Entropic Optimal Transport

On Sinkhorn algorithm and the link with Schröfinger problem

Luca Nenna (LMO, Université Paris-Saclay), 24/01/2024

Monge-Kantorovich problem and some numerics

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

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 unknonws and $2N$ constraints to verify

$O(N^3)$ complexity via standard linear programming

Entropic Optimal Transport: the discrete case

Main idea: penalize the non-negativity of $\gamma_{x,y} \geq 0$ by means of an entropy term

$$\mathrm{Ent}(\gamma) := \sum_{\mathrm{x,y}} \mathrm{e}(\gamma_{\mathrm{x,y}}) \text{ where}$$

$$e(r) = \begin{cases} r(\log r - 1) & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ +\infty & \text{if } r < 0 \end{cases}$$

The regularized problem takes the following form

$$\mathscr{K}_c^{\varepsilon}(\mu, \nu) := \inf \left\{ \langle c, \gamma \rangle + \varepsilon \mathrm{Ent}(\gamma) \mid \sum_{\mathrm{y}} \gamma_{\mathrm{x,y}} = \mu_{\mathrm{x}}, \ \sum_{\mathrm{x}} \gamma_{\mathrm{x,y}} = \nu_{\mathrm{y}} \right\}$$

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

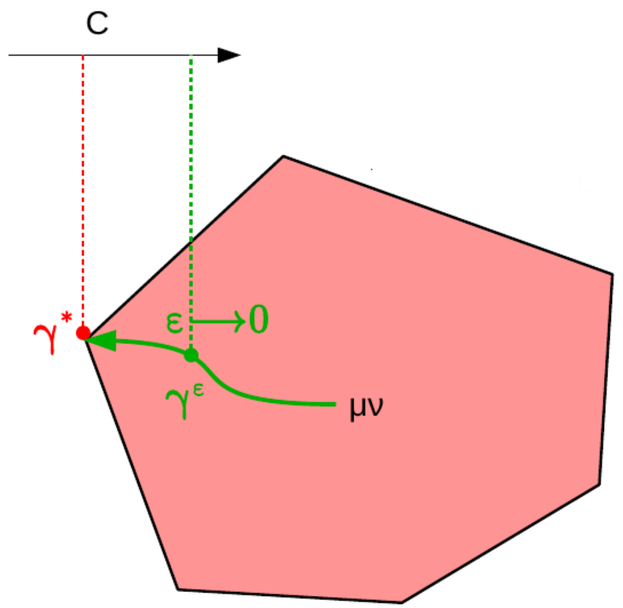
1st good news

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

2nd good news

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

$$\gamma_\varepsilon \to \operatorname{argmin}\{\mathrm{Ent}(\gamma) \mid \gamma \in \Pi(\mu,\nu),\ \langle \mathrm{c},\gamma\rangle = \mathscr{K}_{\mathrm{c}}(\mu,\nu)\}.$$

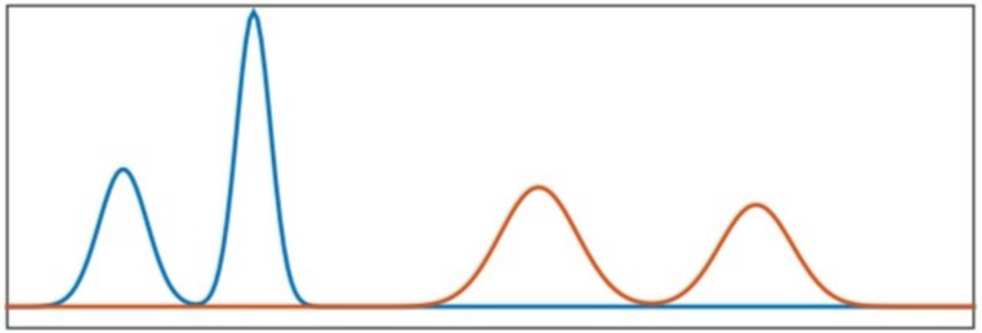


Sketch

- TAKE ε_n S.T. $\varepsilon_n \to 0$ ($n \uparrow +\infty$) AND DENOTE γ_n THE SOLUTION WITH $\varepsilon = \varepsilon_n$
- $\Pi(\mu,\nu)$ IS BOUNDED AND CLOSE $\exists \gamma_{k_n} \to \gamma^* \in \Pi(\mu,\nu)$
- TAKE φ OPTIMAL FOR $K_\varepsilon(\mu,\nu)$

$$0 \leq \langle \gamma_n, C \rangle - \langle \gamma, C \rangle \leq \varepsilon_n \left(\mathrm{Ent}(\gamma) - \mathrm{Ent}(\gamma_n) \right)$$

