Entropic regularization of Optimal Transport

Quentin Bertrand, Mathurin Massias, Titouan Vayer

Last updated: December 6, 2024

Contents

1 Entropic regularization of optimal transport **1**
1.1 Convergence of the Sinkhorn algorithm 4

2 The Sinkhorn divergence **5**
2.1 Renormalizing the quantity 6
2.2 Some properties of the Sinkhorn divergence 6

1 Entropic regularization of optimal transport

References: Peyré et al. (2019, Chap. 4).

Definition 1.1 (Entropy)**.** *The entropy of a positive vector $x \in \mathbb{R}^d_+$ is:*

$$H(x) = -\sum_{i=1}^{d} x_i \log(x_i) \tag{1.1}$$

with the convention $0 \log 0 = 0$. It extends to matrices by summing over both indices i and j: $H(P) = -\sum_{i,j} P_{ij} \log(P_{ij})$.

*The Hessian of the entropy is $\operatorname{diag}(-1/x_i)$ so the entropy is **strictly concave**. In 1D on $[0, 1]$ it looks roughly like a concave parabola, as shown on Figure 1, but with infinite slope at 0.*

The negative entropy *is simply minus the entropy. It is a negative quantity, and strictly convex.*

Exercise 1.1. *Show that the element of the simplex that minimizes the negative entropy (equivalently, that maximizes the entropy) is the constant vector $(1/n, \dots, 1/n)$.*

Definition 1.2. *The entropic-regularized optimal transport problem (EOT) is:*

$$\min_{P \geq 0} \langle C, P \rangle + \varepsilon \sum_{ij} P_{ij}(\log(P_{ij} - 1) \quad s.t. \quad P\mathbb{1} = a, \; P^\top \mathbb{1} = b \tag{1.2}$$

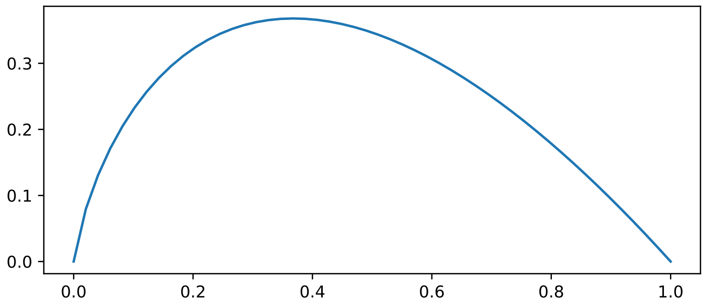

Figure 1: the entropy

Since the negative entropy is strictly convex, the solution of EOT is unique. Note that the -1 in the entropy is here for convenience and can be removed without affecting the solution, since the sum of entries of P is constant for any feasible P.

Proposition 1.3. *When ε goes to 0 the solution P^* of the entropic problem converges to a solution of the Kantorovich problem (which corresponds to $\varepsilon = 0$). If there are multiple solutions to the Kantorovich problem, it converges to the one that has maximal entropy.*

Note that this result is not true for general optimization problems: argmin and limit do not commute.

Proof. Take a sequence of positive regularization strength ε_ℓ going to 0, with associated EOT solutions P_ℓ. Since the constraint set is the same for all P_ℓ, and that it is closed and bounded, we can extract a converging subsequence (say towards P_∞) that we rename P_ℓ for convenience. Let P^* be a solution of the Kantorovich problem. Then, by optimality of P^* for the Kantorovich problem, since P_ℓ is feasible for it too,

$$\langle C, P* \rangle \leq \langle C, P_\ell \rangle \tag{1.3}$$

On the other hand, since P^* is feasible for EOT with $\varepsilon = \varepsilon_\ell$, by optimality of P_ℓ for this problem,

$$\langle C, P_\ell \rangle - \varepsilon_\ell H(P_\ell) \leq \langle C, P^* \rangle - \varepsilon_\ell H(P^*) \tag{1.4}$$

Combining the two we get:

$$0 \leq \langle C, P_\ell \rangle - \langle C, P^* \rangle \leq \varepsilon_\ell (H(P_\ell) - H(P^*)) \tag{1.5}$$

