Entropic Optimal Transport

On Sinkhorn algorithm and the link with Schröfinger problem
Monge-Kantorovich problem and some numerics

\[ \mathcal{K}_c(\mu, \nu) := \inf \left\{ \int_{X \times Y} c(x, y) \gamma(x, y) \, d\gamma(x, y) \mid \gamma \in \Pi(\mu, \nu) \right\} \]

Three main ways to solve numerically this problem:

1) Discrete to Discrete: X and Y are finite set and the measures are supported on diracs;

2) Discrete to continuous: one of the measure is a.c. With respect to Lebesgue (see Quentin Mérigot’s works);

3) Continuous to continuous: both the measures are ac with respect to Lebesgue: the celebrated Benamou-Bernier formulation of Optimal transport.
Some remarks:

- Semi-discrete OT works with quadratic cost (or some close variant such as some power $p$ distance);

- Continuous OT works if the Monge-Kantorovich problem admits a dynamic formulation;

- Other approaches exist: e.g. column generation (Friesecke and Penka), moment constraints relaxation (Virginie, Alfonsi et al), solving Monge-Ampére equation for the quadratic case (Benamou, Mirebeau, Froese, Oberman, Duval)

Today we will main focus on discrete optimal and entropic regularization to solve it
Assume $X$ and $Y$ finite sets (with both cardinality $N$),

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

Then the problem is formulated as follows

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

$N^2$ unknowns and $2N$ constraints to verify

$O(N^3)$ complexity via standard linear programming
Main idea: penalize the non-negativity of $\gamma_{x,y} \geq 0$ by means of an entropy term $\text{Ent}(\gamma) := \sum_{x,y} e(\gamma_{x,y})$ where
\[
e(\gamma) = \begin{cases} 
  r(\log r - 1) & \text{if } r > 0 \\
  0 & \text{if } r = 0 \\
  +\infty & \text{if } r < 0
\end{cases}
\]

The regularized problem takes the following form
\[
\mathcal{K}^e_c(\mu, \nu) := \inf \left\{ \langle c, \gamma \rangle + \varepsilon \text{Ent}(\gamma) \mid \sum_y \gamma_{x,y} = \mu_x, \sum_x \gamma_{x,y} = \nu_y \right\}
\]

Where $\langle c, \gamma \rangle = \sum_{x,y} c(x,y)\gamma_{x,y}$. 
1st good news

**Thm:** problem $\mathcal{K}_c^\varepsilon(\mu, \nu)$ has a unique solution $\gamma_\varepsilon$ which belongs to $\Pi(\mu, \nu)$.

2nd good news

**Thm[convergence in $\varepsilon$]:** Consider the sequence of unique solutions $\gamma_\varepsilon$, then it converges to the optimal solution with the minimal entropy within the set of all optimal solution of $\mathcal{K}_c(\mu, \nu)$ that is

$$
\gamma_\varepsilon \to \arg\min\{\text{Ent}(\gamma) \mid \gamma \in \Pi(\mu, \nu), \langle c, \gamma \rangle = \mathcal{K}_c(\mu, \nu)\}.
$$
Sketch

- Take \( \varepsilon_n \) s.t. \( \varepsilon_n \to 0 \) and denote \( Y \) in the solution with \( \varepsilon = \varepsilon_n \).

- \( \Pi(\mu, \nu) \) is bounded and close \( \exists \gamma_{\varepsilon_n} \to \gamma^* \in \Pi(\mu, \nu) \).