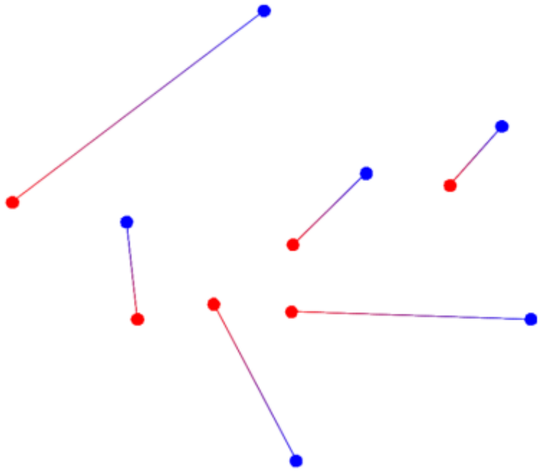
The effect of the regularization



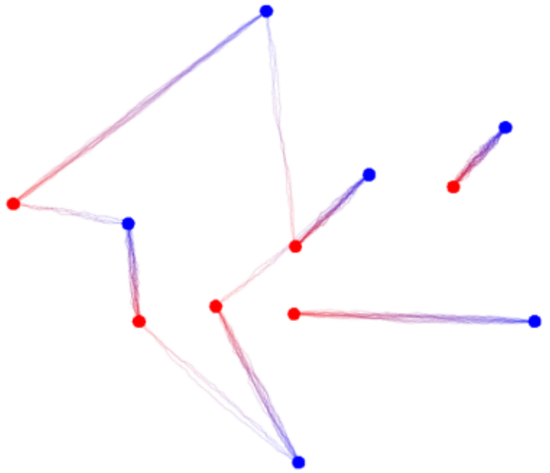
Marginals μ and ν

Support of the optimal γ_ε as $\varepsilon \to 0$

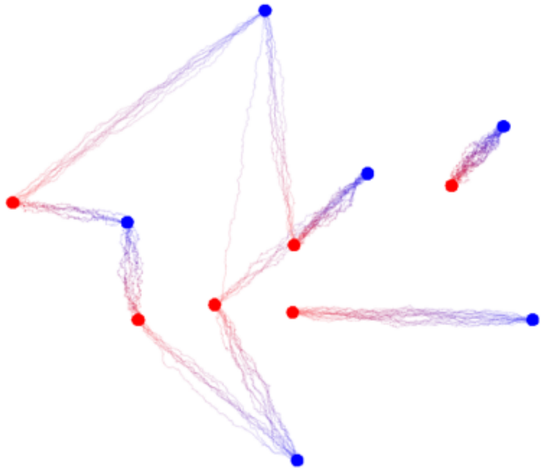
The matching problem and the regularized counterpart



$\varepsilon = 0$

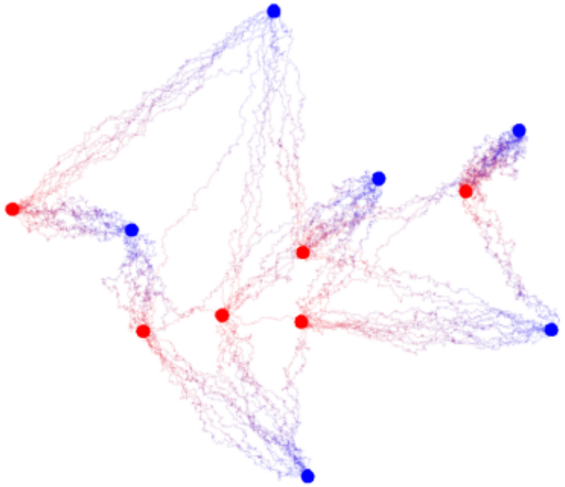


$\varepsilon = .05$



$\varepsilon = 0.2$

(From G. Peyre's twitter)



$\varepsilon = 1$

Deriving the dual problem. First, consider the Lagrangian associated to the entropic problem, that is

$$\mathscr{L}(\gamma, \varphi, \psi) := \langle c, \gamma \rangle + \varepsilon \mathrm{Ent}(\gamma) + \sum_{\mathrm{x}} \varphi(\mathrm{x}) \left(\mu_{\mathrm{x}} - \sum_{\mathrm{y}} \gamma_{\mathrm{x,y}} \right) + \sum_{\mathrm{y}} \psi(\mathrm{y}) \left(\nu_{\mathrm{y}} - \sum_{\mathrm{x}} \gamma_{\mathrm{x,y}} \right)$$

Where, as in the unregularized case, φ and ψ are the Lagrange multipliers.

