Why we should avoid KL while discussing stability:

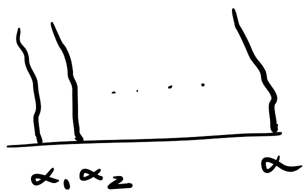
If d is our distance / pseudo distance / some sort of measure of distance, a desirable property for d would be for

$\alpha_n \to \alpha$ i.e. $\int f \, d\alpha_n \to \int f \, d\alpha \quad \forall$ test f,

$d(\alpha_n, \alpha) \to 0$

KL does not have this property.

Counter example:



α_n, α are measures on $\mathbb{R}^2$ and are uniform on vertical line segments

Entropy regularized optimal transport

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

[IOM, eq 1]

α, β are measures on $\mathcal{X}$.

The dual problem for OT_ε

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

[IOM, eq 8]

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

Lenoid Kantorovich won the Nobel prize in economics in 1975 for proving the equivalence of these dual and primal problems.

[GDA, pg-98]

Does OT_ε satisfy our desired property?

No. $OT_\varepsilon(\alpha, \alpha) > 0$ in general.

To get our desired property we can define Sinkhorn divergence.

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

S_ε is symmetric, positive-definite, smooth.

$S_\varepsilon(\alpha, \beta) = 0 \iff \alpha = \beta$ [IOM sec 2.3]

S_ε also possesses our desired property. [IOM Thm 1]

Is S_ε is a distance?

No. The gluing lemma of Villani does not hold for S_ε

and proving triangle inequality in a similar fashion to OT_0 is not possible. (TOT lemma 7.6)

Does S_ε converge to Wasserstein distance?

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

(IOM eq 4, LGM Thm 1)

Cuturi's formulation is not useful for our purposes since α, β are measures on a finite set there and their "dual" formulation is different from IOM dual.

The algorithm for computing S_ε.

Clearly written algorithm does not appear in IOM. LGM does have an algorithm but several sources cite that it might not stable (EROT pg 38-40 (stabilizing Sinkhorn)). To stabilize the algorithm you have to implement in log domain.

References :

1. IOM : Interpolating between Optimal transport and MMD

2. LGM : Learning Generative Models with Sinkhorn divergences

3. EROT : Entropy Regularized Optimal transport

4. TOT : Topics in optimal Transportion