- Take \( \Pi \) optimal for \( K_\varepsilon(\mu, \nu) \).

\[ 0 \leq \langle \gamma_{\varepsilon_n}, c \rangle - \langle \gamma, c \rangle \leq \varepsilon_n (\text{Ent}(\gamma) - \text{Ent}(\nu)) \]
The effect of the regularization

Marginals $\mu$ and $\nu$

Support of the optimal $\gamma_\varepsilon$ as $\varepsilon \to 0$
The matching problem and the regularized counterpart

(From G. Peyre’s twitter)
Deriving the dual problem. First, consider the Lagrangian associated to the entropic
problem, that is

\[ \mathcal{L}(\gamma, \varphi, \psi) := \langle c, \gamma \rangle + \varepsilon \text{Ent}(\gamma) + \sum_x \varphi(x) \left( \mu_x - \sum_y \gamma_{x,y} \right) + \sum_y \psi(y) \left( \nu_y - \sum_x \gamma_{x,y} \right) \]

Where, as in the unregularized case, \( \varphi \) and \( \psi \) are the Lagrange multipliers.

Then,

\[ \mathcal{H}_c^\varepsilon(\mu, \nu) = \inf_{\gamma} \sup_{\varphi, \psi} \mathcal{L}(\gamma, \varphi, \psi) \]
Rmk: by KKT (optimality) conditions $\partial_\gamma \mathcal{L}(\gamma, \varphi, \psi) = 0$, we have the following relation between primal and dual variables

$$c(x, y) - \varphi(x) - \psi(y) + \varepsilon \log(\gamma_{x,y}) = 0$$

Which gives the following form of the optimal $\gamma$

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

Now, by interchanging $\inf$ and $\sup$, as we did last week, we obtain…
\[
\mathcal{D}^\varepsilon_c(\mu, \nu) := \sup_{\phi, \psi} \Phi^\varepsilon(\phi, \psi),
\]

Where

\[
\Phi^\varepsilon(\phi, \psi) := \sum_x \phi(x) \mu_x + \sum_y \psi(y) \nu_y - \varepsilon \sum_{x,y} \exp\left(\frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon}\right)
\]

**Thm:** strong duality holds \( \mathcal{H}^\varepsilon_c(\mu, \nu) = \mathcal{D}^\varepsilon_c(\mu, \nu) \)
Rmk 1: the optimal coupling can be written as

$$\gamma = D_\varphi KD_\psi$$

Where $D_\varphi = \text{diag}(\varphi/\epsilon)$ and $D_\psi = \text{diag}(\psi/\epsilon)$ are diagonal matrices and $K \in \mathbb{R}^N \times \mathbb{R}^N$ is such that $K_{x,y} = \exp\left(\frac{-c(x,y)}{\epsilon}\right)$. This actually makes our problem very similar to a matrix scaling problem.

**Def (matrix scaling problem):** Given a matrix $K$ with positive coefficients find $(D_\varphi, D_\psi)$ such that $D_\varphi KD_\psi$ is doubly stochastic.

Rmk 2: Notice that if $(D_\varphi, D_\psi)$ is a solution then $(cD_\varphi, \frac{1}{c}D_\psi)$ for any $c$. 
Sinkhorn algorithm

Algorithm 1 Sinkhorn-Knopp algorithm for the matrix scaling problem

1: function SINKHORN-KNOPP(K)
2: \[ D^0_\varphi \leftarrow 1_N, \; D^0_\psi \leftarrow 1_N \]
3: for \( 0 \leq k < k_{\text{max}} \) do
4: \[ D^{k+1}_\varphi \leftarrow 1_N ./ (K D^k_\psi) \]
5: \[ D^{k+1}_\psi \leftarrow 1_N ./ (K^T D^{k+1}_\varphi) \]
6: end for
7: end function

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

1: function SINKHORN-KNOPP(K_\varepsilon, \mu, \nu)
2: \[ D^0_\varphi \leftarrow 1_X, \; D^0_\psi \leftarrow 1_Y \]
3: for \( 0 \leq k < k_{\text{max}} \) do
4: \[ D^{k+1}_\varphi \leftarrow \mu ./ (K D^k_\psi) \]
5: \[ D^{k+1}_\psi \leftarrow \nu ./ (K^T D^{k+1}_\varphi) \]
6: end for
7: end function

Rmk: \( \mathcal{H}_C^\varepsilon \) can be recasted as a matrix scaling problem by taking
The importance of being sparse: a multi-scale approach.

