Self-study: Entropic regularization of optimal transport

Tianpei Xie

Aug. 17th., 2022

Contents

1 Recall: Optimal transport problem

Recall the problem of optimal transport:

- The **primal problem** for discrete measures, $\alpha = \sum_{i=1}^{n} a_i \delta_{\boldsymbol{x}_i}$ and $\beta := \sum_{i=1}^{m} b_i \delta_{\boldsymbol{y}_i}$,

$$\begin{aligned} \min_{\boldsymbol{P} \in \mathbb{R}_+^{n \times m}} \langle \boldsymbol{P} \,, \boldsymbol{C} \rangle = \sum_{i,j} C_{i,j} P_{i,j} \qquad (1) \\ \text{s.t. } \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a} \\ \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b} \\ P_{i,j} \geq 0 \end{aligned}$$

where $\boldsymbol{C}_{n,m} := [C_{i,j}]_{i \in [1:n], j \in [1:m]}$, $C_{i,j} := c(\boldsymbol{x}_i, \boldsymbol{y}_j) \geq 0$. The feasible set is defined as

$$U(\boldsymbol{a}, \boldsymbol{b}) := \left\{ \boldsymbol{P} \in \mathbb{R}_+^{n \times m} : \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}, \ \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b} \right\} \qquad (2)$$

We can define

$$\boldsymbol{A} = \left[\begin{array}{c} \mathbf{1}_n^T \otimes \boldsymbol{I}_m \\ \boldsymbol{I}_n \otimes \mathbf{1}_m^T \end{array} \right] = \left[\begin{array}{ccc} \boldsymbol{I}_m & \dots & \boldsymbol{I}_m \\ \boldsymbol{I}_n & \dots & \boldsymbol{I}_n \end{array} \right]_{nm \times nm}$$

and $\boldsymbol{p} = [P_{i*(m-1)+j}] \in \mathbb{R}^{nm \times 1}$, and $\boldsymbol{c} = [C_{i*(m-1)+j}] \in \mathbb{R}^{nm \times 1}$. Then the primal problem can be writen in matrix form

$$\begin{aligned} \min_{\boldsymbol{p} \in \mathbb{R}_+^{nm \times 1}} \boldsymbol{c}^T \boldsymbol{p} \qquad (3) \\ \text{s.t. } \boldsymbol{A}\boldsymbol{p} = \left[\begin{array}{c} \boldsymbol{a} \\ \boldsymbol{b} \end{array} \right] \end{aligned}$$

- The dual problem for optimal transport:

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m} \langle \boldsymbol{\lambda} \,, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu} \,, \boldsymbol{b} \rangle \qquad (4) \\ \text{s.t. } \lambda_i + \mu_j \leq C_{i,j} \quad \forall \, i \in [1:n], j \in [1:m] \end{aligned}$$

where $\boldsymbol{\lambda} = [\lambda_i]_n$, $\boldsymbol{\mu} = [\mu_j]_m$ are **dual variables** (slack variables) for marginal distribution constrain $\boldsymbol{a}$ and $\boldsymbol{b}$. We denote $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} := \boldsymbol{\lambda}\mathbf{1}_m^T + \mathbf{1}_n \boldsymbol{\mu}^T \in \mathbb{R}^{n \times m}$ so that the linear constraints is $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} \leq \boldsymbol{C}$. Such dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are often referred to as "***Kantorovich potentials***." The feasible set of the dual problem is defined as

$$R(\boldsymbol{C}) := \{ \boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m : \boldsymbol{\lambda} \oplus \boldsymbol{\mu} \leq \boldsymbol{C} \} \qquad (5)$$

where $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} = \boldsymbol{\lambda}\mathbf{1}_m + \mathbf{1}_n \boldsymbol{\mu}^T$.

Similarly, we have the dual problem in matrix form, when we define $\boldsymbol{h} = \left[\begin{array}{c} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{array} \right] \in \mathbb{R}^{nm \times 1}$

$$\begin{aligned} \max_{\boldsymbol{h} \in \mathbb{R}^{nm \times 1}} \left[\begin{array}{cc} \boldsymbol{a}^T & \boldsymbol{b}^T \end{array} \right] \boldsymbol{h} \qquad (6) \\ \text{s.t. } \boldsymbol{A}^T \boldsymbol{h} \leq \boldsymbol{c} \end{aligned}$$

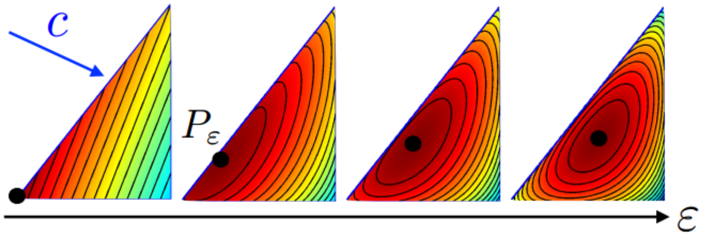


Figure 1: The effect of entropy regularization is to diffuse the results.

2 Entropic regularization of optimal transport

This section introduces a family of numerical schemes to approximate solutions to Kantorovich formulation of optimal transport and its many generalizations. It operates by adding an entropic regularization penalty to the original problem. This regularization has several important advantages:

- the minimization of the regularized problem can be solved using a simple *alternate minimization* scheme, which only required matrix-vector multiplication,
- the resulting approximate distance is *smooth* with respect to input histogram weights and positions of the Diracs and can be differentiated using *automatic differentiation*.

2.1 Entropic regularization

We introduce addtional entropic regularization term to the primal problem (1):

$$H(\boldsymbol{P}) := -\sum_{i,j} P_{i,j}\left(\log(P_{i,j}) - 1\right) \tag{7}$$

The function H is 1-***strongly concave***, because its Hessian is $\partial^2 H(\boldsymbol{P}) = \operatorname{diag}(1/P_{i,j}) > 0$ and $P_{i,j} \le 1$.

The idea of the entropic regularization of optimal transport is to use $-H$ as a regularizing function to obtain approximate solutions to the original transport problem (1):

$$L^\epsilon_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}\,, \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{P}) \tag{8}$$

This objective function is **$\underline{\epsilon}$-convex**, and thus has a unique optimal solution. The idea of adding the entropic regularization is to **diffuse** the solution to form a more *"blurred"* prediction of traffic given marginals and transportation costs. Figure 1 shows the effect of entropic regularization when ϵ is large. The entropy term pushes the optimal solution *away from the boundary* to **interial point of the feasible region**.

The convergence is guaranteed

Proposition 2.1 *The unique solution $\boldsymbol{P}_\epsilon$ of (8) converges to the optimal solution with* ***maximal entropy*** *within the set of all optimal solutions of the Kantorovich problem, namely*

$$\boldsymbol{P}_\epsilon \overset{\epsilon \to 0}{\to} \arg\min_{\boldsymbol{P}} \left\{ -H(\boldsymbol{P}) : \ \boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b}), \langle \boldsymbol{P}\,, \boldsymbol{C} \rangle = L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) \right\} \tag{9}$$

