_k) defined by 0.3 and 0.4 converge in $(C_0, \|\cdot\|_\infty)$ to the unique (up to a constant) couple of potentials (φ, ψ) which maximizes Φ_ε.

Proof.

The convergence of Sinkhorn in the continuous setting

- The proof of convergence relies on some important properties of the logsumexp (LSE) function $\log \int \exp$ which we summarise in the next Lemma. we define the pseudo-norm $\|\cdot\|_{\circ,\infty}$ of uniform convergence as

$$\|f\|_{\circ,\infty} := \frac{1}{2}(\sup f - \inf f) = \inf_{a \in \mathsf{r}} \|f + a\|_\infty .$$

Lemma 1

The LSE function is convex and

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leq \|\varphi_1 - \varphi_2\|_{\circ,\infty} \tag{0.5}$$

Proof.

The convergence of Sinkhorn in the continuous setting

Lemma 2

Let $u, v \in C(X)$ and $\mu \in \mathcal{P}(X)$ and denote ν_u, ν_v the Gibbs measures associated to u and v, that is $\mathrm{d}\nu_u = \frac{1}{Z_u} e^u \,\mathrm{d}\mu$ and $\mathrm{d}\nu_v = \frac{1}{Z_v} e^v \,\mathrm{d}\mu$, where Z_u and Z_v are the normalizing the constants, then

$$\|\nu_u - \nu_v\|_{L^1} \leq 2(1 - e^{-2\|u-v\|_{\circ,\infty}}).$$

Proof.

The convergence of Sinkhorn in the continuous Setting

Theorem 4 (Convergence of Sinkhorm)

The map $S = S_\nu \circ S_\mu$ is a contraction for $\|\cdot\|_{\circ,\infty}$. In particular the sequence (φ_k, ψ_k) defined by the Sinkhorn algorithm linearly converges to the unique (up to a constant) maximizer of the dual problem.

Proof. We actually have to prove that

$$\|S_\mu(\varphi_1) - \S_\mu(\varphi_2)\|_{\circ,\infty} \leq \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}, \tag{0.6}$$

and analogously for S_ν with $\kappa_\mu < 1 (\kappa_\nu < 1)$. Once we have established that $S_\mu(S_\nu)$ is a contraction then is easily follows that

$$\|S(\varphi_1) - S(\varphi_2)\|_{\circ,\infty} \leq \kappa_\mu \kappa_\nu \|\varphi_1 - \varphi_2\|_{\circ,\infty}$$

which would conclude the proof. In order to prove 0.6 we give an estimation of the oscillations of S_μ

$$\frac{1}{2}|S_\mu(\varphi_1)(y) - S_\mu(\varphi_2)(y) - S_\mu(\varphi_1)(x) + S_\mu(\varphi_2)(x)| \leq \frac{1}{2}\left|\int_0^1 \int_X (\varphi_1 - \varphi_2)(\,\mathrm{d}\eta_{t,y} - \,\mathrm{d}\eta_{t,x})\,\mathrm{d}t\right|$$

where $\mathrm{d}\eta_{t,z} := \frac{1}{Z}\exp\left(\frac{t(\varphi_1-\varphi_2)+\varphi_2-c(z,\cdot)}{\varepsilon}\right)\mathrm{d}\mu$ where Z is the normalising constant. Since $\mathrm{d}\eta_{t,z}$ is a Gibbs measure we can apply the L^1 bound of lemma 2 to estimate $\|\eta_{t,y} - \eta_{t,x}\|_{L^1}$ and get

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leq \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}$$

with $\kappa = 2(1 - e^{-2\frac{\|c\|_{\circ,\infty}}{\varepsilon}})$.