Since (P_ℓ) is bounded (the feasible set is the same for all problem, and it is bounded), so is $H(P_\ell)$; letting ℓ to infinity we must have $\langle C, P_\infty \rangle = \langle C, P^* \rangle$, which shows that P_∞ is a solution of the Kantorovich problem (as it is feasible). □

Exercise 1.2. *Show that when $\varepsilon \to \infty$, the solution of EOT goes to the feasible point that has maximal entropy. Show that this point is $ab^\top$.*

Definition 1.4. *The* Gibbs kernel *associated to the EOT problem is $K = \exp(-\frac{C}{\varepsilon})$, where the exponential acts entrywise. It is thus a matrix with strictly positive entries.*

Definition 1.5. *The Kullback-Leibler divergence between two matrices $P, Q \in \mathbb{R}_+^{n\times m}$ (so, with positive entries) is:*

$$\mathrm{KL}(P, Q) \triangleq \sum_{i=1}^{n}\sum_{j=1}^{m} P_{ij} \log \frac{P_{ij}}{Q_{ij}} - Q_{ij} + P_{ij} \tag{1.6}$$

Proposition 1.6. *The KL divergence is the Bregman divergence associated to the negative entropy. It is thus strictly convex in its first variable.*

Proposition 1.7 (EOT as Bregman projection)**.** *The solution of EOT is also the solution of:*

$$\operatorname*{argmin}_P \mathrm{KL}(P, K) \quad s.t. \quad P \geq 0,\ P1 = a,\ P^\top 1 = b \tag{1.7}$$

with K the Gibbs kernel $K = \exp(-C/\varepsilon)$.

This means that solving EOT is the same as projecting the Gibbs kernel onto the feasible set, in the sense of the KL divergence (and not, as in the usual Euclidean projection, in the sense of $\frac{1}{2}\| \cdot - \cdot \|^2$).

Proposition 1.8. *The dual of EOT is the following unconstrained problem:*

$$\max_{f,g} \langle a, f\rangle + \langle b, g\rangle - \varepsilon \sum_{ij} \exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) \tag{1.8}$$

We see that as $\varepsilon \to 0$, the values for which $f_i + g_j > C_{ij}$ are more and more penalized, eventually leading to a hard constraint for $\varepsilon = 0$.

In addition, for any solution f^*, g^* to the dual, the primal solution P^* satisfies:

$$P^*_{ij} = \exp\left(\frac{f^*_i + g^*_j - C_{ij}}{\varepsilon}\right) \tag{1.9}$$

One of the strong benefits of EOT is that its dual is unconstrained, so when solving it one does not have to enforce constraints such as $f_i + g_j \leq C_{ij}$, which are nearly impossible to apply in the continuous case.

Exercise 1.3. *Show that it (f^*, g^*) is a solution to the dual of EOT, so is $(f^* + \delta 1_n, g^* - \delta 1_m)$ for any value of δ.*

Fun fact: show that if a and b have entries all smaller than 1, then any solution (f^*, g^*) of the dual problem must satisfy $f^*_i + g^*_j \leq C_{ij}$ – and this, even though the constraint is not enforced in a hard way, unlike in unregularized OT.

Definition 1.9. *Let $M \in \mathbb{R}_{++}^{n\times m}$, $a \in \mathbb{R}_{++}^n$, and $b \in \mathbb{R}_{++}^m$. The* matrix scaling problem *is to find vectors $u, v \in \mathbb{R}_{++}^n \times \mathbb{R}_{++}^m$ such that*

$$\mathrm{diag}(u) M \,\mathrm{diag}(v) 1 = a \tag{1.10}$$

$$(\mathrm{diag}(u) M \,\mathrm{diag}(v))^\top 1 = b \tag{1.11}$$

It corresponds to multiplying rows and columns of M by scalars so that the rows end up summing to a and the columns end up summing to b.

Exercise 1.4. *Show that if u, v is a solution of the matrix scaling problem, so is $(u/\lambda, \lambda v)$ for any $\lambda > 0$.*

Proposition 1.10. *The matrix scaling problem with $M = K$ the Gibbs kernel is equivalent to the dual of EOT: if (u^*, v^*) solves the matrix scaling problem, then it is equal to $(\exp(f^*/\varepsilon), \exp(g^*/\varepsilon))$ where (f^*, g^*) solves the dual of EOT (and vice versa, any solution of the dual can be mapped to a solution of the matrix scaling problem).*

How do the two types of invariance (by addition-subtraction for f, g, by multiplication/division for (u, v)) relate to this?

