Interpolating between Optimal Transport and MMD with Sinkhorn divergences

De-biasing the Sinkhorn loop to prevent the measures' supports from **shrinking**.

Jean Feydy[1,2] Thibault Séjourné[1] François-Xavier Vialard[3] Shun-ichi Amari[4] Alain Trouvé [2] Gabriel Peyré[1]

[1]DMA, École Normale Supérieure [2]CMLA, ENS Paris-Saclay [3]LIGM, UPEM [4]Brain Science Institute, RIKEN

1 Optimal Transport + Entropy

If α, β are Radon probability measures on a compact feature space $\mathcal{X}$ endowed with a Lipschitz cost function $\mathrm{C} : (x, y) \mapsto \mathrm{C}(x, y)$ (e.g. $\frac{1}{p}\|x-y\|^p$), **Entropy-regularized OT** (Sch32) is defined through:

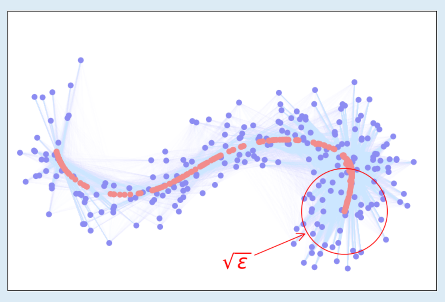
$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \text{Transport Cost} + \varepsilon \cdot \text{Entropy}$$
$$= \min_{0 \leqslant \pi \ll \alpha\otimes\beta} \langle \pi, \mathrm{C}\rangle + \varepsilon\,\mathrm{KL}(\pi, \alpha\otimes\beta) \quad \text{s.t.}\quad \pi\mathbf{1} = \alpha,\ \pi^\intercal\mathbf{1} = \beta$$
$$= \max_{f,g:\mathcal{X}\to\mathbb{R}} \langle \alpha, f\rangle + \langle \beta, g\rangle \quad \text{s.t.}\quad \max^\varepsilon_{\alpha\otimes\beta}\,[\,f \oplus g - \mathrm{C}\,] \leqslant 0$$

This approximation of the linear OT program can be solved efficiently using the iterative **IPFP-SoftAssign-Sinkhorn algorithm** (Wil69; KY94; PC17), i.e. coordinate ascent on the dual pair (f, g).

2 Removing the entropic bias

When $\varepsilon > 0$, **fuzzy transport plans** induce shrinking artifacts (CR03):

Minimize $\mathrm{OT}_\varepsilon(\alpha,\beta)$ with respect to α $\longrightarrow$



$\Longrightarrow$ Use the **unbiased** Sinkhorn divergence (RTC17; GPC18; SZRM18):

$$\mathrm{S}_\varepsilon(\alpha,\beta) = \mathrm{OT}_\varepsilon(\alpha,\beta) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\alpha,\alpha) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(\beta,\beta),$$

$$\underbrace{\mathrm{OT}(\alpha,\beta)}_{\text{Wasserstein}} \xleftarrow{\varepsilon\to 0} \underbrace{\mathrm{S}_\varepsilon(\alpha,\beta)}_{\text{Easy to compute}} \xrightarrow{\varepsilon\to+\infty} \underbrace{\mathrm{MMD}_{-\mathrm{C}}(\alpha,\beta)}_{\text{Kernel MMD}}$$