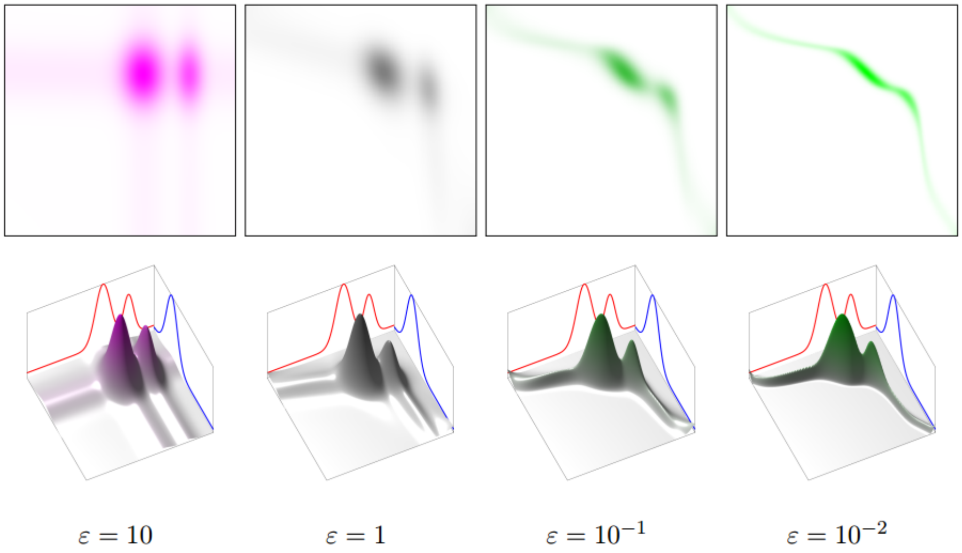


Figure 4.2: Impact of ε on the couplings between two 1-D densities, illustrating Proposition 4.1. Top row: between two 1-D densities. Bottom row: between two 2-D discrete empirical densities with the same number $n = m$ of points (only entries of the optimal $(\mathbf{P}_{i,j})_{i,j}$ above a small threshold are displayed as segments between x_i and y_j).

Figure 2: The effect of entropy regularization on optimal coupling matrix

so that in particular,

$$\lim_{\epsilon \to 0} L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})$$

Also we have

$$\lim_{\epsilon \to \infty} \boldsymbol{P}_{\epsilon} = \boldsymbol{a} \otimes \boldsymbol{b} = \boldsymbol{a}\boldsymbol{b}^T \tag{10}$$

We see that

- For small $\epsilon > 0$, the solution converges to the **maximum entropy optimal transport coupling**, as in (9)
- For large ϵ, however, (10) shows that the solution converges to the coupling with maximal entropy between two prescribed marginals $\boldsymbol{a}$, $\boldsymbol{b}$, namely the joint probability between **two independent random variables** distributed following α, β.

We also define the KL-divergence between $\boldsymbol{P}$ and the **Gibbs kernel** $\boldsymbol{K} = [\exp(-C_{i,j}/\epsilon)]$ as

$$\mathbb{KL}\left(\boldsymbol{P} \,\|\, \boldsymbol{K}\right) = \sum_{i,j} \left(P_{i,j} \log\left(\frac{P_{i,j}}{K_{i,j}}\right) - P_{i,j} + K_{i,j} \right) \tag{11}$$

We can see that the unique solution $\boldsymbol{P}_{\epsilon}$ of (8) is the projection of $\boldsymbol{K}$ onto the feasible region $U(\boldsymbol{a}, \boldsymbol{b})$, i.e.

$$\boldsymbol{P}_{\epsilon} = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \mathbb{KL}\left(\boldsymbol{P} \,\|\, \boldsymbol{K}\right) \tag{12}$$

2.2 Maximum entropy optimal transport

We can extend this to abitrary measures using the relative entropy between π and **product measure** $d(\alpha \otimes \beta)(x,y) = (d\alpha \otimes d\beta)(x,y) := d\alpha(x)d\beta(y)$ and propose a regularized counterpart to (1)

$$\mathcal{L}^{\epsilon}(\alpha,\beta) := \inf_{\pi \in U(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \epsilon\, \mathbb{KL}\left(\pi \parallel \alpha \otimes \beta\right) \tag{13}$$

where the relative entropy is a generalization of the discrete Kullback-Leibler divergence

$$\begin{aligned} \mathbb{KL}\left(\pi \parallel \xi\right) = &\int_{\mathcal{X}\times\mathcal{Y}} \log\left(\frac{d\pi}{d\xi}(x,y)\right) d\pi(x,y) \\ &+ \int_{\mathcal{X}\times\mathcal{Y}} \left(d\xi(x,y) - d\pi(x,y)\right) \end{aligned} \tag{14}$$

and $\frac{d\pi}{d\xi}$ is density of π with respect to ξ. The problem (13) measure the joint probability π again two independent probabilities $\alpha \otimes \beta$.

Using (13) the *Gibbs distributions* $\mathcal{K}$, $d\mathcal{K}(x,y) := \exp\left(-\frac{c(x,y)}{\epsilon}\right) d\alpha(x)d\beta(y)$, we reformulate the ***maximum entropy*** *optimal transport problem* (13) as

$$\mathcal{L}^{\epsilon}(\alpha,\beta) := \min_{\pi \in U(\alpha,\beta)} \mathbb{KL}\left(\pi \parallel \mathcal{K}\right). \tag{15}$$

As $\epsilon \to 0$, the unique solution to (15) converges to the maximum entropy solution $\mathcal{L}(\alpha,\beta)$.

2.3 Probability interpretation using mutual information

The second term in (13) is the ***mutual information*** $I(X;Y) := \mathbb{KL}\left(\pi \parallel \alpha \otimes \beta\right)$. Thus we can reformulate the problem (13) using probability interpretation

$$\begin{aligned} \mathcal{L}^{\epsilon}(\alpha,\beta) := \min_{(X,Y)\sim\pi} &\ \mathbb{E}_{(X,Y)}\left[c(X,Y)\right] + \epsilon\, I(X;Y) \\ \text{s.t. } &X \sim \alpha \\ &Y \sim \beta \end{aligned} \tag{16}$$

A coupling $\pi \in U(\alpha,\beta)$ describes the distribution of a couple of random variables (X,Y) defined on $(\mathcal{X},\mathcal{Y})$, where $X \sim \alpha$ and $Y \sim \beta$. As stated in (10), when $\epsilon \to \infty$, the algorithm minimizes the mutual information, i.e. $\pi_\epsilon \to \alpha \otimes \beta$. This proposes the random variables (X,Y) being **independent**.

In contrast, as $\epsilon \to 0$, π_ϵ convergence to a solution π_0 of the OT problem. Note that we know for $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and c is the euclidean distance, then there exists unique optimal Monge map T so that $Y = T(X)$ and $X = T^{-1}(Y)$. In this case, (X,Y) are in some sense **fully dependent**.

For 1-D case, we can use visualize the change of relationship between X and Y via changing ϵ. Let the c.d.f. of π be $F_\pi(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} d\pi$ and the c.d.f of α, $F_\alpha(x)$ and the c.d.f. of β, $F_\beta(y)$. Specifically, we can define a function $\xi_\pi : \mathbb{R} \to \mathbb{R}$ as

$$\xi_\pi(s,t) := F_\pi(F_\alpha^{-1}(s), F_\alpha^{-1}(t))$$

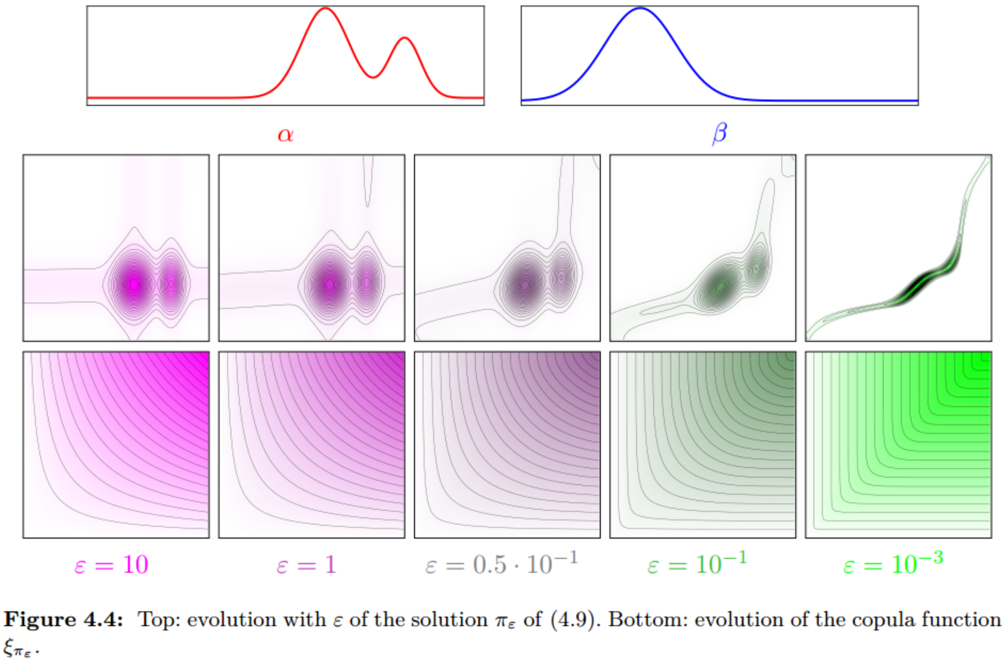


Figure 4.4: Top: evolution with ε of the solution π_ε of (4.9). Bottom: evolution of the copula function ξ_{π_ε}.