Definition 1.11. *The Sinkhorn algorithm starts from a strictly positive v^0 and iterates:*

$$u^{\ell+1} = a/Kv^{\ell} \tag{1.12}$$

$$v^{\ell+1} = b/K^{\top}u^{\ell+1} \tag{1.13}$$

where division is meant pointwise.

Sinkhorn's algorithm has many interpretation. The easiest one if the following:

- for a fixed v, the choice of u that makes $\operatorname{diag}(u)K\operatorname{diag}(v)$ satisfy $\operatorname{diag}(u)K\operatorname{diag}(v)1 = a$ is a/Kv (this is because $K\operatorname{diag}(v)1 = Kv$).
- for a fixed u, the choice of v that makes $\operatorname{diag}(u)K\operatorname{diag}(v)$ satisfy $(\operatorname{diag}(u)M\operatorname{diag}(v))^{\top}1 = b$ is $b/K^{\top}u$.

So, Sinkhorn alternatively modifies u and v to satisfy the first and the second constraints.

Surprisingly, this works: Sinkhorn's algorithm converges to a solution of the matrix scaling problem, and hence can be used to solve the dual of EOT.

Proposition 1.12. *If we define $P^{\ell} = \operatorname{diag}(u^{\ell})K\operatorname{diag}(v^{\ell})$, then the sequence P^{ℓ} corresponds to alternated Bregman projection (in the KL sense) onto the constraint set.*

Alternating projections usually do not converge, but when the constraint sets are affine, it does.

Proposition 1.13. *If one defines $u = \exp(f/\varepsilon)$ and $v = \exp(g/\varepsilon)$, then performing iterations of Sinkhorn is the same as performing alternate maximization (in f and in g) on the dual of EOT.*

1.1 Convergence of the Sinkhorn algorithm

To derive convergence results easily we'll work in a different metric

Definition 1.14. *The* Hilbert projective metric *on $\mathbb{R}^d_{++}$ is:*

$$d_{\mathcal{H}}(u, u) \triangleq \log \max_{i,j} \frac{u_i u'_j}{u_j u'_i} \tag{1.14}$$

Interpretation: it is a distance on the cone $\mathbb{R}^d_{++}$ quotiented by the equivalence relation $u \sim u' \Leftrightarrow u = ru', r > 0$.

This metric is useful because it "does not see" variations in scaling: $d_{\mathcal{H}}(u, \lambda u') = d_{\mathcal{H}}(u, u')$; those variations in scaling in u and v are precisely the one that makes the solutions to the matrix scaling problem not unique.

Exercise 1.5. *Show that the Hilbert projective metric satisfies the triangular inequality* $d_{\mathcal{H}}(u, u'') \leq d_{\mathcal{H}}(u, u') + d_{\mathcal{H}}(u', u'')$.

Proposition 1.15. *Let* $M \in \mathbb{R}_{++}^{n \times m}$. *Then* M *is a contraction in the Hilbert projective metric: for any positive* v, v',

$$d_{\mathcal{H}}(Mv, Mv') \leq \lambda(M) d_{\mathcal{H}}(v, v') \tag{1.15}$$

where $\lambda(K) \triangleq \frac{\sqrt{\eta(M)}-1}{\sqrt{\eta(M)}+1} < 1$ *and* $\eta(M) \triangleq \max_{i,j,k,l} \frac{M_{ik} M_{kl}}{M_{jk} M_{il}} \geq 1$ *is some sort of "conditioning" of* K.

Proposition 1.16. *Consider the iterates of Sinkhorn algorithm* u^ℓ, v^ℓ, *and define* $P^\ell = \operatorname{diag}(u^\ell) K \operatorname{diag}(v^\ell)$. *Then*

1. $u^\ell, v^\ell \to u^*, v^*$ *a solution of the matrix scaling problem associated to* K, *and so* P^ℓ *converges to the solution of EOT. The convergence is at a speed* $d_{\mathcal{H}}(u^\ell, u^*) = \mathcal{O}(\lambda(K)^{2\ell})$, *same for* v.
2. $d_{\mathcal{H}}(u^\ell, u^*) \leq \frac{d_{\mathcal{H}}(P^\ell 1, a)}{1-\lambda(K)^2}$ *and similarly for* v *and* b.
3. $\|P^\ell - P^*\|_\infty \leq d_{\mathcal{H}}(u^\ell, u^*) + d_{\mathcal{H}}(v^\ell, v^*)$.

