Intership proposal: convergence of the Wasserstein gradient flow of the Sinkhorn divergence.

Contact: theo.lacombe@univ-eiffel.fr

In short: This internship project is about studying the Wasserstein gradient flow of the Sinkhorn divergence to a target measure—a regularized version of the Wasserstein distance. The main question, raised in [1], is to study the limit of this flow, which is expected to be the target measure itself in reasonable settings.

This internship is expected to lead to a Ph.D. on related topics at the crossroad of entropic optimal transport, geometry and numerical optimization. The Ph.D. will be supervised by François-Xavier Vialard and T. Lacombe at the *Laboratoire d'Informatique Gaspard Monge* at *Université Gustave Eiffel* and funded via the ANR project ThéATRE[1].

1 Introduction: Optimal transport and Wasserstein Gradient flows.

The *optimal transportation* problem that consists of minimizing the cost of turning a source distribution μ into a target distribution ν belonging to the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures with finite second moment. Formally,

$$\mathrm{OT}(\mu,\nu) \coloneqq \inf_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\iint |x-y|^2 \mathrm{d}\pi(x,y), \tag{1}$$

where $\Pi(\mu,\nu)$ denotes measures on $\mathbb{R}^d\times\mathbb{R}^d$ with first and second marginals being μ and ν, respectively. This quantity is referred to as the (squared) *Wasserstein distance* between μ and ν. Minimizers in (1) are referred to as *optimal transport plans* between μ and ν. Should they be of the form $\pi = (\mathrm{id}, T)_*\mu$ for some (measurable) map $T:\mathbb{R}^d\to\mathbb{R}^d$, T is said to be a *Monge map* between μ and ν induced by π. Here, $\varphi_*\mu$ denotes the pushforward of a measure $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ by a (measurable) map $\varphi:\mathbb{R}^d\to\mathbb{R}^d$, i.e. $\varphi_*\mu(A) = \mu(\varphi^{-1}(A))$ for all Borel $A\subset\mathbb{R}^d$.

About two decades ago, it was observed that several important differential equations could be phrased as minimization problems of functionals $\mu\mapsto\mathcal{F}(\mu)$ with respect to *the geometry induced by the optimal transportation problem* (1) [7, 6, 8]. To do so, one defines a sequence starting from μ_0 by

$$\mu_{k+1} = \arg\min_{\mu}\left\{\mathcal{F}(\mu) + \frac{1}{2\tau}\mathrm{OT}(\mu,\mu_k)\right\} \tag{2}$$

for some step size $\tau>0$ and obtains under suitable assumptions an absolutely continuous curve $t\mapsto\mu_t$ in the limit $\tau\to 0$, called the *Wasserstein gradient flow* (WGF) of $\mathcal{F}$.

In particular, when $\mathcal{F}(\mu) = \mathrm{OT}(\mu,\mu_{\text{target}})$ for some μ_{target} and the sequence is initialized at $\mu_0 = \mu_{\text{source}}$, one retrieves the so-called *McCann interpolation* between μ_{source} and μ_{target}.

Entropy-regularized Optimal Transport. Cuturi popularized in [2] a variant of (1) by introducing an *entropic regularization* term, yielding the Entropic Optimal Transport (EOT) formulation that can

[1] Théorie et Application du Transport avec Régularisation Entropique

be expressed through two optimization problems dual of each other:

$$\mathrm{OT}_\varepsilon(\mu,\nu) \coloneqq \min_{\pi\in\Pi(\mu,\nu)} \frac{1}{2}\iint |x-y|^2 \mathrm{d}\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\mu\otimes\nu) \tag{3}$$

$$= \sup_{f,g\in\mathcal{C}(\mathbb{R}^d)} \int f(x)\mathrm{d}\mu(x) + \int g(y)\mathrm{d}\nu(y) - \varepsilon\iint \left(e^{\frac{f(x)+g(y)-\frac{1}{2}|x-y|^2}{\varepsilon}} - 1\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y) \tag{4}$$

where $\varepsilon > 0$ is the regularization strength, and $\mathrm{KL}(\alpha|\beta) \coloneqq \int \log\left(\frac{\mathrm{d}\alpha}{\mathrm{d}\beta}\right)\mathrm{d}\alpha$ denotes the Kullback-Leibler divergence between two probability measures. The seminal motivation was to turn (1) into a *strictly* convex optimization problem that can be—whenever μ and ν have finite support—solved on GPU via the *Sinkhorn algorithm*. Precisely, the Sinkhorn algorithm solves (4) by seeking for a fixed point of