Figure 3: From independent to fully dependent: the change of ξ_π.

For *independent* variables, $\epsilon = +\infty$, i.e. $\pi = \alpha \otimes \beta$, one has $\xi_{\pi,\infty}(s,t) = s\,t$. In contrast, for *fully dependent* variables, $\epsilon \to 0$, one has $\xi_{\pi,0}(s,t) = \min(s,t)$. Figure 3 shows how entropic regularization generates copula $\xi_{\pi,\epsilon}$ interpolating between these two extreme cases.

2.4 Solution of maximum entropy optimal transport

Now we focus on solving the maximum entropy optimization problem (12)

$$\begin{aligned} &\min_{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})} \langle \boldsymbol{P}\,, \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{P}) \\ \Leftrightarrow &\min_{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})} \mathbb{KL}\left(\boldsymbol{P} \,\|\, \boldsymbol{K}\right) \end{aligned}$$

where $\boldsymbol{K} = \exp\left(-\boldsymbol{C}/\epsilon\right)$ is the Gibbs distribution.

The following proposition is for its solution

Proposition 2.2 *[Peyr and Cuturi, 2019] The solution to* (8) *is* ***unique*** *and has the form*

$$\forall i \in [1:n], j \in [1:m], \quad P_{i,j} = u_i\, K_{i,j}\, v_j \tag{17}$$

for two (unknown) scaling variable $(\boldsymbol{u}, \boldsymbol{v}) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$.

Proof: We introduce the dual variable $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ for each marginal constraint. The Lagrangian is

$$\mathcal{L}(\boldsymbol{P}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \langle \boldsymbol{P}\,, \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{P}) - \langle \boldsymbol{\lambda}\,, \boldsymbol{P}\mathbf{1}_m - \boldsymbol{a} \rangle - \langle \boldsymbol{\mu}\,, \boldsymbol{P}^T\mathbf{1}_n - \boldsymbol{b} \rangle$$

The first order condition yields

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{P}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{P}} &= \boldsymbol{C} + \epsilon \left[\log \boldsymbol{P}\right] - \boldsymbol{\lambda} \mathbf{1}_m^T - \mathbf{1}_n \boldsymbol{\mu}^T = \mathbf{0} \\ \Leftrightarrow -\frac{1}{\epsilon}\left(\boldsymbol{C} - \boldsymbol{\lambda} \mathbf{1}_m^T - \mathbf{1}_n \boldsymbol{\mu}^T\right) &= \log \boldsymbol{P} \\ \boldsymbol{P} &= \exp\left(-\frac{1}{\epsilon}\left(\boldsymbol{C} - \boldsymbol{\lambda} \mathbf{1}_m^T + \mathbf{1}_n \boldsymbol{\mu}^T\right)\right) \\ \Leftrightarrow P_{i,j} &= \exp(\frac{\lambda_i}{\epsilon}) \exp(-\frac{C_{i,j}}{\epsilon}) \exp(\frac{\mu_j}{\epsilon}) := u_i K_{i,j} v_j\end{aligned}$$

where the ***matrix scaling factor*** $\boldsymbol{u} = [\exp(\lambda_i/\epsilon)] = \exp(\boldsymbol{\lambda}/\epsilon)$ and $\boldsymbol{v} = [\exp(\mu_j/\epsilon)] = \exp(\boldsymbol{\mu}/\epsilon)$. $\blacksquare$

From (17), we can write the optimal solution in **matrix form**

$$\boldsymbol{P}^* = \underline{\operatorname{diag}(\boldsymbol{u})\, \boldsymbol{K} \operatorname{diag}(\boldsymbol{v})} \tag{18}$$

which is a **doule stochastic matrix**.

2.5 Dual problem for maximum entropy optimal transport

We can find the dual problem corresponding to the primal entropic regularized optimal transport problem in (8). In particular, we have the ***dual formulation*** for *entropic regularized optimal transport problem*

$$L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a}, \boldsymbol{b}) = \max_{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m} \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\, \boldsymbol{K} \exp\left(\boldsymbol{\mu}/\epsilon\right) \right\rangle \tag{19}$$

where $\boldsymbol{K} = \exp\left(-\boldsymbol{C}/\epsilon\right)$ is the Gibbs distribution. Note that the optimal dual solution $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ links to the scaling factor $\boldsymbol{u} = \exp(\boldsymbol{\lambda}/\epsilon)$ and $\boldsymbol{v} = \exp(\boldsymbol{\mu}/\epsilon)$ in (18).

Proof: Note that the primal optimal solution $\boldsymbol{P} = \operatorname{diag}(\boldsymbol{u})\, \boldsymbol{K} \operatorname{diag}(\boldsymbol{v})$ based on derivation in Proposition 2.2. We substitute this into the Lagrangian function

$$\begin{aligned}&\mathcal{L}(\boldsymbol{P}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \langle \boldsymbol{P}\,,\, \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{P}) - \langle \boldsymbol{\lambda}\,,\, \boldsymbol{P} \mathbf{1}_m - \boldsymbol{a} \rangle - \left\langle \boldsymbol{\mu}\,,\, \boldsymbol{P}^T \mathbf{1}_n - \boldsymbol{b} \right\rangle \\ &\Rightarrow Q((\boldsymbol{\lambda}, \boldsymbol{\mu}) = \left\langle \exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\, \left(\boldsymbol{K} \odot \boldsymbol{C}\right) \exp\left(\boldsymbol{\mu}/\epsilon\right) \right\rangle - \epsilon H\left(\operatorname{diag}\left(\exp\left(\boldsymbol{\lambda}/\epsilon\right)\right) \boldsymbol{K} \operatorname{diag}\left(\exp\left(\boldsymbol{\mu}/\epsilon\right)\right)\right)\end{aligned}$$

For the second term, we see that

$$\begin{aligned}-\epsilon H(\boldsymbol{P}) &= \left\langle \boldsymbol{P}\,,\, \log \boldsymbol{P} - \mathbf{1}\mathbf{1}^T \right\rangle \\ &= \left\langle \operatorname{diag}\left(\exp\left(\boldsymbol{\lambda}/\epsilon\right)\right) \boldsymbol{K} \operatorname{diag}\left(\exp\left(\boldsymbol{\mu}/\epsilon\right)\right)\,,\, -\boldsymbol{C} + \boldsymbol{\lambda} \mathbf{1}_m^T + \mathbf{1}_n \boldsymbol{\mu}^T - \epsilon \mathbf{1}\mathbf{1}^T \right\rangle \\ &= -\left\langle \operatorname{diag}\left(\exp\left(\boldsymbol{\lambda}/\epsilon\right)\right) \boldsymbol{K} \operatorname{diag}\left(\exp\left(\boldsymbol{\mu}/\epsilon\right)\right)\,,\, \boldsymbol{C} \right\rangle + \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\, \boldsymbol{K} \exp\left(\boldsymbol{\mu}/\epsilon\right) \right\rangle \\ &= -\left\langle \exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\, \left(\boldsymbol{K} \odot \boldsymbol{C}\right) \exp\left(\boldsymbol{\mu}/\epsilon\right) \right\rangle + \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\, \boldsymbol{K} \exp\left(\boldsymbol{\mu}/\epsilon\right) \right\rangle\end{aligned}$$

which cancels out the first term in equation above. So we have the form of dual objective functions. $\blacksquare$