2 The Sinkhorn divergence

When using also as a loss between probability distributions, the entropic regularization is also very useful.

Consider two (discrete) probability distributions $\alpha = \sum_{i=1}^n a_i \delta_{x_i}, \beta = \sum_{j=1}^m b_j \delta_{y_j}$ and a cost matrix $C = (c(x_i, y_j))_{ij}$. We can have a look at the minimum value

$$\min_{\substack{P \geq 0 \\ P1 = a \\ P^\top 1 = b}} \langle C, P \rangle - \varepsilon H(P) \,, \tag{2.1}$$

where $H(P) = -\sum_{ij} P_{ij}(\log(P_{ij}) - 1)$. When $\varepsilon = 0$ and when $c(x, y) = \|x - y\|$ this gives the Wasserstein distance between α and β, noted $W(\alpha, \beta)$. $W(\cdot, \cdot)$ is in fact a true metric between probability distributions: it satisfies the triangular inequality, is non-negative and vanishes if and only if $\alpha = \beta$. However when $\varepsilon > 0$ the behavior is quite different: due to the fact that the entropy is always > 0 we have that (2.1) > 0 even when $\alpha = \beta$! This leads to bad behavior when one attempts to use this quantity as a loss for example.

2.1 Renormalizing the quantity

In Feydy et al. (2019) the authors define the so-called *Sinkhorn divergence* which is a well-behaved object when one wants to use entropic regularization to compare probability distributions. First we need to see that the entropic regularized problem described in (2.1) is in fact equivalent to[1]

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \overset{def}{=} \min_{\substack{P \geq 0 \\ P\mathbb{1}=a \\ P^\top \mathbb{1}=b}} \langle C, P\rangle + \varepsilon \, \mathrm{KL}(P, ab^\top)\,. \tag{2.2}$$

This formulation will have its importance when looking at the behavior of the minimum at $+\infty$. As an exercise you can check that (2.2) admits the dual formulation

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \max_{f,g} \langle a, f\rangle + \langle b, g\rangle - \varepsilon \sum_{ij} \left(\exp(\frac{f_i + g_j - C_{ij}}{\varepsilon}) - 1 \right) a_i b_j\,. \tag{2.3}$$

Then the definition follows.

Definition 2.1 (Sinkhorn divergence)**.** *Let α, β be two probability distributions. The Sinkhorn divergence between α, β is the quantity*

$$\mathrm{S}_\varepsilon(\alpha, \beta) = \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}(\mathrm{OT}_\varepsilon(\alpha, \alpha) + \mathrm{OT}_\varepsilon(\beta, \beta))\,. \tag{2.4}$$

This simple renormalization first implies that $\mathrm{S}_\varepsilon(\alpha, \alpha) = 0$, but it is now not so clear that this quantity remains non-negative. For this we need to define the notion of kernel, which is an object used to measure similarity between points.

Definition 2.2 (Kernel)**.** *A function $\kappa : X \times X \to \mathbb{R}$ is a kernel if for any $n \in \mathbb{N}$ and $(x_1, \cdots, x_n) \in X^n$ and $(c_1, \cdots, c_n) \in \mathbb{R}^n$ we have that $\sum_{ij} \kappa(x_i, x_j) c_i c_j \geq 0$. In other words, if for any n and $(x_1, \cdots, x_n) \in X^n$ the matrix $K = (\kappa(x_i, x_j))_{ij}$ is positive semi-definite.*

A simple example is the Gaussian kernel $\kappa(x, y) = \exp(-\|x - y\|_2^2/\varepsilon)$ which is indeed a kernel.

2.2 Some properties of the Sinkhorn divergence

The following results proved in Feydy et al. (2019) shows that the Sinkhorn divergence is a "good object" for defining similarities with a little assumption on the cost.