In order to reduce the number of grid points used one can apply a multiscale approach and refine the mesh where the solution is supported.

Figure: support of the optimal $\gamma_\varepsilon$ for the Coulomb cost.
Consider the entropic with a kernel $K$. Then we can re-write the Sinkhorn iterations on log-domain and obtain an iterative methods acting on the dual variable, that is

$$\varphi^k = \varepsilon \log(\mu) - \varepsilon \log(KD^k_{\psi}),$$

$$\psi^k = \varepsilon \log(\nu) - \varepsilon \log(KD^k_{\varphi}).$$

It turns out that $-\varepsilon \log(KD^k_{\psi})$ (resp. $-\varepsilon \log(KD^k_{\varphi})$) is the soft c-transform of $\psi$ (resp. $\varphi$).

In particular the relations above still hold for the optimal dual variables and we have

$$\varphi^* = \varepsilon \log(\mu) - \varepsilon \log(KD^*_{\psi}) \to \min_y c(x, y) - \psi(y) \text{ as } \varepsilon \to 0.$$
Convergence of Sinkhorn by using the Hilbert metric

**Def (Hilbert projective metric):** the Hilbert projective metric on $\mathbb{R}_+^*$ is defined as

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

Where $\| x \|_V = \max_{i} x_i - \min_{i} x_i$.

**Theorem A.2 ([2, 12]).** Let $K \in \mathbb{R}_{+,*}^{n \times n}$, then for $(u, v) \in (\mathbb{R}_+^*)^2$

$$d_H(Ku, Kv) \leq \lambda(K) d_H(u, v),$$

where

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

and

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


Thm(Franklin and Lorenz ‘89): One has $(D^k_{\phi}, D^k_{\psi}) \rightarrow (D^*_\phi, D^*_\psi)$ and

$$d_H(D^k_{\phi}, D^*_\phi) = O(\lambda(K)^{2k}), \quad d_H(D^k_{\psi}, D^*_\psi) = O(\lambda(K)^{2k}).$$

Moreover,

$$d_H(D^k_{\phi}, D^*_\phi) \leq \frac{d_H(\gamma^k1_n, \mu)}{1 - \lambda(K)^2},$$

$$d_H(D^k_{\psi}, D^*_\psi) \leq \frac{d_H(\gamma^kT1_n, \nu)}{1 - \lambda(K)^2},$$

Where $\gamma^k = D^k_{\phi}KD^k_{\psi}$. Last, one has

$$||\log(\gamma^k) - \log(\gamma_\epsilon)||_\infty \leq d_H(D^k_{\phi}, D^*_\phi) + d_H(D^k_{\psi}, D^*_\psi).$$
Sketch of proof:

First, notice that

\[ d_H(u, v) = d_H(u/v, 1_n) = d_H(1_n/u, 1_n/v). \]

Together with the previous theorem, this gives

\[ d_H(D^k_\psi, D^*_\psi) = d_H\left(\frac{\mu}{KD^k_\psi}, \frac{\mu}{KD^*_\psi}\right) = d_H(KD^k_\psi, KD^*_\psi) \leq \lambda(K)d_H(D^k_\psi, D^*_\psi). \]
Then by using the triangular inequality we have

\[
\begin{align*}
    d_H(D^k_\varphi, D^*_\varphi) &\leq d_H(D^{k+1}_\varphi, D^k_\varphi) + d_H(D^{k+1}_\varphi, D^*_\varphi) \\
    &\leq d_H\left(\frac{\mu}{KD^k_\psi}, D^k_\varphi\right) + \lambda(K)d_H(D^k_\varphi, D^*_\varphi) \\
    &= d_H(\mu, D^k_\varphi \odot (KD^k_\psi)) + \lambda(K)^2d_H(D^k_\varphi, D^*_\varphi) \\
    &= d_H(\mu, \gamma^k 1_n) + \lambda(K)^2d_H(D^k_\varphi, D^*_\varphi),
\end{align*}
\]