For arbitrary measures we have the **dual formulation** for $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, constant $\epsilon > 0$

$$\sup_{(\lambda, \mu) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \lambda d\alpha + \int_{\mathcal{Y}} \mu d\beta - \epsilon \int_{\mathcal{X} \times \mathcal{Y}} \left(\exp\left(\frac{-c + \lambda \oplus \mu}{\epsilon}\right) - 1\right) d\alpha d\beta. \tag{20}$$

where $(\lambda \oplus \mu)(x,y) = \lambda(x) + \mu(y)$. In the case $\int_{\mathcal{X}} d\alpha = 1$ and $\int_{\mathcal{Y}} d\beta = 1$, we have dual formulation

$$\sup_{(\lambda,\mu)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \lambda d\alpha + \int_{\mathcal{Y}} \mu d\beta + \text{soft-min}_{\epsilon}\left\{c - \lambda \oplus \mu\right\} \tag{21}$$

where the soft-min operator on $f \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$

$$\text{soft-min}_{\epsilon}\left\{f\right\} := -\epsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(-f/\epsilon\right) d\alpha d\beta. \tag{22}$$

As $\epsilon \to 0$, soft-min$_{\epsilon} \to$ min, as used in the unregularized and unconstrained formulation.

The probability interpretation of dual is as below

$$\sup_{(\lambda,\mu)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \mathbb{E}_{X\sim\alpha}\left[\lambda(X)\right] + \mathbb{E}_{Y\sim\beta}\left[\mu(Y)\right] - \epsilon\mathbb{E}_{X\sim\alpha,Y\sim\beta}\left[\exp\left(\frac{-c(X,Y) + \lambda(X) + \mu(Y)}{\epsilon}\right)\right]. \tag{23}$$

The entropic dual (19) and (20) is a smooth unconstrained concave maximization problem, which approximates the original Kantorovich dual (4). The following proposition states that the optimal solution of regularized problem is still feasible for original dual problem.

Proposition 2.3 *Any pair of optimal solution to* (19) $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ *is a feasible solution to the original dual problem* (4)*, i.e.* $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*) \in R(\boldsymbol{C})$*. For any* $\epsilon > 0$*,*

$$\langle \boldsymbol{\lambda}_* \,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}_* \,,\, \boldsymbol{b} \rangle \leq L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}).$$

On the other hand, a chief **advantage** of the regularized transportation cost $L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})$ defined in (8) is that it is *smooth* and *convex*, which makes it a perfect fit for integrating as a loss function in variational problems.

Proposition 2.4 *Given* $\boldsymbol{C}$*, the optimal value* $L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})$ *is* ***convex*** *function with respective to* $(\boldsymbol{a}, \boldsymbol{b})$ *for any* $\epsilon \geq 0$*. If* $\epsilon > 0$*, then*

$$\nabla_{(\boldsymbol{a},\boldsymbol{b})} L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \begin{bmatrix} \boldsymbol{\lambda}_* \\ \boldsymbol{\mu}_* \end{bmatrix} \tag{24}$$

where $(\boldsymbol{\lambda}_*, \boldsymbol{\mu}_*)$ *are the optimal solutions of regularized dual problem* (19) *chosen so that their coordinates sum to* 0*.*

Note that $L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b})$ is referred as ***Sinkhorn divergence*** in many literature [Li et al., 2021].

2.6 Barycentric projection

Consider the entropic regularized optimal transport problem (8). In order to obtain the map $T : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \mathbb{R}^d$, one can define the so-called ***barycentric projection map***:

$$T : \boldsymbol{x}_i \in \mathcal{X} \to \sum_j P_{i,j} y_j \in \mathcal{Y}, \tag{25}$$

where the input measures are discrete of the form $\alpha = \sum_{i=1}^{n} a_i \delta_{\boldsymbol{x}_i}$. Note that this map is only defined for points $\boldsymbol{x}_i$ in the support of α. For T as a permutation map, we can see that T converges to the *Monge map* when $\epsilon \to 0$.

Similarly, consider the arbitrary measure and optimization problem in (13). The solution $\pi \in U(\alpha, \beta)$ defines a density $\frac{d\pi}{d\alpha \otimes d\beta}$. The ***barycentric projection map*** is defined as

$$T : x \in \mathcal{X} \to \int_{\mathcal{Y}} y \frac{d\pi(x, y)}{d\alpha(x)\, d\beta(y)} d\beta(y), \tag{26}$$

for $\epsilon = 0$, π is supported on the graph of the Monge map. For $\epsilon > 0$, it is a smooth map.

This map has been used in imaging. It has also been used to compute approximations of principal geodesics in the space of probability measures endowed with the Wasserstein metric.

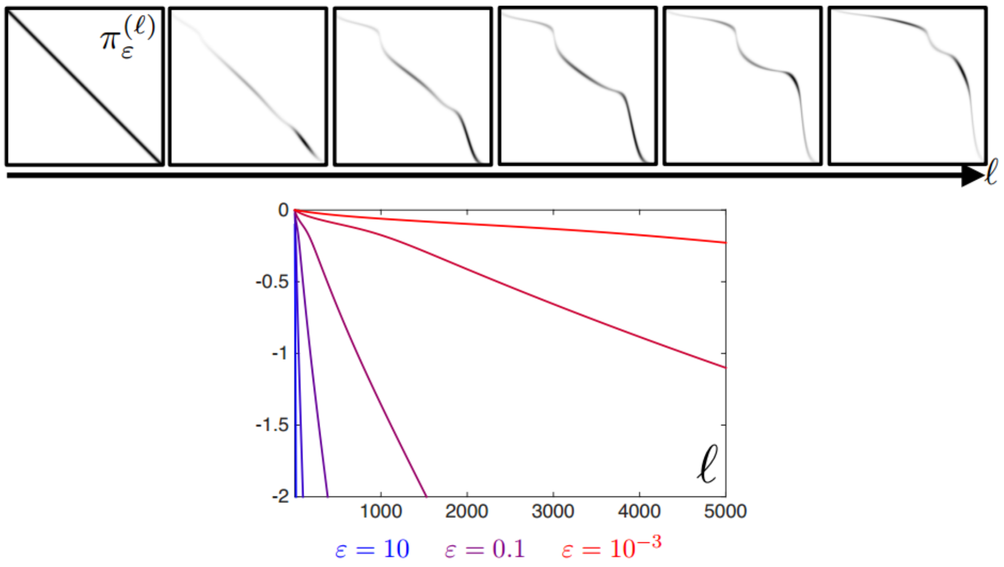


Figure 4.5: Top: evolution of the coupling $\pi_\varepsilon^{(\ell)} = \operatorname{diag}(\mathbf{u}^{(\ell)})\mathbf{K}\operatorname{diag}(\mathbf{v}^{(\ell)})$ computed at iteration ℓ of Sinkhorn's iterations, for 1-D densities on $\mathcal{X} = [0,1]$, $c(x,y) = |x-y|^2$, and $\varepsilon = 0.1$. Bottom: impact of ε the convergence rate of Sinkhorn, as measured in term of marginal constraint violation $\log(\|\pi_\varepsilon^{(\ell)}\mathbb{1}_m - \mathbf{b}\|_1)$.