3 Our contributions

Theorem: If $e^{-\mathrm{C}(x,y)/\varepsilon}$ is a positive definite kernel,

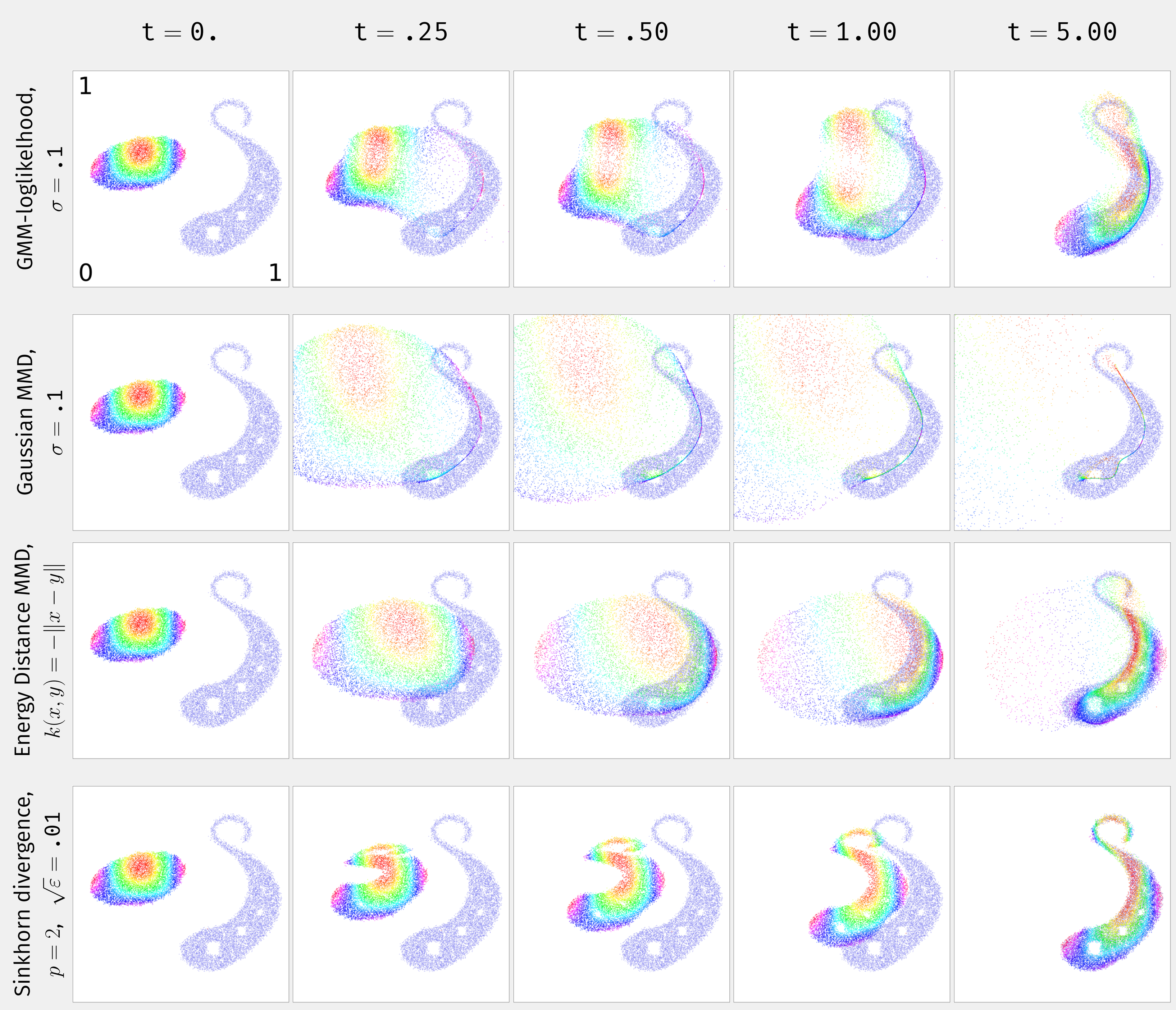
$$\mathrm{S}_\varepsilon(\beta,\beta) = 0 \leqslant \mathrm{S}_\varepsilon(\alpha,\beta)$$
$$\mathrm{S}_\varepsilon(\alpha,\beta) = 0 \iff \alpha = \beta$$
$$\mathrm{S}_\varepsilon(\alpha_n,\beta) \to 0 \iff \alpha_n \rightharpoonup \beta$$
$$\mathrm{Loss}_\beta : \alpha \mapsto \mathrm{S}_\varepsilon(\alpha,\beta) \ \text{ is \textbf{convex}.}$$

In practice: Our PyTorch+KeOps implementation has a **linear memory footprint** and outperforms the standard Sinkhorn loop by **two orders of magnitude**. It is freely available on pip and at

`www.kernel-operations.io/geomloss`

4 Geometric Loss functions for measure-fitting applications: GMM-loglikelihood vs. Kernel MMDs vs. Sinkhorn divergences

Wasserstein gradient flow: $\alpha = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} \delta_{x_i}$, $\beta = \frac{1}{\mathrm{M}}\sum_{j=1}^{\mathrm{M}} \delta_{y_j}$, minimize $\mathrm{Loss}(\alpha,\beta)$ through gradient descent on the x_i's. **Toy model for generative networks** and shape registration, without regularizing prior.