Where \( \odot \) denotes the element-wise multiplication.
**Rmk (stopping criteria):** the bounds $d_H(D^k_\varphi, D^*_\varphi) \leq \frac{d_H(\gamma^k 1_n, \mu)}{1 - \lambda(K)^2}$ and $d_H(D^k_\psi, D^*_\psi) \leq \frac{d_H(\gamma^k, T 1_n, \nu)}{1 - \lambda(K)^2}$ shows that some error measures on the marginal constraints violation, for instance $||\gamma^k 1_n - \mu||_1$ and $||\gamma^k, T 1_n - \nu||_1$, are useful stopping criteria to monitor the convergence.

**Rmk:** This theorem shows that Sinkhorn algorithm converges linearly, but the rates becomes exponentially bad as $\epsilon \to 0$, since it scales like $e^{-1/\epsilon}$. 
Back to the continuous case

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

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

Where

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

And $\mu, \nu$ are probability measures on the compact sets $X$ and $Y$. 
Linear convergence of Sinkhorn for bounded cost

Consider the following variant of Sinkhorn algorithm

\[
\varphi^{k+1}(x) = -\varepsilon \log \left( \int_X \exp \left( \frac{1}{\varepsilon} (\psi^k(y) - c(x, y)) \right) d\nu(y) \right) + \lambda^k
\]

\[
\psi^{k+1}(y) = -\varepsilon \log \left( \int_X \exp \left( \frac{1}{\varepsilon} (\varphi^{k+1}(x) - c(x, y)) \right) d\mu(x) \right),
\]

Where \( \lambda^k = \varepsilon \int_X \log \left( \int_X \exp \left( \frac{1}{\varepsilon} (\psi_k(y) - c(x, y)) \right) d\nu(y) \right) d\mu(x). \)

This is equivalent to previous Sinkhorn...I am just fixing a constant.
The algorithm in the previous slide is equivalent to the following coordinate ascent method

\[ \varphi^{k+1} = \arg\max_{\varphi, \int \varphi d\mu = 0} \Phi_{\varepsilon}(\varphi, \psi^k), \]

\[ \psi^{k+1} = \arg\max_{\psi} \Phi_{\varepsilon}(\varphi^{k+1}, \psi). \]

Where

\[ \Phi_{\varepsilon}(\phi, \psi) := \int_{X} \varphi(x) d\mu(x) + \int_{Y} \psi(y) d\nu(y) \]

\[ -\varepsilon \int_{X \times Y} \exp\left( \frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon} \right) d\mu \otimes d\nu(x, y). \]
**Rmk [alternative dual formulation]:** one can show that the dual problem can also be rewritten in the following way

$$\sup_{\phi, \psi} \tilde{\Phi}_\varepsilon(\phi, \psi),$$

Where

$$\tilde{\Phi}_\varepsilon(\phi, \psi) := \int_X \phi(x) d\mu(x) + \int_Y \psi(y) d\nu(y)$$

$$-\varepsilon \log \left( \int_{X \times Y} \exp \left( \frac{\phi(x) + \psi(y) - c(x, y)}{\varepsilon} \right) d\mu \otimes d\nu(x, y) \right).$$

To see it just use the variational representation of the relative entropy, that is

$$\mathcal{H}(\rho | \pi) = \sup_{\phi} \left( \int \phi d\rho - \log \left( \int e^{\phi} d\pi \right) \right).$$
**Lemma:** For every $k \geq 0$ we have

$$||\varphi^k||_\infty \leq 2||c||_\infty \quad \text{and} \quad ||\psi^k||_\infty \leq 3||c||_\infty.$$ 