Figure 4: Sinkhorn's update from Gibbs distribution to optimal transport

3 Sinkhorn's algorithm and Its convergence

3.1 Sinkhorn's algorithm

From (17), we can write the optimal solution in matrix form

$$\boldsymbol{P}^* = \underline{\operatorname{diag}(\boldsymbol{u})\,\boldsymbol{K}\operatorname{diag}(\boldsymbol{v})}$$

This solution need to satisfiy the equality constrain in $U(\boldsymbol{a},\boldsymbol{b})$:

$$\begin{aligned} \operatorname{diag}(\boldsymbol{u})\,\boldsymbol{K}\operatorname{diag}(\boldsymbol{v})\,\mathbf{1}_m = \boldsymbol{a},\;\; \operatorname{diag}(\boldsymbol{v})\,\boldsymbol{K}^T\operatorname{diag}(\boldsymbol{u})\,\mathbf{1}_n = \boldsymbol{b} \\ \Rightarrow \boldsymbol{u}\odot(\boldsymbol{K}\boldsymbol{v}) = \boldsymbol{a},\;\; \boldsymbol{v}\odot\left(\boldsymbol{K}^T\boldsymbol{u}\right) = \boldsymbol{b} \end{aligned} \tag{27}$$

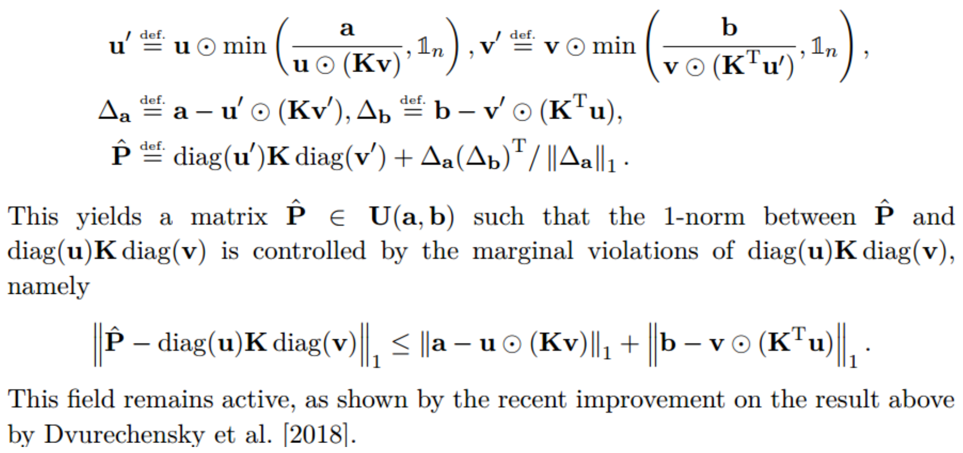
where $\odot$ means element-wise multiplication. This problem is known in the numerical analysis community as the ***matrix scaling problem*** [Franklin and Lorenz, 1989, Nemirovski and Rothblum, 1999].

The solution $(\boldsymbol{u},\boldsymbol{v})$ for equations (27) can be obtained via two iterative updates. We have the ***Sinkhorn*'s algorithm** [Sinkhorn, 1967]

$$\boldsymbol{u}_{t+1} \leftarrow \frac{\boldsymbol{a}}{\boldsymbol{K}\boldsymbol{v}_t} := \boldsymbol{a}\oslash\boldsymbol{K}\boldsymbol{v}_t \tag{28}$$

$$\boldsymbol{v}_{t+1} \leftarrow \frac{\boldsymbol{b}}{\boldsymbol{K}^T\boldsymbol{u}_{t+1}} := \boldsymbol{b}\oslash\boldsymbol{K}^T\boldsymbol{u}_{t+1} \tag{29}$$

where the right hand side is element wise division i.e. *Hadamard division* ($\oslash$). Also initialized with an arbitrary positive vector $\boldsymbol{v}_0 = \mathbf{1}_m$.



$$\mathbf{u}' \overset{\text{def.}}{=} \mathbf{u} \odot \min\left(\frac{\mathbf{a}}{\mathbf{u} \odot (\mathbf{K}\mathbf{v})}, \mathbb{1}_n\right), \mathbf{v}' \overset{\text{def.}}{=} \mathbf{v} \odot \min\left(\frac{\mathbf{b}}{\mathbf{v} \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u}')}, \mathbb{1}_n\right),$$
$$\Delta_{\mathbf{a}} \overset{\text{def.}}{=} \mathbf{a} - \mathbf{u}' \odot (\mathbf{K}\mathbf{v}'), \Delta_{\mathbf{b}} \overset{\text{def.}}{=} \mathbf{b} - \mathbf{v}' \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u}),$$
$$\hat{\mathbf{P}} \overset{\text{def.}}{=} \operatorname{diag}(\mathbf{u}')\mathbf{K}\operatorname{diag}(\mathbf{v}') + \Delta_{\mathbf{a}}(\Delta_{\mathbf{b}})^{\mathrm{T}} / \|\Delta_{\mathbf{a}}\|_1 .$$

This yields a matrix $\hat{\mathbf{P}} \in \mathbf{U}(\mathbf{a}, \mathbf{b})$ such that the 1-norm between $\hat{\mathbf{P}}$ and $\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$ is controlled by the marginal violations of $\operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})$, namely

$$\left\|\hat{\mathbf{P}} - \operatorname{diag}(\mathbf{u})\mathbf{K}\operatorname{diag}(\mathbf{v})\right\|_1 \leq \|\mathbf{a} - \mathbf{u} \odot (\mathbf{K}\mathbf{v})\|_1 + \left\|\mathbf{b} - \mathbf{v} \odot (\mathbf{K}^{\mathrm{T}}\mathbf{u})\right\|_1 .$$

This field remains active, as shown by the recent improvement on the result above by Dvurechensky et al. [2018].

Figure 5: Rouding scheme for Sinkhorn's algorithm

Note that a different initialization will likely lead to a different solution for $(\boldsymbol{u}, \boldsymbol{v})$, since $(\boldsymbol{u}, \boldsymbol{v})$ are only defined up to a multiplicative constant. It turns out, however, that these iterations converge and all result in the same optimal coupling $\operatorname{diag}(\boldsymbol{u})\boldsymbol{K}\operatorname{diag}(\boldsymbol{v})$. Figure 4 top row, shows the evolution of the coupling $\operatorname{diag}(\boldsymbol{u}_t)\boldsymbol{K}\operatorname{diag}(\boldsymbol{v}_t)$ computed by Sinkhorn iterations. It evolves from the Gibbs kernel $\boldsymbol{K}$ toward the optimal coupling solving $L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a}, \boldsymbol{b})$ by progressively shifting the mass away from the diagonal.

Note that Sinkhorn's algorithm has **advantages** in computation since it **only requires matrix-vector multiplication** and each **column** is updated **independently**. We can even combine N optimal transport problems together to form matrix-matrix multiplication and update simultaneously. All of these allow it to be computed efficiently using ***parallel computing*** such as GPUs.

3.2 Numerical analysis of Sinkhorn's algorithm

It can be showed [Peyr and Cuturi, 2019] that by setting $\epsilon = \frac{4\log(n)}{\tau}$, $O(\|\boldsymbol{C}\|_{\infty}^{3}\log(n)\tau^{-3})$ Sinkhorn iterations (with an additional rounding step to compute a valid coupling $\boldsymbol{P}' \in U(\boldsymbol{a}, \boldsymbol{b})$) are enough to ensure that $\langle \boldsymbol{P}', \boldsymbol{C}\rangle \leq L_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) + \tau$. This implies that Sinkhorn computes a τ-approximate solution of the unregularized OT problem in $\underline{O(n^2\log(n)\tau^{-3})}$ operations.

Altschuler et al introduced the rounding scheme in Figure 5.

The convergence of Sinkhorns algorithm deteriorates as $\epsilon \to 0$. In numerical practice, however, that slowdown is rarely observed in practice. This is because Sinkhorns algorithm will often fail to terminate as soon as some of the elements of the kernel $\boldsymbol{K}$ become too negligible to be stored in memory as positive numbers, and become instead null. This then affects the denominator of the updates (28) and (29), causing division by 0 error. This concern can be alleviated to some extent by carrying out computations in the **log domain**.