Proposition 2.3. *Suppose that the cost c is symmetric, i.e., $c(x, y) = c(y, x)$, and that for any $\varepsilon > 0, \kappa(x, y) = \exp(-c(x, y)/\varepsilon)$ defines a kernel. Then $\mathrm{S}_\varepsilon(\cdot, \cdot)$ is symmetric, non-negative and satisfies $\mathrm{S}_\varepsilon(\alpha, \alpha) = 0$ for any probability distribution α. In other words, S_ε is a* divergence[2] *between probability distributions.*

[1] Indeed $\mathrm{KL}(P, ab^\top) = \sum_{ij} P_{ij} \log(P_{ij}/a_i b_j) - a_i b_j + P_{ij} = \sum_{ij} P_{ij} \log(P_{ij}) - \sum_{ij} P_{ij} \log(a_i) - \sum_{ij} P_{ij} \log(b_j) = -H(P) - \sum_i a_i \log(a_i) - \sum_j b_j \log(b_j) = -H(P) + H(a) + H(b)$. So the optimization problem only changes by constant terms, so both are equivalent.

[2] A divergence D is a function of two variables that satisfies $\forall x, D(x, x) = 0, \forall (x, y), D(x, y) \geq 0$.

This hypothesis on c is quite natural for many cost functions, for example when $c(x,y) = \|x-y\|_2^2$ since $\kappa(x,y) = \exp(-\|x-y\|_2^2/\varepsilon)$ is a Gaussian kernel. We will prove this proposition by using the following result.

Proposition 2.4. *Let $(f^\star, g^\star)$ be optimal dual variables for (2.3), then $\mathrm{OT}_\varepsilon(\alpha,\beta) = \langle a, f^\star\rangle + \langle b, g^\star\rangle$. Moreover, when $\alpha = \beta$ and when the cost c is symmetric, i.e. $c(x,y) = c(y,x)$, there is a pair $(f^\star, g^\star)$ of optimal dual variables of problem (2.3) with $f^\star = g^\star$.*

Proof. We note $S(f,g) = \langle a, f\rangle + \langle g, b\rangle - \varepsilon \sum_{ij} \left(\exp(\frac{f_i+g_j-C_{ij}}{\varepsilon}) - 1\right) a_i b_j$ the dual loss. Since the problem is concave we have that $\nabla_f S(f^\star, g^\star) = 0, \nabla_g S(f^\star, g^\star) = 0$. These conditions implies that (do the calculus)

$$\begin{aligned} \forall j,\ \sum_i \exp(\frac{f_i^\star + g_j^\star - C_{ij}}{\varepsilon}) a_i = 1 \\ \forall i,\ \sum_j \exp(\frac{f_i^\star + g_j^\star - C_{ij}}{\varepsilon}) b_j = 1 \end{aligned} \tag{2.5}$$

In other words by multiplying the first equation by b_j and summing all the terms we get that $\sum_{ij} \exp(\frac{f_i^\star+g_j^\star-C_{ij}}{\varepsilon}) a_i b_j = 1$. Thus $-\varepsilon \sum_{ij} \left(\exp(\frac{f_i^\star+g_j^\star-C_{ij}}{\varepsilon}) - 1\right) a_i b_j = 0$ which concludes the first point of the proof. For the second point consider the problem (2.3) with $\alpha = \beta = \sum_{i=1}^n a_i \delta_{x_i}$ and suppose the cost is symmetric. In this case

$$S(f,g) = \langle a, f\rangle + \langle a, g\rangle - \varepsilon \sum_{ij} \left(\exp(\frac{f_i + g_j - C_{ij}}{\varepsilon}) - 1\right) a_i a_j \,. \tag{2.6}$$

Since C is symmetric we can easily check that $S(f,g) = S(g,f)$. We will show that this implies $\max_{f,g} S(f,g) = \max_f S(f,f)$ thus there will be optimal dual variables $(f^\star, g^\star)$ with $f^\star = g^\star$ that solve the problem. To see this we first have $\max_{f,g} S(f,g) \geq \max_f S(f,f)$. Now since the dual problem is (jointly) concave, *i.e.* $(f,g) \to S(f,g)$ is concave, we have that

$$S(\frac{(f,g)}{2} + \frac{(g,f)}{2}) \geq \frac{1}{2}(S(f,g) + S(g,f)) = S(f,g)\,. \tag{2.7}$$

But $S(\frac{(f,g)}{2} + \frac{(g,f)}{2}) = S(\frac{f+g}{2}, \frac{f+g}{2}) \leq \max_f S(f,f)$. Thus $\max_f S(f,f) \geq \max_{f,g} S(f,g)$ which concludes. □

We can use this result to prove Proposition 2.3.

