SLMath Summer School on OT: Day 5 - Problem Set 1

1. (**First order method to solve OT: Alternative of Sinkhorn**) Consider the MK optimal transport problem between two discrete measures $\mu := \sum_{i=1} a_i\delta_{x_i}$ and $\nu := \sum_{j=1}^m b_j\delta_{y_j}$ where $\sum_{i=1}^n a_i = 1$ and $\sum_{j=1}^m b_j = 1$:

$$\min_{\pi\in\mathbb{R}_+^{m\times n}} \sum_{i=1}^n\sum_{j=1}^m c(x_i,y_j)\pi_{i,j}, \text{ s.t. } \mathbf{1}^T\pi = \mathbf{a}^T, \pi\mathbf{1} = \mathbf{b}.$$

Suppose we stack all the columns of π in a single vector x. Let encode all the constraints $\mathbf{1}^T\pi = \mathbf{a}^T, \pi\mathbf{1} = \mathbf{b}$ on π matrix in the above optimization problem as $Ax = d$ for some $A \in \mathbb{R}^{(n+m)\times nm}$ and $d \in \mathbb{R}^{n+m}$. One can then rewrite the MK problem as

$$\min_{x\in\mathbb{R}_+^{mn}} \mathbf{c}^T x, \text{ s.t. } Ax = d$$

where $\mathbf{c} \in \mathbb{R}^{nm}$ is obtained by stacking all the columns of the cost matrix c in a single vector. Let us reparametrize x as $x = u\odot u$ for some $u \in \mathbb{R}^{nm}$, i.e., $x_i = u_i^2$ for $1 \le i \le nm$. Suppose u_t is the continuous time gradient descent to solve this minimization problem $\min_{u\in\mathbb{R}^{nm}} \|A(u\odot u) - d\|^2$ starting from the initial condition u_0 where $u_0(i) = \exp(-\frac{\mathbf{c}_i}{2\epsilon})$, i.e.,

$$\frac{d}{dt}u_t = -u_t \odot \Big(A^T(A(u\odot u) - d)\Big).$$

Show that if $u_t \to u_\infty$, then,

$$u_\infty \odot u_\infty = \text{argmin}_{x\in\mathbb{R}_+^{mn}, Ax=d}\Big\{\mathbf{c}^T x + \epsilon\sum_{i=1}^{nm} x_i \log(x_i/e)\Big\}.$$

Also code this gradient descent to see if it actually work.

2. (**Linear Convergence of Sinkhorn for bounded cost**) For any $\phi \in L^1(\mu)$ and $\psi \in L^1(\nu)$, define

$$G(\phi,\psi) = \int \phi d\mu + \int \psi d\nu - \int e^{\phi(x)+\psi(y)-c(x,y)} d(\mu(x)\otimes\nu(y)) + 1.$$

Let (ϕ_*,ψ_*) be the unique EOT potential between μ and ν w.r.t. an uniformly bounded cost function c. Let (ϕ_t,ψ_t) be the Sinkhorn potential upto the time step t. We seek to show that

$$G(\phi_*,\psi_*) - G(\phi_t,\psi_t) \le \beta^t\big(G(\phi_*,\psi_*) - G(\phi_0,\psi_0)\big)$$
$$\|\phi_* - \phi_t\|^2_{L^2(\mu)} + \|\psi_* - \psi_t\|^2_{L^2(\nu)} \le \beta_0\beta^t\big(G(\phi_*,\psi_*) - G(\phi_0,\psi_0)\big)$$

where $\beta := 1 - e^{-24\|c\|_\infty} \in (0,1)$ and $\beta_0 := 2e^{6\|c\|_\infty}$. We break it down into few stages:

- **Step 1** Show that for any $\phi, \phi' \in L^2(\mu)$ and $\psi, \psi' \in L^2(\nu)$ satisfying $\phi \oplus \psi - c \geq -\alpha$ and $\phi' \oplus \psi' - c \geq -\alpha$,

$$G(\phi', \psi') - G(\phi, \psi) \geq \int \partial_1 G(\phi', \psi')(x)[\phi'(x) - \phi(x)]\mu(dx) + \int \partial_2 G(\phi', \psi')(y)[\psi'(y) - \psi(y)]\nu(dy) + \frac{e^{-\alpha}}{2}(\|\phi' - \phi\|^2_{L(\mu)} + \|\psi - \psi'\|^2_{L^2(\nu)})$$

where

$$\partial_1 G(\phi', \psi')(x) = 1 - \int e^{\phi'(x)+\psi'(y)-c(x,y)}\nu(dx)$$
$$\partial_2 G(\phi', \psi')(y) = 1 - \int e^{\phi'(x)+\psi'(y)-c(x,y)}\mu(dx)$$