3.3 Bregman iterative projections

Note that $U(\boldsymbol{a},\boldsymbol{b}) = \mathcal{S}_{\boldsymbol{a}} \cup \mathcal{S}_{\boldsymbol{b}}$ where $\mathcal{S}_{\boldsymbol{a}} = \{\boldsymbol{P} : \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a}\}$ and $\mathcal{S}_{\boldsymbol{b}} = \{\boldsymbol{P} : \boldsymbol{P}^T\mathbf{1}_n = \boldsymbol{b}\}$. An alternative algorithm for solving problem (8) is via **Bregman iterative projections**:

$$\boldsymbol{P}_{t+1} \leftarrow \mathrm{Proj}_{\mathcal{S}_{\boldsymbol{a}}}\left(\boldsymbol{P}_t\right) := \arg\min_{\boldsymbol{P}\in\mathcal{S}_{\boldsymbol{a}}} \mathbb{KL}\left(\boldsymbol{P} \,\|\, \boldsymbol{P}_t\right) \tag{30}$$

$$\boldsymbol{P}_{t+2} \leftarrow \mathrm{Proj}_{\mathcal{S}_{\boldsymbol{b}}}\left(\boldsymbol{P}_{t+1}\right) := \arg\min_{\boldsymbol{P}\in\mathcal{S}_{\boldsymbol{b}}} \mathbb{KL}\left(\boldsymbol{P} \,\|\, \boldsymbol{P}_{t+1}\right) \tag{31}$$

where $\mathrm{Proj}_{\mathcal{A}}(\cdot)$ is the projection via Bregman divergence (KL divergence in this case.) onto the affine set $\mathcal{A}$. These iterates are equivalent to Sinkhorn iterations (28) and (29) since defining

$$\begin{aligned} \boldsymbol{P}_{2t} &= \mathrm{diag}\left(\boldsymbol{u}_t\right)\boldsymbol{K}\mathrm{diag}\left(\boldsymbol{v}_t\right) \\ \Rightarrow \boldsymbol{P}_{2t+1} &= \mathrm{diag}\left(\boldsymbol{u}_{t+1}\right)\boldsymbol{K}\mathrm{diag}\left(\boldsymbol{v}_t\right) \\ \Rightarrow \boldsymbol{P}_{2t+2} &= \mathrm{diag}\left(\boldsymbol{u}_{t+1}\right)\boldsymbol{K}\mathrm{diag}\left(\boldsymbol{v}_{t+1}\right) \end{aligned}$$

In practice, however, one should prefer using (28) and (29), which only requires *manipulating* ***scaling vectors*** and ***multiplication*** *against a* ***Gibbs kernel***, which can often be accelerated.

3.4 Proximal point algorithm

In order to approximate a solution of the original unregularized ($\epsilon = 0$) problem (1), it is possible to use ***iteratively*** *the Sinkhorn algorithm*, using the so-called **proximal point algorithm** for the KL metric.

Denote the unconstrained objective function as

$$F(\boldsymbol{P}) := \langle \boldsymbol{P}\,,\, \boldsymbol{C} \rangle + \mathbb{1}\left\{\boldsymbol{P} \in U(\boldsymbol{a},\boldsymbol{b})\right\}$$

The ***proximal point iterations*** for the KL divergence computes a minimizer of F

$$\boldsymbol{P}_{t+1} \leftarrow \mathrm{Proj}_{F(\boldsymbol{P})}\left(\boldsymbol{P}_t\right) := \arg\min_{\boldsymbol{P}\in\mathbb{R}_+^{n\times m}} \mathbb{KL}\left(\boldsymbol{P} \,\|\, \boldsymbol{P}_t\right) + \frac{1}{\epsilon}F(\boldsymbol{P}) \tag{32}$$

The proximal point algorithm is the most basic *proximal splitting method.* Initially introduced for the Euclidean metric, it extends to any Bregman divergence, so in particular it can be applied here for the KL divergence.

The optimization appearing in (32) is very similar to the entropy regularized problem (8), with the relative entropy $\mathbb{KL}\left(\cdot \,\|\, \boldsymbol{P}_t\right)$ used in place of the negative entropy $-H$. Since the optimal solution for each subproblem is of similar form, we can still use the Sinkhorn algorithm, with $\boldsymbol{K} := \exp(-\frac{1}{\epsilon}\boldsymbol{C}) \odot \boldsymbol{P}_t$. Iterations (32) can thus be implemented by **running the Sinkhorn algorithm at each iteration**. Assuming for simplicity $\boldsymbol{P}_0 = \mathbf{1}_n\mathbf{1}_m^T$, these iterations thus have the form

$$\boldsymbol{P}_{t+1} = \mathrm{diag}\left(\boldsymbol{u}_t \odot \ldots \odot \boldsymbol{u}_0\right)\left[\exp(-\frac{1}{\epsilon/(1+t)}\boldsymbol{C}) \odot \boldsymbol{P}_t\right]\mathrm{diag}\left(\boldsymbol{v}_t \odot \ldots \odot \boldsymbol{v}_0\right)$$

Here the kernel is $\exp(-\frac{1}{\epsilon/(1+t)}\boldsymbol{C})$ and the temperature of the kernel decays with parameter ϵ/t. This method is thus tightly connected to a series of works which combine Sinkhorn with some **decaying schedule** on the regularization.

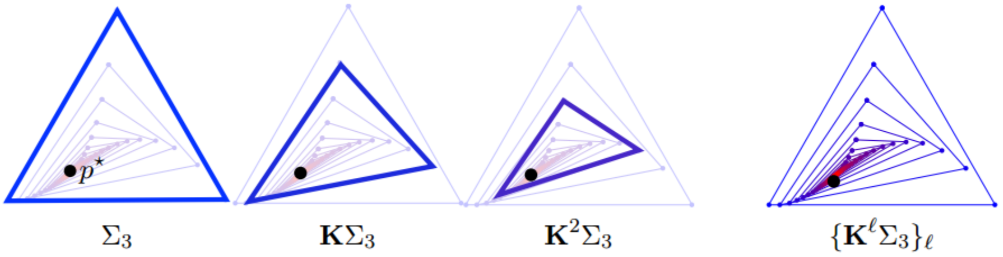


Figure 4.8: Evolution of $\mathbf{K}^\ell\Sigma_3 \to \{p^\star\}$ the invariant probability distribution of $\mathbf{K} \in \mathbb{R}^{3\times 3}_{+,*}$ with $\mathbf{K}^\top \mathbb{1}_3 = \mathbb{1}_3$.

Figure 6: Evolution of boundary of $\boldsymbol{K}_t\Delta_3$

3.5 Convergence analysis of Sinkhorn's algorithm

The analysis used a metric called **Hilbert projective metric**.

Definition The ***Hilbert metric***, also known as the ***Hilbert projective metric***, is an explicitly defined distance function on a *bounded convex subset* of the n-dimensional Euclidean space $\mathbb{R}^n$.

For the space of positive vectors $\mathbb{R}^n_{+,*} \subset \mathbb{R}^n$, Hilbert projective metric on $\mathbb{R}^n_{+,*}$

$$\begin{aligned} d_{\mathcal{H}}(\boldsymbol{u},\boldsymbol{v}) &= \log \max_{i,j} \frac{u_i v_j}{u_j v_i} \quad \forall \boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^n_{+,*} \\ &= \|\log \boldsymbol{u} - \log \boldsymbol{v}\|_{\text{var}}\,, \end{aligned}$$

where $\|\boldsymbol{f}\|_{\text{var}} = \max f_i - \min f_i$. It can be shown to be a distance on the ***projective cone*** $\mathbb{R}^n_{+,*}/\sim$, where $\boldsymbol{u} \sim \boldsymbol{v}$ means that $\exists r > 0, \boldsymbol{u} = r\boldsymbol{v}$. This means that $d_{\mathcal{H}}$ satisfies the triangular inequality and $d_{\mathcal{H}}(\boldsymbol{u},\boldsymbol{v}) = 0$ if and only if $\boldsymbol{u} \sim \boldsymbol{v}$.

