Lecture on Sinkhorn

Flavien LÉGER

math+econ+code,

June 14 2022

1. Introduction and notation

Exogenous: $(\mathcal{X}, n), (\mathcal{Y}, m), \Phi_{xy}, T > 0$.

Assume that

$$\sum_{x \in \mathcal{X}} n_x = 1, \quad \sum_{y \in Y} m_y = 1.$$

Choo-Siow model = entropic optimal transport in applied math

$$\min_{\mu \in \Pi(n,m)} \sum_{x,y} -\Phi_{xy}\mu_{xy} + 2T\mu_{xy} \log \mu_{xy}, \tag{P}$$

where

$$\Pi(n, m) = \{\mu \in \mathbb{R}^{\mathcal{X}\times\mathcal{Y}} : \mu_{xy} \geq 0, \sum_y \mu_{xy} = n_x, \sum_x \mu_{xy} = m_y\}.$$

2. Minimum entropy

Definition. Let $\mu, \tilde{\mu}$ be two measures over a finite set Z such that

$$\sum_{z \in Z} \mu_z = \sum_{z \in Z} \tilde{\mu}_z = 1.$$

The *Kullback--Leibler divergence* or *relative entropy* is defined by

$$\mathrm{KL}(\mu|\tilde{\mu}) = \sum_{z \in Z} \mu_z \log(\mu_z/\tilde{\mu}_z)$$

Fact.

- $\mathrm{KL}(\mu|\tilde{\mu}) \geq 0$
- $\mathrm{KL}(\mu|\tilde{\mu}) \neq \mathrm{KL}(\tilde{\mu}|\mu)$.

Proof. Use Jensen's inequality

$$\mathrm{KL}(\mu|\tilde{\mu}) = \sum_z -\log(\tilde{\mu}_z/\mu_z)\mu_z \geq -\log(\sum_z \tilde{\mu}_z/\mu_z \cdot \mu_z) = -\log(1) = 0.$$

Idea: KL is like a distance between μ and $\tilde{\mu}$.

We now rewrite (P) as a minimum entropy problem:

$$
\begin{aligned}
\min_{\mu\in\Pi(n,m)} \sum_{x,y} -\Phi_{xy}\mu_{xy} + 2T\mu_{xy}\log\mu_{xy} &= \min_{\mu\in\Pi(n,m)} \sum_{x,y}(C-\Phi_{xy})\mu_{xy} + 2T\mu_{xy}\log\mu_{xy} - C \\
&= \min_{\mu\in\Pi(n,m)} 2T\sum_{x,y} \frac{C-\Phi_{xy}}{2T}\mu_{xy} + \mu_{xy}\log\mu_{xy} - C \\
&= \min_{\mu\in\Pi(n,m)} 2T\sum_{x,y} \mu_{xy}\log\left(\frac{\mu_{xy}}{e^{\frac{\Phi_{xy}-C}{2T}}}\right) - C
\end{aligned}
$$

Choose $C > 0$ such that

$$\sum_{x,y} e^{\frac{\Phi_{xy}-C}{2T}} = 1,$$

i.e. $C = 2T\log(\sum_{xy} e^{\frac{\Phi_{xy}}{2T}})$,and let

$$R_{xy} = e^{\frac{\Phi_{xy}-C}{2T}}$$

We showed:

$$(\mathrm{P}) = \min_{\mu\in\Pi(n,m)} 2T\,\mathrm{KL}(\mu|R) - C$$

From now on we suppose that $2T = 1$ and $C = 0$

Fact. Problem (P) is equivalent to the minimum entropy problem

$$\min_{\mu\in\Pi(n,m)} \mathrm{KL}(\mu|R)$$

with R a reference measure such that $\sum_{x,y} R_{xy} = 1$.

3. Dual problem and Sinkhorn

Any convex minimization problem admits a dual concave maximization problem. Here:

$$\max_{u,v} \sum_x u_x n_x + \sum_y v_y m_y - \sum_{x,y} e^{u_x+v_y} R_{xy} \tag{D}$$

Duality link. Primal variable μ and dual variables u, v are linked by the formula

$$\mu_{xy} = e^{u_x+v_y} R_{xy}.$$

We can recover μ from u and v.

Algorithm. Good thing about (D): unconstrained maximization.

Let $D(u,v) = \sum_x u_x n_x + \sum_y v_y m_y - \sum_{x,y} e^{u_x+v_y} R_{xy}$.

The Sinkhorn algorithm solves (D) by alternating maximization of $D(u,v)$:

Given iterates (u^k, v^k), compute

$$u^{k+1} = \operatorname*{argmax}_u D(u, v^k) \tag{Sink1}$$

$$v^{k+1} = \operatorname*{argmax}_v D(u^{k+1}, v) \tag{Sink2}$$

Question: are the updates simple to compute and in close form? YES

Recall the primal problem

$$\min_{\mu\in\Pi(n,m)} \mathrm{KL}(\mu|R),$$

where $\mu \in \Pi(n,m)$ means $\sum_y \mu_{xy} = n_x$ and $\sum_x \mu_{xy} = m_y$. We define

$$\Pi(n,*) = \{\mu_{xy} \geq 0 : \sum_y \mu_{xy} = n_x\}$$

and

$$\Pi(*,m) = \{\mu_{xy} \geq 0 : \sum_x \mu_{xy} = m_y\}$$

$\Pi(n,*)$ consists of nonnegative matrices whose sums along the rows is n.