Proof. We note $f^\alpha \in \mathbb{R}^n$ an optimal dual variable for $\mathrm{OT}_\varepsilon(\alpha,\alpha)$ and $g^\beta \in \mathbb{R}^m$ for $\mathrm{OT}_\varepsilon(\beta,\beta)$ (which exist since we can consider optimal dual variables that are equal from Proposition 2.4). Now the couple (f^α, g^β) is admissible for $\mathrm{OT}_\varepsilon(\alpha,\beta)$. Thus

$$\mathrm{OT}_\varepsilon(\alpha,\beta) \geq \langle a, f^\alpha\rangle + \langle g^\beta, b\rangle - \varepsilon \sum_{ij} \left(\exp(\frac{f_i^\alpha + g_j^\beta - c(x_i,y_j)}{\varepsilon}) - 1\right) a_i b_j \,. \tag{2.8}$$

Using Proposition 2.4 we also have $\mathrm{OT}_\varepsilon(\alpha,\alpha) = \langle f^\alpha, a\rangle + \langle f^\alpha, a\rangle = 2\langle f^\alpha, a\rangle$ (same for β). Hence

$$\begin{aligned} \mathrm{S}_\varepsilon(\alpha,\beta) &= \mathrm{OT}_\varepsilon(\alpha,\beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta,\beta) \\ &= \mathrm{OT}_\varepsilon(\alpha,\beta) - \langle f^\alpha, a\rangle - \langle g^\beta, b\rangle \\ &\geq -\varepsilon \sum_{ij} \left(\exp(\frac{f_i^\alpha + g_j^\beta - c(x_i,y_j)}{\varepsilon}) - 1 \right) a_i b_j \\ &= \varepsilon(1 - \overline{u}^\top K_{xy}\overline{v}) \,, \end{aligned} \tag{2.9}$$

where we defined $\overline{u}_i = a_i \exp(f_i^\alpha/\varepsilon), \overline{v}_j = \exp(g_j^\beta/\varepsilon)$ and $K_{xy} = (\exp(-c(x_i,y_j)/\varepsilon))_{ij}$. We note also $K_{yx} = (\exp(-c(y_j,x_i)/\varepsilon))_{ji}$. Now using Proposition 2.4 we have also, for the problems related to $\mathrm{OT}_\varepsilon(\alpha,\alpha)$ and $\mathrm{OT}_\varepsilon(\beta,\beta)$,

$$\begin{aligned} &\sum_{ii'} \exp(\frac{f_i^\alpha + f_{i'}^\alpha - c(x_i,x_{i'})}{\varepsilon}) a_i a_{i'} = 1 \implies \overline{u}^\top K_{xx}\overline{u} = 1 \\ &\sum_{jj'} \exp(\frac{g_j^\beta + g_{j'}^\beta - c(y_j,y_{j'})}{\varepsilon}) b_j b_{j'} = 1 \implies \overline{v}^\top K_{yy}\overline{v} = 1 \,, \end{aligned} \tag{2.10}$$

where $K_{xx} = (\exp(-c(x_i,x_{i'})/\varepsilon))_{ii'}\,, K_{yy} = (\exp(-c(y_j,y_{j'})/\varepsilon))_{jj'}$. Thus

$$\begin{aligned} S_\varepsilon(\alpha,\beta) &\geq \varepsilon(1 - \overline{u}^\top K_{xy}\overline{v}) = \frac{\varepsilon}{2}(1 + 1 - 2\overline{u}^\top K_{xy}\overline{v}) \\ &= \frac{\varepsilon}{2}(\overline{u}^\top K_{xx}\overline{u} + \overline{v}^\top K_{yy}\overline{v} - 2\overline{u}^\top K_{xy}\overline{v}) \\ &= \begin{pmatrix}\overline{u} & \overline{v}\end{pmatrix}^\top \begin{pmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy}\end{pmatrix} \begin{pmatrix}\overline{u} \\ \overline{v}\end{pmatrix} \,. \end{aligned} \tag{2.11}$$