Theorem 3.1 *(**Global linear convergence**) [Franklin and Lorenz, 1989]*
For Sinkhorn algorithm, one has $(\boldsymbol{u}_t,\boldsymbol{v}_t) \to (\boldsymbol{u}_*,\boldsymbol{v}_*)$ and

$$d_{\mathcal{H}}(\boldsymbol{u}_t,\boldsymbol{u}_*) = O(\lambda(\boldsymbol{K})^{2t}) \qquad d_{\mathcal{H}}(\boldsymbol{v}_t,\boldsymbol{v}_*) = O(\lambda(\boldsymbol{K})^{2t}) \tag{33}$$

where $\lambda(\boldsymbol{K}) := \frac{\sqrt{\eta(\boldsymbol{K})}-1}{\sqrt{\eta(\boldsymbol{K})}+1}$ and $\eta(\boldsymbol{K}) = \max_{i,j,k,l} \frac{K_{i,k}K_{j,l}}{K_{j,k}K_{i,l}}$

One also has

$$\begin{aligned} d_{\mathcal{H}}(\boldsymbol{u}_t,\boldsymbol{u}_*) &\le \frac{d_{\mathcal{H}}(\boldsymbol{P}_t\mathbf{1}_m,\boldsymbol{a})}{1-\lambda(\boldsymbol{K})^2}, \\ d_{\mathcal{H}}(\boldsymbol{v}_t,\boldsymbol{v}_*) &\le \frac{d_{\mathcal{H}}(\boldsymbol{P}_t^T\mathbf{1}_n,\boldsymbol{b})}{1-\lambda(\boldsymbol{K})^2} \end{aligned} \tag{34}$$

where we denoted $\boldsymbol{P}_t = diag(\boldsymbol{u}_t)\,\boldsymbol{K}\,diag(\boldsymbol{v}_t)$. Last, one has

$$\|\log \boldsymbol{P}_t - \log \boldsymbol{P}_*\|_\infty \le d_{\mathcal{H}}(\boldsymbol{u}_t,\boldsymbol{u}_*) + d_{\mathcal{H}}(\boldsymbol{v}_t,\boldsymbol{v}_*) \tag{35}$$

where $\boldsymbol{P}_*$ follows (17) *is the unique solution of* (8).

3.6 Log-domain Sinkhorn to solve the dual problem

Recall the dual objective function

$$Q(\boldsymbol{\lambda}, \boldsymbol{\mu}) := \langle \boldsymbol{\lambda} \, , \, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu} \, , \, \boldsymbol{b} \rangle - \epsilon \, \langle \exp \left(\boldsymbol{\lambda} / \epsilon \right) \, , \, \boldsymbol{K} \exp \left(\boldsymbol{\mu} / \epsilon \right) \rangle \tag{36}$$

3.6.1 Sinkhorn as block coordinate ascent in dual domain

A simple approach to solving the unconstrained maximization problem (19) is to use an exact ***block coordinate ascent*** strategy, namely to update alternatively $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ to cancel the respective gradients in these variables of the objective of (19).

The gradient of dual objective w.r.t $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{\lambda}} &= \boldsymbol{a} - \epsilon \, \exp \left(\boldsymbol{\lambda} / \epsilon \right) \odot \left(\boldsymbol{K} \exp \left(\boldsymbol{\mu} / \epsilon \right) \right) & (37) \\ \frac{\partial Q(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} &= \boldsymbol{b} - \epsilon \, \exp \left(\boldsymbol{\mu} / \epsilon \right) \odot \left(\boldsymbol{K}^T \exp \left(\boldsymbol{\lambda} / \epsilon \right) \right) & (38) \end{aligned}$$

Thus <u>***block coordinate ascent in dual domain***</u> can therefore be implemented in a closed form by applying successively the following updates, starting from any arbitrary $\boldsymbol{\mu}_0$, Solving the equation $\frac{\partial Q(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{\lambda}} = \mathbf{0}$ and $\frac{\partial Q(\boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \mathbf{0}$, the updated on $\boldsymbol{\lambda}$ and on $\boldsymbol{\mu}$, respectively are obtained

$$\begin{aligned} \boldsymbol{\lambda}_{t+1} &\leftarrow \arg \min_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda}, \boldsymbol{\mu}_t) \\ &= \epsilon \log \boldsymbol{a} - \epsilon \log \left(\boldsymbol{K} \exp \left(\boldsymbol{\mu}_t / \epsilon \right) \right) & (39) \\ \boldsymbol{\mu}_{t+1} &\leftarrow \arg \min_{\boldsymbol{\mu}} Q(\boldsymbol{\lambda}_{t+1}, \boldsymbol{\mu}) \\ &= \epsilon \log \boldsymbol{b} - \epsilon \log \left(\boldsymbol{K}^T \exp \left(\boldsymbol{\lambda}_{t+1} / \epsilon \right) \right) & (40) \end{aligned}$$

Such iterations are mathematically equivalent to the Sinkhorn iterations (28) and (29) when considering the primal-dual relations highlighted in $\boldsymbol{u} = \exp(\boldsymbol{\lambda}/\epsilon)$ and $\boldsymbol{v} = \exp(\boldsymbol{\mu}/\epsilon)$. Indeed, we recover that at any iteration

$$(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t) = \epsilon(\log \boldsymbol{u}_t, \log \boldsymbol{v}_t)$$

3.6.2 Log-domain Sinkhorn

Note that the second term in (39)

$$\begin{aligned} -\epsilon [\log \left(\boldsymbol{K} \exp \left(\boldsymbol{\mu}_t / \epsilon \right) \right)]_{i,j} &= -\epsilon \log \left(\sum_j \exp(- \left(C_{i,j} - \mu_{j,t} \right) / \epsilon) \right) \\ &:= \text{soft-min}_\epsilon \left(\left[\boldsymbol{C} - \mathbf{1}_n \boldsymbol{\mu}_t^T \right]_i \right) = \text{soft-min}_{\epsilon, \text{row}} \left(\boldsymbol{C} - \mathbf{1}_n \boldsymbol{\mu}_t^T \right) \end{aligned}$$

actually choose the ***soft-min*** of difference $C_{i,j} - \mu_{j,t}$ over index j, where the **soft-min operator**

$$\text{soft-min}_\epsilon(\boldsymbol{z}) := -\epsilon \log \sum_i \exp \left(-\boldsymbol{z}_i / \epsilon \right).$$

These operations are equivalent to the entropic c-transform. Using this notation, **Sinkhorns iterates** (39) and (40) read

$$\boldsymbol{\lambda}_{t+1} \leftarrow \text{soft-min}_{\epsilon,\text{row}}\left\{\boldsymbol{C} - \mathbf{1}_n\boldsymbol{\mu}_t^T\right\} + \epsilon\log\boldsymbol{a} \tag{41}$$

$$\boldsymbol{\mu}_{t+1} \leftarrow \text{soft-min}_{\epsilon,\text{column}}\left\{\boldsymbol{C} - \boldsymbol{\lambda}_{t+1}\mathbf{1}_m^T\right\} + \epsilon\log\boldsymbol{b} \tag{42}$$

Here, we introduce the ***log-sum-exp** stabilization* trick to avoid underflow for small values of ϵ. That trick suggests that we subtract the minimum term in the soft-min operator:

$$\min_\epsilon(\boldsymbol{z}) := \bar{z} - \epsilon\log\sum_i \exp\left(-(\boldsymbol{z}_i - \bar{z})/\epsilon\right)$$

