A desirable property when talking about filter stability

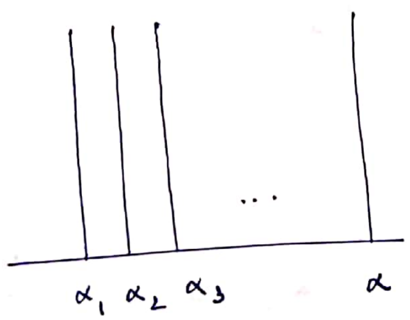
If $\tilde{d}$ is our distance or proxy for distance between measures on $\mathbb{R}^d$, we would like $\tilde{d}$ to have the following property

if $\alpha_n \xrightarrow{\text{weakly}} \alpha$ i.e. $\int f \, d\alpha_n \to \int f \, d\alpha \quad \forall$ test f

then $\tilde{d}(\alpha_n, \alpha) \to 0$ and vice versa.

KL divergence does not satisfy the if condition in the above iff statement.

<u>Counterexample</u> :



α_n, α are measures on $\mathbb{R}^2$ with supports on vertical line segments of equal length.

$\alpha_n \to \alpha$ but $KL(\alpha_n | \alpha) = +\infty \quad \forall n$

Entropy Regularized Optimal Transport

Let α, β be measures on $\mathcal{X}$,

$$OT_\varepsilon(\alpha, \beta) = \min_{\pi_1 = \alpha, \pi_2 = \beta} \left[\int c(x,y) \, d\pi(x,y) + \varepsilon \, KL(\pi | \alpha \otimes \beta) \right]$$

[IOM, eq 1]

The dual problem for OT_ε

Let $f \oplus g\,(x,y) = f(x) + g(y)$.

$$OT_\varepsilon(\alpha,\beta) = \max_{f,g \in C(X)} \left[\int_X f\,d\alpha + \int_X g\,d\beta - \varepsilon \int_{X^2} \left(e^{\frac{f\oplus g - c}{\varepsilon}} - 1 \right) d(\alpha\otimes\beta) \right]$$

[IOM, eq 8]

Lenoid Kantorovich won the Nobel prize in economics in 1975 for proving the equivalence of these primal and dual problems. [GDA, pg 98]

Does OT_ε satisfy our desired property?

No. In general $OT_\varepsilon(\alpha,\alpha) > 0$ so it can't satisfy the property. [IOM, pg 3]

Sinkhorn Divergence

$$S_\varepsilon(\alpha,\beta) := OT_\varepsilon(\alpha,\beta) - \frac{1}{2} OT_\varepsilon(\alpha,\alpha) - \frac{1}{2} OT_\varepsilon(\beta,\beta)$$

[IOM, eq 3]

S_ε satisfies $S_\varepsilon(\alpha,\alpha) = 0$. Not only that, it satisfies something stronger.

If $\mathcal{X}$ is compact and the cost function $c(x,y)$ is Lipschitz (i.e. Lipschitz in both variables separately) then S_ε is symmetric, positive-definite, convex in each input and for all Radon measures $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ (positive unit mass measures on $\mathcal{X}$),

$$0 = S_\varepsilon(\beta,\beta) \leq S_\varepsilon(\alpha,\beta)$$

$$\alpha = \beta \iff S_\varepsilon(\alpha,\beta) = 0$$

$$\alpha_n \xrightarrow{\text{weakly}} \alpha \iff S_\varepsilon(\alpha_n,\alpha) \to 0$$

These results hold for measures with bounded support on $\mathbb{R}^d$ and $c(x,y) = \|x-y\|_2$ or $\|x-y\|_2^2$ satisfy the Lipschitz property. [IOM, Thm 1]

This is good enough for us since our filtering distributions are supported on bounded attractors. From now on we'll assume c is always Lipschitz.

Is S_ε a distance?

No. The gluing lemma of Villani that's crucial in proving the triangle inequality of $OT(\alpha,\beta) = OT_0(\alpha,\beta)$ does not seem to have a counterpart for S_ε for $\varepsilon>0$. [TOT, lemma 7.6]

Does S_ε converge to Wasserstein distance ?