GMM-loglikelihoods ≈ Chamfer distance ≈ Soft-Hausdorff:

If k is a Gaussian kernel of deviation σ, $\mathrm{ML\text{-}Loss}(\alpha,\beta) = 2\sigma^2\langle\, \alpha-\beta\,,\ \log(k\star\alpha) - \log(k\star\beta)\,\rangle$
and generalizes the **Chamfer distance** $\langle\, \alpha-\beta\,,\ \mathrm{dist}(\,\cdot\,,\mathrm{supp}(\beta)) - \mathrm{dist}(\,\cdot\,,\mathrm{supp}(\alpha))\,\rangle$
with a SoftMin estimation of the **distances to the measures' supports**:

$$\mathrm{dist}^2(x, \mathrm{supp}(\beta)) \simeq -2\sigma^2 \log \int_y \exp(-\|x-y\|^2/2\sigma^2)\,\mathrm{d}\beta(y)$$

Kernel MMDs ≈ generalized Sobolev norms ≈ Electrostatic energies:

If k is a positive, translation-invariant kernel:

$$2\,\mathrm{MMD}_k(\alpha,\beta) = \sup_f \langle\, \alpha-\beta\,,\, f\,\rangle \quad \text{s.t.} \quad \|f\|_k^2 = \int_{\omega\in\mathbb{R}^d} \tfrac{1}{\widehat{k}(\omega)}|\widehat{f}(\omega)|^2\,\mathrm{d}\omega \leqslant 1$$
$$= \sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{N}} \alpha_i\alpha_j\,k(x_i,x_j) - 2\sum_{i=1}^{\mathrm{N}}\sum_{j=1}^{\mathrm{M}} \alpha_i\beta_j\,k(x_i,y_j) + \sum_{i=1}^{\mathrm{M}}\sum_{j=1}^{\mathrm{M}} \beta_i\beta_j\,k(y_i,y_j)$$
$$= \text{Generalization of the \textbf{Electrostatic Energy}}(+\alpha, -\beta) \text{ to potentials } k(x,y) \neq \frac{1}{\|x-y\|}.$$

$\Longrightarrow$ **Screening artifacts**, as in Coulombian physics: **dampening** of the attractive force generated by the y_j's through the set α of positive charges.

Optimal Transport ≈ Linear Assignment ≈ SoftAssign:

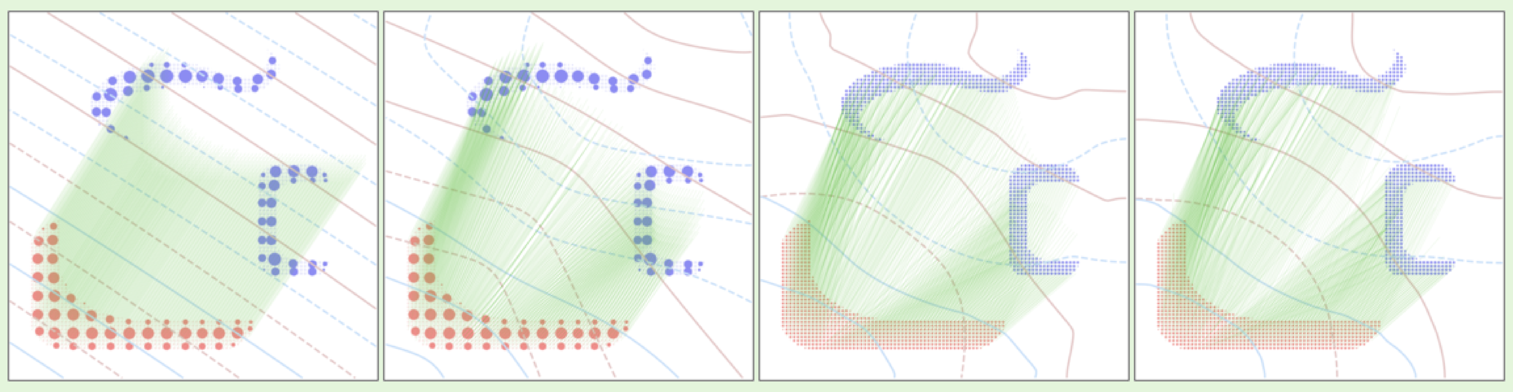
Sinkhorn divergences are positive and definite generalizations of the Earth-Mover's distance:

$$\mathrm{Wasserstein}_1(\alpha,\beta) = \sup_f \langle\, \alpha-\beta\,,\, f\,\rangle \quad \text{s.t.} \quad f \ \text{ is 1-Lipschitz.}$$

They perfectly retrieve **global translations and small deformations** in the feature space.

5 The multiscale Sinkhorn algorithm

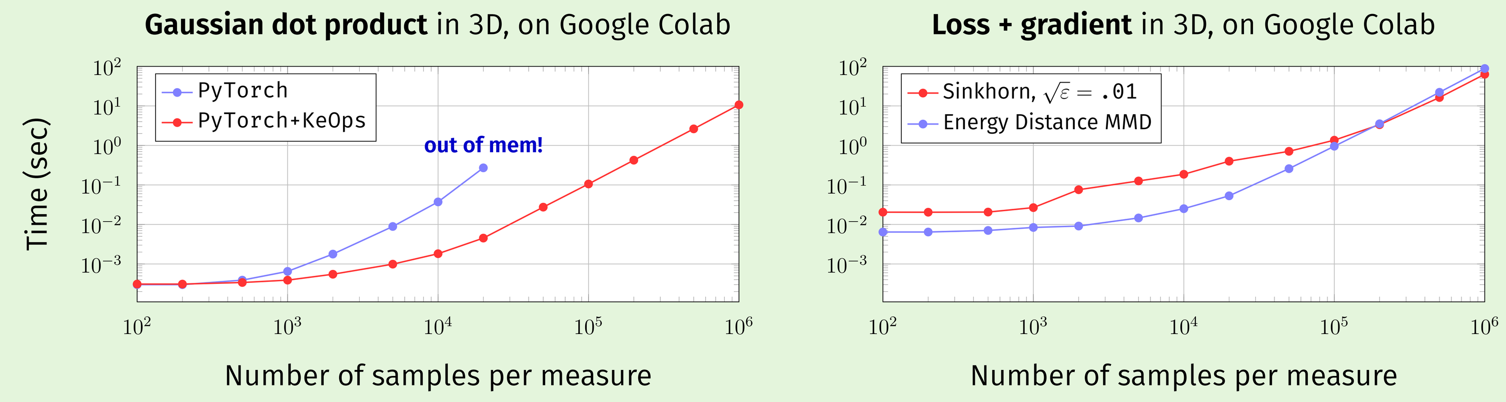
Baseline IPFP-Sinkhorn loop $\xrightarrow[\text{speedup}]{\text{x10}}$ ε-scaling heuristic (KY94) $\xrightarrow[\text{speedup}]{\text{x10}}$ Coarse-to-fine decomposition + Kernel truncation (Sch16)



it = 1, $\sqrt{\varepsilon} = 1.00$ it = 4, $\sqrt{\varepsilon} = .125$ it = 6, $\sqrt{\varepsilon} = .031$ it = 8, $\sqrt{\varepsilon} = .01$

6 Scaling up to millions of samples on the GPU

KeOps library: Kernel Operations on the GPU, with autodiff, **without memory overflows**.
Provides efficient, **online** map-reduce CUDA routines through a simple PyTorch interface:



`pip install pykeops` `pip install geomloss`

References

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. arXiv preprint arXiv:1701.07875, 2017.
[BBR06] F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. Statistics & probability letters, 76(12):1298–1302, 2006.
[BPC16] N. Bonneel, G. Peyré, and M. Cuturi. Wasserstein barycentric coordinates: Histogram regression using optimal transport. ACM Transactions on Graphics, 35(4), 2016.
[Bre67] L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR computational mathematics and mathematical physics, 7(3):200–217, 1967.
[CFG18] B. Charlier, J. Feydy, and J. Glaunès. Kernel operations on the gpu, with autodiff, without memory overflows. http://www.kernel-operations.io, 2018. Accessed: 2019-04-01.
[CR03] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. Computer Vision and Image Understanding, 89(2-3):114–141, 2003.
[Cut13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Adv. in Neural Information Processing Systems, pages 2292–2300, 2013.
[DRG15] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, pages 258–267, 2015.
[FL89] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. Linear Algebra and its applications, 114:717–735, 1989.
[FZM+15] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In Advances in Neural Information Processing Systems, pages 2053–2061, 2015.
[GBR+07] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In Advances in neural information processing systems, pages 513–520, 2007.
[GPAM+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.
[GPC18] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with sinkhorn divergences. In International Conference on Artificial Intelligence and Statistics, pages 1608–1617, 2018.
[GS10] A. Galichon and B. Salanié. Matching with trade-offs: Revealed preferences over competing characteristics. Preprint hal-00473173, 2010.
[GTY04] J. Glaunes, A. Trouvé, and L. Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, volume 2, pages II–II. Ieee, 2004.
[Kan42] L. Kantorovich. On the transfer of masses (in Russian). Doklady Akademii Nauk, 37(2):227–229, 1942.
[KCC17] I. Kaltenmark, B. Charlier, and N. Charon. A general framework for curve and surface comparison and registration with oriented varifolds. In Computer Vision and Pattern Recognition (CVPR), 2017.
[KY94] J. Kosowsky and A. L. Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. Neural networks, 7(3):477–490, 1994.
[Léo13] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. arXiv preprint arXiv:1308.0215, 2013.
[LSZ15] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In Proceedings of the 32nd International Conference on Machine Learning (ICML-15), pages 1718–1727, 2015.
[MMC16] G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of restricted boltzmann machines. In Advances in Neural Information Processing Systems, pages 3718–3726, 2016.
[MXZ06] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. Journal of Machine Learning Research, 7(Dec):2651–2667, 2006.
[PC17] G. Peyré and M. Cuturi. Computational optimal transport. arXiv:1610.06519, 2017.
[PGC+17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.
[RTC17] A. Ramdas, N. G. Trillos, and M. Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. Entropy, 19(2), 2017.
[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 40(2):99–121, November 2000.
[San15] F. Santambrogio. Optimal Transport for applied mathematicians, volume 87 of Progress in Nonlinear Differential Equations and their applications. Springer, 2015.
[SBRL18] M. Sanjabi, J. Ba, M. Razaviyayn, and J. D. Lee. On the convergence and robustness of training GANs with regularized optimal transport. arXiv preprint arXiv:1802.08249, 2018.
[Sch32] E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In Annales de l'institut Henri Poincaré, volume 2, pages 269–310, 1932.
[Sch16] B. Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. arXiv preprint arXiv:1610.06519, 2016.
[SR04] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. InterStat, 5(16.10), 2004.
[SZRM18] T. Salimans, H. Zhang, A. Radford, and D. Metaxas. Improving GANs using optimal transport. In International Conference on Learning Representations, 2018.
[TCDP17] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis. Overrelaxed sinkhorn-knopp algorithm for regularized optimal transport. arXiv preprint arXiv:1711.01851, 2017.
[VG05] M. Vaillant and J. Glaunès. Surface matching via currents. In Biennial International Conference on Information Processing in Medical Imaging, pages 381–392. Springer, 2005.
[Wil69] A. G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. Journal of Transport Economics and Policy, pages 108–126, 1969.

Contact Information

Jean Feydy,
PhD student with Prof. Alain Trouvé,
DMA, ENS Paris + CMLA, ENS Cachan.

Mail : jean.feydy@ens.fr
Web : www.math.ens.fr/~feydy