Then,

$$\mathscr{K}_c^{\varepsilon}(\mu, \nu) = \inf_{\gamma} \sup_{\varphi, \psi} \mathscr{L}(\gamma, \varphi, \psi)$$

Rmk: by KKT (optimality) conditions $\partial_\gamma \mathscr{L}(\gamma, \varphi, \psi) = 0$, 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 γ

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

Now, by interchanging $\inf$ and $\sup$, as we did last week, we obtain…

$$\mathscr{D}_c^{\varepsilon}(\mu,\nu) := \sup_{\varphi,\psi} \Phi_{\varepsilon}(\varphi,\psi),$$

Where

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

Thm: *strong duality holds* $\mathscr{K}_c^{\varepsilon}(\mu,\nu) = \mathscr{D}_c^{\varepsilon}(\mu,\nu)$

Rmk 1: the optimal coupling can be written as

$$\gamma = D_\varphi K D_\psi$$

Where $D_\varphi = \mathrm{diag}(\varphi/\varepsilon)$ and $D_\psi = \mathrm{diag}(\psi/\varepsilon)$ are diagonal matrices and $K \in \mathbb{R}^N \times \mathbb{R}^N$ is such that $K_{x,y} = \exp\left(\frac{-c(x,y)}{\varepsilon}\right)$. 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_φ, D_ψ) such that $D_\varphi K D_\psi$ is doubly stochastic.

Rmk 2: Notice that if (D_φ, D_ψ) 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

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

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

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

Rmk: $\mathscr{K}_c^\varepsilon$ 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

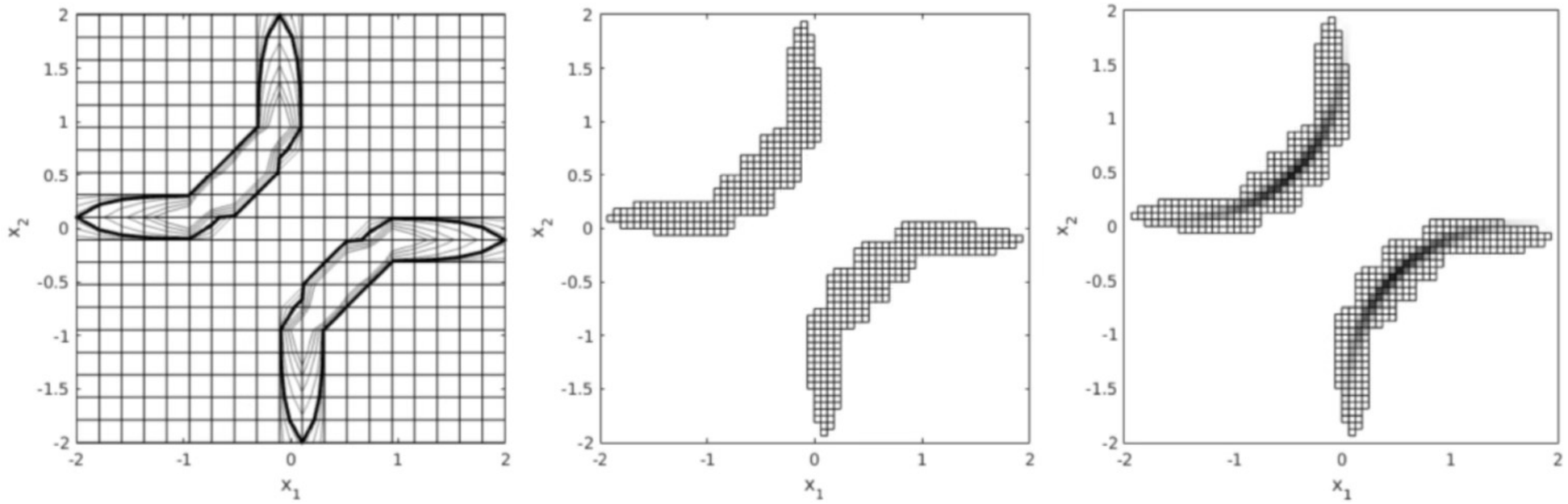


Figure: support of the optimal γ_ε for the Coulomb cost.

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

$$\varphi^k = \varepsilon \log(\mu) - \varepsilon \log(KD_\psi^k),$$

$$\psi^k = \varepsilon \log(\nu) - \varepsilon \log(KD_\varphi^k).$$

It turns out that $-\varepsilon \log(KD_\psi^k)$ (resp. $-\varepsilon \log(KD_\varphi^k)$) is the soft c-transform of ψ (resp. φ).

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 $\mathbb{R}^n_{+,\star}$ is defined as

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

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

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

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

where

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

and

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

[2] Birkhoff, Extensions of Jentzsch's theorem, Transaction of the American Mathematical Society 85 (1957), no. 1, 219-227.