Yes. $\lim_{\varepsilon \to 0} S_\varepsilon(\alpha, \beta) = OT_0(\alpha,\beta) = W(\alpha,\beta)^p$ (for the right p)

Here of course we have $c(x,y) = \|x-y\|_p^p$.

[IOM, eq 4
LGM, Thm 1]

So S_ε is indeed a good proxy for Wasserstein distance or EMD for small ε.

Cuturi's formulation is not useful for us

Cuturi's paper proves that when α, β are discrete measures with identical support or α, β are measures on a discrete, finite space X then they can be just characterized by their weights say r, c respectively and $\mathbb{1}_{r \neq c} OT_\varepsilon$ then turns out to be a distance on the space of weights (and therefore measures on X).

[LCOT, Thm 1]

This view is not useful for our purposes.
But once again triangle inequality for $\mathbb{1}_{r \neq c} OT_\varepsilon$ is proven with the help of Villani's gluing lemma.

[LCOT, lemma 1]

Why is the dual formulation of OT_ε useful?

$$OT_\varepsilon(\alpha,\beta) = \max_{f,g \in C(\mathcal{X})} \left[\int_{\mathcal{X}} f \, d\alpha + \int_{\mathcal{X}} g \, d\beta - \varepsilon \int_{\mathcal{X}^2} \left(e^{\frac{f \oplus g - c}{\varepsilon}} - 1 \right) d(\alpha \otimes \beta) \right]$$

Turns out the dual problem is concave in each variable i.e. f and g and therefore we can solve it by iteratively optimizing f and g. [EROT, section 4.2]

The Sinkhorn Algorithm

For discrete measures $\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$ and $\beta = \sum_{i=1}^{m} \beta_i \delta_{y_i}$ where $x_i, y_i \in \mathbb{R}^d$, we just need to keep track of the vectors $[f(x_1), \dots, f(x_n)]$ and $[g(y_1), \dots, g(y_m)]$.

initialize : $f_i = f(x_i) = 0 \quad \forall \; i = 1, \dots, n \; ; \; g_j = g(y_j) = 0 \quad \forall \; j = 1, \dots, m$

iterate until convergence (or for a certain number of times):

$$f_i = -\varepsilon \log \sum_{k=1}^{m} \exp\left(\log \beta_k + \frac{1}{\varepsilon} g_k - \frac{1}{\varepsilon} c(x_i, y_k) \right) \quad \forall \; i = 1, \dots, n$$

$$g_j = -\varepsilon \log \sum_{k=1}^{n} \exp\left(\log \alpha_k + \frac{1}{\varepsilon} f_k - \frac{1}{\varepsilon} c(x_k, y_j) \right) \quad \forall \; j = 1, \dots, m$$

final output : $OT_\varepsilon(\alpha, \beta) = \sum_{i=1}^{n} \alpha_i f_i + \sum_{j=1}^{m} \beta_j g_j$

[IOM, section 3.1
EROT, proposition 10]

How expensive is S_ε compared to OT_ε ?

If $\alpha = \beta$ then $f = g$ and the Sinkhorn iteration for f can be written as

$$f_i = \frac{1}{2}\left[f_i - \varepsilon \log \sum_{k=1}^{n} \exp\left(\log \alpha_k + \frac{1}{\varepsilon} f_k - \frac{1}{\varepsilon} c(x_i, x_k)\right)\right]$$

Turns out this iteration converges much faster than alternate updates to f and g and 3 iterations are enough to compute f ! [IOM, section 3.1]

So computation of $OT_\varepsilon(\alpha,\alpha)$ and $OT_\varepsilon(\beta,\beta)$ only requires a few iterations and S_ε is not all expensive compared to OT_ε.