**Proof** Just compute $\varphi^k(x_1) - \varphi^k(x_2)$.

**Thm:** Let $(\varphi^*, \psi^*)$ the unique solution of the dual entropic problem with

$$\int_X \varphi^*(x)d\mu(x) = 0.$$ 

The iterates of Sinkhorn satisfy

$$\Phi_\varepsilon(\varphi^*, \psi^*) - \Phi_\varepsilon(\varphi^k, \psi^k) \leq \beta^k(\Phi_\varepsilon(\varphi^*, \psi^*) - \Phi_\varepsilon(\varphi^0, \psi^0)),$$

$$||\varphi^* - \varphi^k||_{L^2}^2 + ||\psi^* - \psi^k||_{L^2}^2 \leq \eta \beta^k(\Phi_\varepsilon(\varphi^*, \psi^*) - \Phi_\varepsilon(\varphi^0, \psi^0)),$$

Where $\beta := 1 - e^{-24||c||_\infty}/\varepsilon$ and $\eta = 2e^6||c||_\infty/\varepsilon$. 
**Sketch of proof:**

The basic idea is to use strong convexity of the exponential function on an interval $[-\alpha, +\infty)$, that is

$$e^b - e^a \geq (b - a)e^a + \frac{e^{-\alpha}}{2} |b - a|^2, \quad \text{for} \quad a, b \in [-\alpha, +\infty).$$

**Step 1:** Given $\sigma = e^{-6\|c\|_\infty/\epsilon}$ we have

$$\Phi_\epsilon(\phi^{k+1}, \psi^{k+1}) - \Phi_\epsilon(\phi^k, \psi^k) \geq \frac{\sigma}{2}(\|\phi^{k+1} - \phi^k\|^2_{L^2} + \|\psi^{k+1} - \psi^k\|^2_{L^2}).$$
Step 2: Given the optimal \((\varphi^*, \psi^*)\) we obtain

\[
\Phi_\varepsilon(\varphi^k, \psi^k) - \Phi_\varepsilon(\varphi^*, \psi^*) \geq \int_X \partial_1 \Phi_\varepsilon(\varphi^k, \psi^k)(x)[\varphi^k(x) - \varphi^*(x)]d\mu(x) \\
+ \int_Y \partial_2 \Phi_\varepsilon(\varphi^k, \psi^k)(y)[\psi^k(y) - \psi^*(y)]d\nu(y) \\
+ \frac{\sigma^2}{2} (||\varphi^k - \varphi^*||_{L^2}^2 + ||\psi^k - \psi^*||_{L^2}^2)
\]

Where

\[
\partial_1 \Phi_\varepsilon(\varphi, \psi) = 1 - \varepsilon \int_Y e^{\frac{\varphi + \psi - c}{\varepsilon}} d\nu(y)
\]
Step 3: By exploring the zero mean iterates and Young’s inequality we get

$$\Phi_\varepsilon(\varphi^*, \psi^*) - \Phi_\varepsilon(\varphi^k, \psi^k) \leq \frac{1}{2\sigma} \left| \| \partial_1 \Phi_\varepsilon(\varphi^k, \psi^k) - \partial_1 \Phi_\varepsilon(\varphi^{k+1}, \psi^k) \|_{L^2} \right|^2.$$

Now by Lipschitz continuity of the exponential and step 1 we have

$$\Phi_\varepsilon(\varphi^*, \psi^*) - \Phi_\varepsilon(\varphi^k, \psi^k) \leq \frac{1}{\sigma^4} (\Phi_\varepsilon(\varphi^{k+1}, \psi^{k+1}) - \Phi_\varepsilon(\varphi^k, \psi^k))$$

Taking $\Delta^k = \Phi_\varepsilon(\varphi^*, \psi^*) - \Phi_\varepsilon(\varphi^k, \psi^k)$ the above inequality can be expressed as

$$\Delta^{k+1} \leq (1 - \sigma^4) \Delta^k$$

Iterating we get the result.