To conclude it suffices to use that $\kappa(x,y) = \exp(-c(x,y)/\varepsilon)$ so that $\overline{K} = \begin{pmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy}\end{pmatrix} = (\kappa(z_i,z_j))_{ij}$ where $(z_1,\cdots,z_n,z_{n+1},\cdots z_{n+m}) = (x_1,\cdots,x_n,y_1,\cdots,y_m)$. Since κ is a kernel the matrix $\overline{K}$ is a PSD matrix and thus $\begin{pmatrix}\overline{u} & \overline{v}\end{pmatrix}^\top \begin{pmatrix} K_{xx} & K_{xy} \\ K_{yx} & K_{yy}\end{pmatrix} \begin{pmatrix}\overline{u} \\ \overline{v}\end{pmatrix} \geq 0$. □

Last, but not least, the Sinkhorn divergence has very interesting asymptotic properties as detailed below.

Proposition 2.5. *Let c be a symmetric cost. The Sinkhorn divergence interpolates between the unregularized OT distance and the so-called* Maximum Mean Discrepancy *(MMD)*

$$\begin{aligned} &\mathrm{S}_\varepsilon(\alpha,\beta) \underset{\varepsilon\to 0}{\to} \min_{\substack{P\geq 0 \\ P1=a,\, P^\top 1=b}} \langle C, P\rangle \\ &\mathrm{S}_\varepsilon(\alpha,\beta) \underset{\varepsilon\to+\infty}{\to} \frac{1}{2}\left(-\sum_{ii'} c(x_i,x_{i'})a_i a_{i'} - \sum_{jj'} c(y_j,y_{j'})b_j b_{j'} + 2\sum_{ij} c(x_i,y_j)a_i b_j \right) \end{aligned} \tag{2.12}$$

Under the hypothesis on c of Proposition 2.3 the complicated term when $\varepsilon \to +\infty$ is non-negative and defines also a very useful distance between probability distributions.

We will prove only the convergence result, we leave the clean definition of the MMD (and the fact that it is non-negative) for another time...

Proof. Consider $(f^\varepsilon, g^\varepsilon)$ optimal dual variables for the problem $\mathrm{OT}_\varepsilon(\alpha, \beta)$, that is

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \langle a, f^\varepsilon \rangle + \langle b, g^\varepsilon \rangle - \varepsilon \sum_{ij} \left(\exp(\frac{f_i^\varepsilon + g_j^\varepsilon - c(x_i, y_j)}{\varepsilon}) - 1 \right) a_i b_j \,. \tag{2.13}$$

As $\varepsilon \to +\infty$ we have $\exp(\frac{f_i^\varepsilon + g_j^\varepsilon - c(x_i,y_j)}{\varepsilon}) = 1 + \frac{f_i^\varepsilon + g_j^\varepsilon - c(x_i,y_j)}{\varepsilon} + O(\frac{1}{\varepsilon^2})$ (everything is discrete so the dual variables are uniformly bounded, independently of ε). Thus

$$\begin{aligned} \mathrm{OT}_\varepsilon(\alpha,\beta) &= \langle a, f^\varepsilon \rangle + \langle b, g^\varepsilon \rangle - \varepsilon \sum_{ij} \left(\frac{f_i^\varepsilon + g_j^\varepsilon - c(x_i, y_j)}{\varepsilon} + O(\frac{1}{\varepsilon^2}) \right) a_i b_j \\ &= \sum_{ij} c(x_i, y_j) a_i b_j + O(\frac{1}{\varepsilon}). \end{aligned} \tag{2.14}$$

Thus $\mathrm{OT}_\varepsilon(\alpha,\beta) \to_{\varepsilon \to +\infty} \sum_{ij} c(x_i,y_j) a_i b_j$. In the same vein we have $\mathrm{OT}_\varepsilon(\alpha,\alpha) \to_{\varepsilon \to +\infty} \sum_{ii'} c(x_i, x_{i'}) a_i a_{i'}$ and $\mathrm{OT}_\varepsilon(\beta,\beta) \to_{\varepsilon \to +\infty} \sum_{jj'} c(y_j, y_{j'}) b_j b_{j'}$ hence the result. The setting $\varepsilon \to 0$ is clear. □

References

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690. PMLR, 2019.