[12] Samelson et al, On the Perron-Frobenieus theorem, The Michigan Mathematical Journal 4 (1957), no. 1, 57-59

Thm(Franklin and Lorenz '89): *One has* $(D_\varphi^k, D_\psi^k) \to (D_\varphi^*, D_\psi^*)$ *and*

$$d_H(D_\varphi^k, D_\varphi^*) = O(\lambda(K)^{2k}), \quad d_H(D_\psi^k, D_\psi^*) = O(\lambda(K)^{2k})\,.$$

Moreover,

$$d_H(D_\varphi^k, D_\varphi^*) \leq \frac{d_H(\gamma^k \mathbf{1}_n, \mu)}{1-\lambda(K)^2},$$

$$d_H(D_\psi^k, D_\psi^*) \leq \frac{d_H(\gamma^{k,T} \mathbf{1}_n, \nu)}{1-\lambda(K)^2},$$

Where $\gamma^k = D_\varphi^k K D_\psi^k$. Last, one has

$$||\log(\gamma^k) - \log(\gamma_\varepsilon)||_\infty \leq d_H(D_\varphi^k, D_\varphi^*) + d_H(D_\psi^k, D_\psi^*).$$

Sketch of proof:

First, notice that

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

Together with the previous theorem, this gives

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

Then by using the triangular inequality we have

$$
\begin{aligned}
d_H(D_\varphi^k, D_\varphi^\star) &\leqslant d_H(D_\varphi^{k+1}, D_\varphi^k) + d_H(D_\varphi^{k+1}, D_\varphi^\star) \\
&\leqslant d_H(\frac{\mu}{KD_\psi^k}, D_\varphi^k) + \lambda(K) d_H(D_\varphi^k, D_\varphi^\star) \\
&= d_H(\mu, D_\varphi^k \odot (KD_\psi^k)) + \lambda(K)^2 d_H(D_\varphi^k, D_\varphi^\star) \\
&= d_H(\mu, \gamma^k \mathbf{1}_n) + \lambda(K)^2 d_H(D_\varphi^k, D_\varphi^\star),
\end{aligned}
$$

Where $\odot$ denotes the element-wise multiplication.

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

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

Back to the continuous case

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

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

Where

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

And μ,ν are probability measures on the compact sets X and Y.

Linear convergence of Sinkhorn for bounded cost

Consider the following variant of Sinkhorn algorithm

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

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

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

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

$$\varphi^{k+1} = \operatorname{argmax}_{\varphi, \int \varphi \mathrm{d}\mu = 0} \Phi_\varepsilon(\varphi, \psi^{\mathrm{k}}),$$
$$\psi^{k+1} = \operatorname{argmax}_{\psi} \Phi_\varepsilon(\varphi^{\mathrm{k+1}}, \psi)\,.$$

Where

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

Rmk [alternative dual formulation]: one can sho that the dual problem can also re-written in the following way

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

Where

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

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

$$\mathscr{H}(\rho\,|\,\pi) = \sup_\phi \left(\int \phi\mathrm{d}\rho - \log\left(\int \mathrm{e}^\phi \mathrm{d}\pi\right)\right).$$

Lemma: *For every $k \geq 0$ we have*

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

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

Thm: *Let (φ^*, ψ^*) the unique solution of the dual entropic problem with $\int_X \varphi^*(x)\mathrm{d}\mu(\mathrm{x}) = 0$. 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 $\sigma = e^{-6||c||_\infty/\varepsilon}$ we have

$$\Phi_\varepsilon(\varphi^{k+1}, \psi^{k+1}) - \Phi_\varepsilon(\varphi^k, \psi^k) \geq \frac{\sigma}{2}(||\varphi^{k+1} - \varphi^k||_{L^2}^2 + \psi^{k+1} - \psi^k||_{L^2}^2).$$

Step 2: Given the optimal (φ^*, ψ^*) we obtain

$$\Phi_\varepsilon(\varphi^k, \psi^k) - \Phi_\varepsilon(\varphi^*, \psi^*) \geq \int_X \partial_1 \Phi_\varepsilon(\varphi^k, \psi^k)(x)[\varphi^k(x) - \varphi^*(x)]\mathrm{d}\mu(\mathrm{x})$$

$$+ \int_Y \partial_2 \Phi_\varepsilon(\varphi^k, \psi^k)(y)[\psi^k(y) - \psi^*(y)]\mathrm{d}\nu(\mathrm{y})$$

$$+ \frac{\sigma}{2}(||\varphi^k - \varphi^*||_{L^2}^2 + ||\psi^k - \psi^*||_{L^2}^2)$$

Where

$$\partial_1 \Phi_\varepsilon(\varphi, \psi) = 1 - \varepsilon \int_Y e^{\frac{\varphi + \psi - c}{\varepsilon}} \mathrm{d}\nu(\mathrm{y})$$

Step 3: By exploring the zero mean iterates and Young's inequality we get

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

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.