If α, β are in $\mathcal{M}_1^+(\mathcal{X})$ where $\mathcal{X}$ is compact and α_n, β_n are empirical estimates for α, β respectively, $\alpha_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$, $\beta_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$ where x_i are iid samples from α and y_i are iid samples from β and $S_\varepsilon(\alpha_n, \beta_n) \to 0$ then is it true that $S_\varepsilon(\alpha,\beta) = 0$?

Turns out if (f_n, g_n) are the optimal dual potentials for $OT_\varepsilon(\alpha_n, \beta_n)$ then they uniformly converge to (f, g), the optimal dual potentials for $OT_\varepsilon(\alpha, \beta)$ whenever $\alpha_n \xrightarrow{\text{weakly}} \alpha$ and $\beta_n \xrightarrow{\text{weakly}} \beta$ and $\mathcal{X}$ is compact, all of which are true in our question.

[IOM, proposition 13]

Consequently $OT_\varepsilon(\alpha_n, \beta_n) \to OT_\varepsilon(\alpha, \beta)$, $OT_\varepsilon(\alpha_n, \alpha_n) \to OT_\varepsilon(\alpha, \alpha)$ and $OT_\varepsilon(\beta_n, \beta_n) \to OT_\varepsilon(\beta, \beta)$

$\Rightarrow$ $S_\varepsilon(\alpha_n, \beta_n) \to S_\varepsilon(\alpha, \beta)$ $\Rightarrow$ $S_\varepsilon(\alpha, \beta) = 0 \Leftrightarrow$ $\alpha = \beta$.

Filter stability: Let $\alpha_n, \beta_n \in \mathcal{M}_1^+(\mathcal{X})$ where $\mathcal{X}$ is compact. $\hat{\alpha}_{n,m} = \frac{1}{m}\sum_{i=1}^{m} \delta_{x_{i,n}}$, $\hat{\beta}_{n,m} = \frac{1}{m}\sum_{i=1}^{m} \delta_{y_{i,n}}$ are empirical estimates for α, β.

If $\lim_{m,n\to\infty} S^\varepsilon_{m,n} = \lim_{n\to\infty}\lim_{m\to\infty} S^\varepsilon_{m,n} = \lim_{m\to\infty}\lim_{n\to\infty} S^\varepsilon_{m,n} = 0$ then is it true

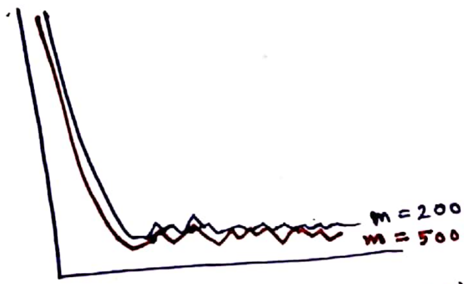
where $S^\varepsilon_{m,n} = S_\varepsilon(\hat{\alpha}_{n,m}, \hat{\beta}_{n,m})$

that $\lim_{n\to\infty} S_\varepsilon(\alpha_n, \beta_n) = 0$?

As saw in the last section S_ε is continuous w.r.t. both inputs and therefore

$$\lim_{n\to\infty} S_\varepsilon(\alpha_n, \beta_n) = \lim_{n\to\infty} \lim_{m\to\infty} S_\varepsilon(\hat{\alpha}_{n,m}, \hat{\beta}_{n,m}) = 0$$

Although we assumed both $\lim_{m\to\infty} \lim_{n\to\infty}$ and $\lim_{n\to\infty} \lim_{m\to\infty}$ exist and the limits can be swapped we only need $\lim_{n\to\infty} \lim_{m\to\infty}$ to be 0 which is true in our experiments.



S_ε plot for stability

References :

1. IOM : Interpolating between Optimal Transport and MMD using Sinkhorn Divergence (2018)

2. GDA : Geometric Data Analysis, beyond Convolutions (2020)

3. TOT : Topics in Optimal Transportation, Villani

4. LGM : Learning Generative Models with Sinkhorn Divergences (2017)

5. EROT : Entroy Regularized Optimal Transport (2019)

6. LCOT : Sinkhorn distances : Lightspeed Computation of Optimal Transportation Distances (2013)