$$\Phi : (f,g) \mapsto \left(-\varepsilon\log\int e^{\frac{g(y)-\frac{1}{2}|\bullet-y|^2}{\varepsilon}}\mathrm{d}\nu(y), -\varepsilon\log\int e^{\frac{f(x)-\frac{1}{2}|x-\bullet|^2}{\varepsilon}}\mathrm{d}\mu(x)\right) \tag{5}$$

by simply building a sequence $(f_t,g_t)_t$ starting from f_0, g_0 by setting $f_{t+1} = \Phi(f_t,g_t)$ and $g_{t+1} = \Phi(f_{t+1},g_t)$. Under mild assumptions, this sequence is guaranteed to converge toward a solution $(f^\varepsilon, g^\varepsilon)$ of (4) [9], called a pair of *(Schrödinger) potentials* for the measures μ and ν. A solution to the primal problem (3) is then obtained through the relation $\pi^\varepsilon = \exp\left(\frac{f^\varepsilon(x)+g^\varepsilon(y)-\frac{1}{2}|x-y|^2}{\varepsilon}\right)\mathrm{d}\mu(x)\mathrm{d}\nu(y)$.

It then appeared that the interest of EOT goes way beyond computational efficiency. In particular, it was proved in [3] that the *Sinkhorn divergence* [10, 4]

$$\mathrm{Sk}_\varepsilon(\mu,\nu) \coloneqq \mathrm{OT}_\varepsilon(\mu,\nu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\mu,\mu) - \frac{1}{2}\mathrm{OT}_\varepsilon(\nu,\nu) \tag{6}$$

defines a proper discrepancy between probability measures: $\mathrm{Sk}_\varepsilon(\mu,\nu) \geq 0$ and equals 0 if and only if $\mu = \nu$.[2] Given that OT_ε (thus Sk_ε) goes to OT as $\varepsilon \to 0$, the Sinkhorn divergence Sk_ε is a good candidate to mimic the OT-geometry while being more computationally and statistically efficient in numerical applications.

2 Wasserstein gradient flow of the Sinkhorn divergence

The goal of this internship is to study an open question raised in [1] is the following: for two measures $\mu_{\text{source}}, \mu_{\text{target}} \in \mathcal{P}_2(\mathbb{R}^d)$, in which case does the Wasserstein gradient flow of $\mathrm{Sk}_\varepsilon(\bullet, \mu_{\text{target}})$, initialized at $\mu_0 = \mu_{\text{source}}$ converges globally, i.e. $\mu_t \to \mu_{\text{target}}$ as $t \to \infty$?

As a starting point, the work [1] establishes the existence and uniqueness of this gradient flow (at least when μ_{source} and μ_{target} are compactly supported). Global convergence is expected to occur generically in specific (yet standard) situations only, and we propose to study two of them. We expect to treat the first case during the internship; the second case is more likely to be a first project for the Ph.D. (but of course, any preliminary results or intuition will be welcome).

The Gaussian Case. When μ_{source} and μ_{target} are Gaussian (probability) distributions, many formula are accessible in close form (see [5]). This makes the study significantly simpler and we hope that an accurate description of the Wasserstein flow of $\mathrm{Sk}_\varepsilon(\bullet, \mu_{\text{target}})$ will be accessible. With (*a priori*) increasing level of difficulty, we plan to consider the following questions:

1. Since Gaussian distributions are not compactly supported, the results of [1] does not apply straightforwardly. Given the simple explicit formula we have in that case, how can we adapt the main results of this work?

2. Can we prove global convergence, i.e. $\mu_t^\varepsilon \to \mu_{\text{target}}$, of the flow? This would provide a partial answer to the open problem raised in [1, §4.2].

[2] These properties are not satisfied by OT_ε whenever $\varepsilon > 0$.