$\Pi(*,m)$ consists of nonnegative matrices whose sums along the columns is m.

$\Pi(n,m) = \Pi(n,*) \cap \Pi(*,m)$.

Fact. Sinkhorn's updates correspond to alternating rescaling of the rows/columns.

Proof.

At step k we have constructed u^k, v^k. The corresponding primal quantity is

$$\mu^k_{xy} = e^{u^k_x+v^k_y} R_{xy}$$

Compute

$$\frac{\partial D(u,v)}{\partial u_x} = n_x - \sum_y e^{u_x+v_y} R_{xy}$$

So (Sink1) is equivalent to

$$\sum_y e^{u^{k+1}_x+v^k_y} R_{xy} = n_x.$$

Write it as

$$e^{u^{k+1}_x-u^k_x} \sum_y e^{u^k_x+v^k_y} R_{xy} = n_x,$$

i.e.

$$e^{u_x^{k+1}-u_x^k} \sum_y \mu_{xy}^k = n_x.$$

Let

$$\tilde{\mu}_{xy}^k = e^{u_x^{k+1}-u_x^k} \mu_{xy}^k$$

Then $\tilde{\mu}^k$ is a rescaling of the rows of μ^k such that

$$\tilde{\mu}^k \in \Pi(n, *).$$

Similarly the update (Sink2) is

$$\sum_x e^{u_x^{k+1}+v_y^{k+1}} R_{xy} = m_y,$$

i.e.

$$e^{v_y^{k+1}-v_y^k} \sum_x \tilde{\mu}_{xy}^k = m_y,$$

We let

$$\mu_{xy}^{k+1} = e^{v_y^{k+1}-v_y^k} \tilde{\mu}_{xy}^k$$

which is a rescaling of the columns of $\tilde{\mu}^k$ such that

$$\mu_{k+1} \in \Pi(*, m).$$

Remark. we thus see that Sinkhorn can be implemented with primal variables μ^k or dual variables u^k, v^k.

In optimal transport Φ and thus R is often given by a formula. In this case dual variables are better.

4. Sinkhorn as entropic projections

Recall our primal problem

$$\min_\mu \mathrm{KL}(\mu|R)$$

over the constraint

$$\mu \in \Pi(n, *) \cap \Pi(*, m)$$

Recall Sinkhorn: given μ^k, it computes

- $\tilde{\mu}^k \in \Pi(n, *)$ by scaling the rows of μ^k,
- $\mu^{k+1} \in \Pi(*, m)$ by scaling the columns of $\tilde{\mu}^k$.

Fact. Sinkhorn can be seen as the entropic projections

$$
\begin{aligned}
\tilde{\mu}^k &= \underset{\mu \in \Pi(n,*)}{\operatorname{argmin}} \operatorname{KL}(\mu|\mu^k), \\
\mu^{k+1} &= \underset{\mu \in \Pi(*,m)}{\operatorname{argmin}} \operatorname{KL}(\mu|\tilde{\mu}^k).
\end{aligned}
$$

Proof. Let's look at

$$
\min_{\mu \in \Pi(n,*)} \operatorname{KL}(\mu|\mu^k) = \min_{\mu \in \Pi(n,*)} \sum_{x,y} \mu_{xy} \log(\mu_{xy}/\mu^k_{xy}),
$$

and recall that the constraint means $\sum_y \mu_{xy} = n_x$. Introduce Lagrange multiplier λ_x and write

$$
\min_{\mu \in \Pi(n,*)} \sum_{x,y} \mu_{xy} \log(\mu_{xy}/\mu^k_{xy}) = \min_{\mu \geq 0} \max_{\lambda} \sum_{x,y} \mu_{xy} \log(\mu_{xy}/\mu^k_{xy}) + \sum_x \lambda_x (n_x - \sum_y \mu_{xy})
$$

We find

$$
\max_{\lambda} \min_{\mu \geq 0} \sum_{x,y} \mu_{xy} \log(\mu_{xy}/\mu^k_{xy}) - \lambda_x \mu_{xy} + \sum_x \lambda_x n_x.
$$

Derivative w.r.t. μ_{xy}:

$$
\log(\mu_{xy}/\mu^k_{xy}) = \lambda_x - 1,
$$

i.e.

$$
\mu_{xy} = \mu^k_{xy} e^{\lambda_x - 1}
$$

So $\tilde{\mu}^k_{xy}$ is a rescaling of the rows of μ^k and belongs to $\Pi(n,*)$, this is exactly step (Sink1).

POCS. Sinkhorn is thus a generalization to KL of the classical projection onto convex sets (POCS) algorithm.

C and D two convex subsets of $\mathbb{R}^d$ with $C \cap D \neq \emptyset$.

Euclidean norm $||q||^2 = \sum_i q_i^2$.

Want to find $q^* \in C \cap D$.

$$
\tilde{q}^k = \underset{q \in C}{\operatorname{argmin}} ||q - q^k||^2,
$$

$$q^{k+1} = \underset{q \in D}{\operatorname{argmin}} \|q - \tilde{q}^k\|^2.$$

Then $q^k \to q^*$ with $q^* \in C \cap D$ closest to initial q^0.

Sinkhorn is POCS where regular projections are replaced with *entropic projections*. For Sinkhorn $C = \Pi(n, *)$ and $D = \Pi(*, m)$ which are convex

1