- **Step 2** Now show that $\phi_t \oplus \psi_t - c \geq -6\|c\|_\infty$ and $\phi_t \oplus \psi_{t+1} - c \geq -6\|c\|_\infty$ for all $t \geq 0$.
- **Step 3** Set $\sigma := e^{-6\|c\|_\infty}$. Use the previous two steps to show that

$$G(\phi_t, \psi_t) - G(\phi_*, \psi_*) \geq -\frac{1}{2\sigma}\|\partial_2 G(\phi_t, \psi_t) - \partial_2 G(\phi_t, \psi_{t+1})\|^2 + \frac{\sigma}{2}\|\phi_t - \phi_*\|_2^2.$$

- **Step 4** Now show that

$$\|\partial_2 G(\phi_t, \psi_t) - G(\phi_t, \psi_{t+1})\|^2_{L^2(\mu)} \leq \frac{1}{\sigma^2}\|\psi_t - \psi_{t+1}\|^2_{L^2(\nu)}.$$

- **Step 5** Use again the Taylor expansion of G like as in the first step to show that

$$G(\phi_{t+1}, \psi_{t+1}) - G(\phi_t, \psi_t) \geq \frac{\sigma}{2}(\|\phi_{t+1} - \phi_t\|^2_{L(\mu)} + \|\psi_{t+1} - \psi_t\|^2_{L^2(\nu)})$$

- **Step 6** Combine last three steps to show that

$$G(\phi_*, \psi_*) - G(\phi_t, \psi_t) \leq \frac{1}{\sigma^4}(G(\phi_{t+1}, \psi_{t+1}) - G(\phi_t, \psi_t)).$$

From Step 6, derive the recursive relation $\Delta_t \leq \frac{1}{\sigma^4}(\Delta_{t+1} - \Delta_t)$ for $\Delta_t := G(\phi_*, \psi_*) - G(\phi_t, \psi_t)$ and hence, $\Delta_t \leq (1-\sigma^4)^t \Delta_0$. Use similar argument in Step 5 to show that

$$(\|\phi_t - \phi_*\|^2_{L(\mu)} + \|\psi_t - \psi_*\|^2_{L^2(\nu)}) \leq \frac{2}{\sigma}\big(G(\phi_*, \psi_*) - G(\phi_t, \psi_t)\big) \leq \frac{2}{\sigma}(1-\sigma^4)^t.$$

3. (**Sinkhorn as Mirror Gradient Descent**) Consider a differentiable function $H : \mathcal{P}^+(\mathcal{Y}) \to \mathbb{R}$ that we wish to minimize without constraints ($\mathcal{P}^+(\mathcal{Y})$ denotes the set of positive measure on $\mathcal{Y}$). Let $G : \mathcal{P}^+(\mathcal{Y}) \to \mathbb{R}$ be a differentiable strictly convex function. The *Bregman divergence* of G is defined by

$$G(\rho_1|\rho_2) = G(\rho_1) - G(\rho_2) - \langle G'(\rho_2), \rho_1 - \rho_2 \rangle.$$

Here G' denotes the first variation of G, i.e., $G'(\rho)(h) = \lim_{\epsilon \to 0}(G(\rho + \epsilon\delta_h) - G(\rho))/\epsilon$. The gradient descent iteration with a Bregman divergence based on G, also called *mirror descent*, takes the form

$$\rho_{n+1} = \operatorname{argmin}_\rho \Big\{ H(\rho_n) + \langle H'(\rho_n), \rho - \rho_n \rangle + G(\rho|\rho_n), \Big\}$$

for all $n \geq 0$, where H' denotes the Frechét derivative of H. The optimality conditions are given by

$$G'(\rho_{n+1}) - G'(\rho_n) = -H'(\rho_n).$$

Recall the Sinkhorn updates (ϕ_t, ψ_t) (to compute EOT between μ and ν) where $\phi_0 = 0$. For any probability measure $\omega \in \mathcal{P}(\mathcal{Y})$, define

$$F(\omega) := \sup_{\phi,\psi} \Big\{ \int \phi d\mu + \int \psi d\omega - \int \int (e^{\phi(x)+\psi(y)-c(x,y)} - 1) d\mu(x) d\omega(y) \Big\}.$$

Let ν_t be the marginal of $\pi_{2t} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ where $\frac{d\pi_{2t}}{d(\mu \otimes \nu)}(x,y) \propto e^{-c(x,y)+\phi_{t-1}(x)+\psi_t(y)}$. Show that

$$F'(\nu_{t+1}) - F'(\nu_t) = -\mathrm{KL}'(\nu_t||\nu)$$