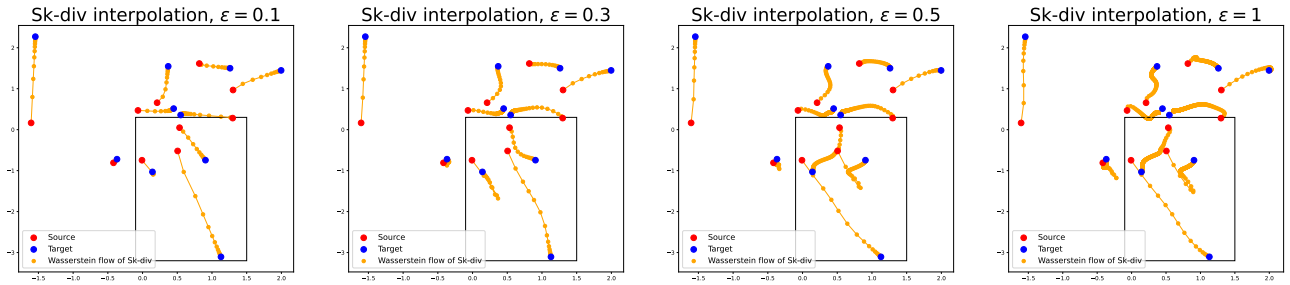

Figure 1: The Wasserstein flow of $\mathrm{Sk}_\varepsilon(\bullet, \mu_{\text{target}})$ with $n = 10$ points, for different values of ε. When $\varepsilon = 0.1$, the flow is close to following straight lines, i.e. the McCann interpolation. However, when ε increases, looking inside the black box, one can observe that the matching between μ_{source} and μ_{target} described by the flow changes, suggesting a critical value $\varepsilon_{\text{crit}} \in (0.3, 0.5)$.

3. Assuming convergence toward μ_{target} as $t \to \infty$, can we get a *convergence rate*?

This initial study will provide a playground showing what can be done in simple cases and will hopefully help us to derive general principles to study this gradient flow. Additionally, having access to close formula allows us to factor out all the possible numerical challenges related to computing (3) and thus to benchmark numerical schemes with respect to the accessible ground truth.

The Particle Case. This second study acts as a counterpart of the Gaussian case where the source and target measures are instead given by uniform measures supported on n Dirac masses, that is $\mu_{\text{source}} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\mu_{\text{target}} = \frac{1}{n}\sum_{i=1}^n \delta_{y_j}$, where $x_i, y_j \in \mathbb{R}^d$. The motivation to study this setting naturally stems from numerical applications where it is very common that one only observes i.i.d. samples from (unknown) ground distributions.

Through preliminary experiments, we numerically observe that (i) global convergence hold for generic $\mu_{\text{source}}, \mu_{\text{target}}$ and ε, (ii) when ε is small, the interpolation curve approximates the McCann interpolation and in particular the matching induced between the x_is and the y_js coincides with the permutation σ, (iii) when ε increases, a "phase transition" occurs: while the flow still globally converges, the induced matching changes, as illustrated in Figure 1. Expecting some regularity of the interpolation with respect to ε, this suggests that there is a critical value $\varepsilon_{\text{crit}}$ around which the matching changes, and in particular global convergence may fail at $\varepsilon = \varepsilon_{\text{crit}}$. Therefore, the realistic conjecture that we consider is the following:

Conjecture: *Assuming that there exists a unique optimal transport plan between* μ_{source} *and* μ_{target} *for the unregularized problem* (1)*, the Wasserstein flow of* $\mathrm{Sk}_\varepsilon(\bullet, \mu_{\text{target}})$ *initialized at* μ_{source} *converges toward* μ_{target} *for almost every* ε*.*

As for the Gaussian case, we propose to investigate the following tracks:

1. An exhaustive study of the case $n = 2$. A preliminary investigation suggests that the sign of $\langle x_1 - x_2, y_1 - y_2 \rangle$ is preserved through the flow (the edge case $= 0$ reflecting symmetric configurations) and dictates the global convergence.

2. Establish the convergence of the flow toward μ_{target} for sufficiently small ε.

3. Study the convergence for general ε, possibly characterizing the critical value $\varepsilon_{\text{crit}}$ for which the induced matching changes.

References

[1] Guillaume Carlier, Lenaic Chizat, and Maxime Laborde. Lipschitz continuity of the schrodinger map in entropic optimal transport. *arXiv preprint arXiv:2210.00225*, 2022.

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 26, 2013.

[3] 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 AISTATS*, pages 2681–2690. PMLR, 2019.

[4] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *AISTATS*, pages 1608–1617. PMLR, 2018.

[5] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between unbalanced gaussian measures has a closed form. *NeurIPS*, 33:10468–10479, 2020.

[6] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[7] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[8] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[9] Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[10] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.