Instead of substracting $\bar{z} := \min_i z_i$ to stabilize the log-domain iterations as above, one can actually substract the previously computed scalings. This leads to the *stabilized* iteration. The **log-domain *Sinkhorn algorithm*** is described as below:

$$\boldsymbol{\lambda}_{t+1} \leftarrow \boldsymbol{\lambda}_t + \text{soft-min}_{\epsilon,\text{row}}\left\{\boldsymbol{C} - \boldsymbol{\lambda}_t\mathbf{1}^T - \mathbf{1}\boldsymbol{\mu}_t^T\right\} + \epsilon\log\boldsymbol{a} \tag{43}$$

$$\boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t + \text{soft-min}_{\epsilon,\text{column}}\left\{\boldsymbol{C} - \boldsymbol{\lambda}_{t+1}\mathbf{1}^T - \mathbf{1}\boldsymbol{\mu}_t^T\right\} + \epsilon\log\boldsymbol{b} \tag{44}$$

where $\text{soft-min}_\epsilon(\boldsymbol{z}) = -\epsilon\log\sum_i\exp\left(-\boldsymbol{z}_i/\epsilon\right)$ is the soft-min operator. The notation $\text{soft-min}_{\epsilon,\text{row}}$ means that we take soft-min for each row of matrix of $\mathbb{R}^{n\times m}$ and returning a column of size n.

In contrast to the original iterations (28) and (29), these log-domain iterations (43) and (44) are stable for arbitrary $\epsilon > 0$, because the quantity $S(\boldsymbol{\lambda},\boldsymbol{\mu}) := \boldsymbol{C} - \boldsymbol{\lambda}_{t+1}\mathbf{1}^T - \mathbf{1}\boldsymbol{\mu}_t^T$ stays bounded during the iterations.

The **downside** is that it requires nm computations of exp at each step. Computing a $\text{soft-min}_{\epsilon,\text{row}}$ or $\text{soft-min}_{\epsilon,\text{column}}$ is typically substantially ***slower*** than matrix multiplications and requires computing line by line soft-minima of matrices $\boldsymbol{S}$. There is therefore *no efficient way* to parallelize the application of Sinkhorn maps for several marginals *simultaneously.*

3.7 Sinkhorn divergence

Definition (**Sinkhorn divergences**) [Cuturi, 2013, Peyr and Cuturi, 2019].
Let $(\boldsymbol{\lambda}_*,\boldsymbol{\mu}_*)$ be optimal solutions to dual problemm (19) and $\boldsymbol{P}_*$ be the solution to primal problem (8). The *Wasserstein distance* is approximated using the following **primal** and **dual Sinkhorn divergences**:

$$\mathfrak{P}^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) := \langle\boldsymbol{P}_*\,,\,\boldsymbol{C}\rangle = \langle\exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\,(\boldsymbol{K}\odot\boldsymbol{C})\exp\left(\boldsymbol{\mu}/\epsilon\right)\rangle \tag{45}$$

$$\mathfrak{D}^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) := \langle\boldsymbol{\lambda}_*\,,\,\boldsymbol{a}\rangle + \langle\boldsymbol{\mu}_*\,,\,\boldsymbol{b}\rangle \tag{46}$$

where $\odot$ stands for the elementwise product of matrices.

Proposition 3.2 *The following relationship holds:*

$$\mathfrak{D}^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) \le L^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) \le \mathfrak{P}^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) \tag{47}$$

Furthermore

$$\mathfrak{P}^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) - \mathfrak{D}^\epsilon_{\boldsymbol{C}}(\boldsymbol{a},\boldsymbol{b}) = \epsilon(H(\boldsymbol{P}_*) - 1) \tag{48}$$

Proof: By definition of optimality, we have

$$\begin{aligned} L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) &= \mathfrak{P}_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) - \epsilon H(\boldsymbol{P}_*) \leq \mathfrak{P}_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) \\ L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) = \mathfrak{D}_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}_*/\epsilon\right) \,,\, \boldsymbol{K}\exp\left(\boldsymbol{\mu}_*/\epsilon\right)\right\rangle &\geq \mathfrak{D}_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) \end{aligned}$$

Thus

$$\begin{aligned} \mathfrak{P}_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) - \mathfrak{D}_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) &= \epsilon H(\boldsymbol{P}_*) - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}_*/\epsilon\right) \,,\, \boldsymbol{K}\exp\left(\boldsymbol{\mu}_*/\epsilon\right)\right\rangle \\ &= \epsilon(H(\boldsymbol{P}_*) - 1) \end{aligned}$$

Note that

$$\begin{aligned} \left\langle \exp\left(\boldsymbol{\lambda}_*/\epsilon\right) \,,\, \boldsymbol{K}\exp\left(\boldsymbol{\mu}_*/\epsilon\right)\right\rangle &= \sum_i \sum_j e^{\lambda_{i,*}/\epsilon} e^{-C_{i,j}/\epsilon} e^{\mu_{j,*}/\epsilon} \\ &= \sum_i \sum_j e^{(\lambda_{i,*}+\mu_{j,*}-C_{i,j})/\epsilon} \\ &= \sum_i \sum_j P_{i,j,*} = 1 \quad \text{(by K.K.T condition. see (17))} \end{aligned}$$

$\blacksquare$

Proposition 3.3 *(Finite Sinkhorn divergences).*
For finite Sinkhorn iterations:

$$\mathfrak{D}_{\boldsymbol{C}}^{(t)}(\boldsymbol{a},\boldsymbol{b}) := \langle \boldsymbol{\lambda}_t \,,\, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}_t \,,\, \boldsymbol{b}\rangle \tag{49}$$

The following relationship holds:

$$\mathfrak{D}_{\boldsymbol{C}}^{(t)}(\boldsymbol{a},\boldsymbol{b}) \leq L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b}) \tag{50}$$

Note that during the Sinkhorn iterations, $(\boldsymbol{\lambda}_t, \boldsymbol{\mu}_t)$ is still dual feasbile for (19). Therefore the inequality holds. Note that the primal iteration $\boldsymbol{P}_t$ is not primal feasible, since, by definition, these iterates are designed to satisfy upon convergence marginal constraints.

Unlike the regularized expression $L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a},\boldsymbol{b})$, the finite Sinkhorn divergence $\mathfrak{D}_{\boldsymbol{C}}^{(t)}(\boldsymbol{a},\boldsymbol{b})$ is **not**, in general, a **convex** function of its arguments (this can be easily checked numerically). $\mathfrak{D}_{\boldsymbol{C}}^{(t)}(\boldsymbol{a},\boldsymbol{b})$ is, however, a ***differentiable*** *function* which can be differentiated using **automatic differentiation** techniques with respect to any of its arguments, notably $\boldsymbol{C}$, $\boldsymbol{a}$, or $\boldsymbol{b}$.

Note in many literatures, we have the following definition of Sinkhorn divergence:

Definition [Li et al., 2021]
The ***Sinkhorn divergence*** is formally defined as:

$$\mathcal{L}^{\epsilon}(\alpha,\beta) := \inf_{\pi\in U(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \epsilon\, \mathbb{KL}\left(\pi \parallel \alpha\otimes\beta\right) \tag{51}$$

In [Vialard, 2019], it is defined as

$$\mathcal{L}^{\epsilon}(\alpha,\beta) + \frac{1}{2}\left(\mathcal{L}^{\epsilon}(\alpha,\gamma) + \mathcal{L}^{\epsilon}(\gamma,\beta)\right)$$

References

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735, 1989.

Qian Li, Zhichao Wang, Gang Li, Jun Pang, and Guandong Xu. Hilbert sinkhorn divergence for optimal transport. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2021.

Arkadi Nemirovski and Uriel Rothblum. On complexity of matrix scaling. *Linear Algebra and its Applications*, 302:435–460, 1999.

Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

François-Xavier Vialard. An elementary introduction to entropic regularization and proximal methods for numerical optimal transport. 2019.