Wasserstein Gradient Flows of Moreau Envelopes of f-Divergences in Reproducing Kernel Hilbert Spaces

joint work with

Sebastian Neumayer, TU Chemnitz

Gabriele Steidl, TU Berlin

Nicolaj Rux, TU Berlin

UCLA Level set seminar (Stan Osher) 19.08.2024

Goal. Recover $\nu \in \mathcal{P}(\mathbb{R}^d)$ from samples by minimizing f-divergence $D_{f,\nu}$ to ν, e.g. $\mathrm{KL}(\cdot \mid \nu)$.

Problem. Only samples $\rightsquigarrow$ empirical measures, but

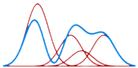
$$\mu \not\ll \nu \implies D_{f,\nu}(\mu) = \infty.$$

weak convergence

Our Solution. Regularize $D_{f,\nu} \colon \mathcal{M}(\mathbb{R}^d) \to [0,\infty]$.

1. "Kernel trick"

$$m \colon \mathcal{M}(\mathbb{R}^d) \to \mathcal{H}_K, \qquad \mu \mapsto \int_{\mathbb{R}^d} K(x,\cdot)\,\mathrm{d}\mu(x)$$



pointwise convergence

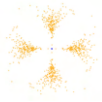
$$\text{“}D_{f,\nu} \circ m^{-1}\text{”} = G_{f,\nu} \colon \mathcal{H}_K \to [0,\infty]$$

2. Moreau envelope regularization

$$^{\lambda}G_{f,\nu}(m(\mu)) = \min_{\sigma \in \mathcal{M}_+(\mathbb{R}^d)} D_{f,\nu}(\sigma) + \frac{1}{2\lambda}\|m(\sigma) - m(\mu)\|^2_{\mathcal{H}_K}, \qquad \lambda > 0.$$

We prove existence & uniqueness of W_2 gradient flows of $({}^{\lambda}G_{f,\nu}) \circ m$.

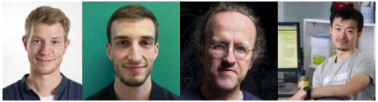
Simulate particle flows $=$ W_2 gradient flows starting at empirical measure



Literature review of prior work

- KALE functional = MMD-regularized KL divergence
  [Glaser, Arbel, Gretton. NeurIPS'21]
  No Moreau envelope interpretation.

- Kernel methods of moments = f-divergence-regularized MMD
  [Kremer, Nemmour, Schölkopf, Zhu. ICML'23]
  Doesn't cover all f-divergences.



- (f, Γ)-divergence = Pasch-Hausdorff envelope of f-divergences
  [Birrell, Dupuis, Katsoulakis, Pantazis, Rey-Bellet, JMLR'23]
  Yields only Lipschitz, not differentiable functional.

- W_1-Moreau envelope of f-divergences [Terjék. ICML'21]
  No RKHS, which makes optimization finite-dimensional, hence tractable.

1. RKHS & MMD

2. Moreau envelopes

3. f-divergences

4. MMD-Moreau envelopes of f-divergences

5. Wasserstein gradient flow

6. WGF of MMD-Moreau envelopes of f-divergences

"Kernel trick": embed data into high-dimensional Hilbert space.

$K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **symmetric, positive definite**.

We consider **radial** kernels $K(x,y) = \phi(\|x-y\|_2^2)$ with

$\phi \in \mathcal{C}^\infty((0,\infty)) \cap \mathcal{C}^2([0,\infty))$, $(-1)^k \phi^{(k)}(r) \geq 0$, $\forall k \in \mathbb{N}$, $r > 0$.

$\rightsquigarrow$ **reproducing kernel Hilbert space (RKHS)**

$\mathcal{H}_K := \overline{\operatorname{span}}(\{K(x,\cdot) : x \in \mathbb{R}^d\})$. Key property: $h \mapsto h(x)$ cts.

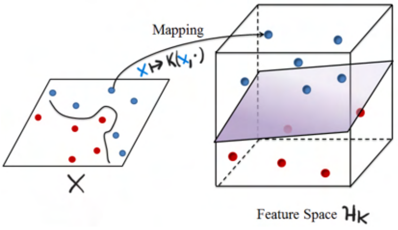
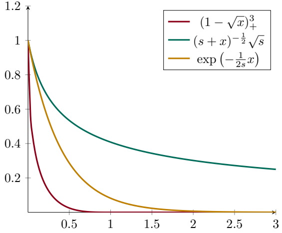


FIG. 1: "Kernel trick". Source: songcy.net/posts/story-of-basis-and-kernel-part-2/



Examples (with parameter $s > 0$).

- Gaussian $\phi(r) = \exp\left(-\frac{1}{2s} r\right)$
- inverse multiquadric $\phi(r) := (s+r)^{-\frac{1}{2}}$
- spline $\phi(r) = \max(0, (1-\sqrt{r})^{s+2})$.

Nonexamples.

- Laplace $\phi(r) = \exp(-\frac{1}{2s}\sqrt{r})$ (not smooth enough)
- $K(x,y) = \|x\| + \|y\| - \|x-y\|$ (not radial)

Kernel mean embedding and Maximum Mean Discrepancy

"Kernel trick for signed measures" $\mu \in \mathcal{M}(\mathbb{R}^d)$ (instead of points): **kernel mean embedding (KME)**

$$m\colon \mathcal{M}(\mathbb{R}^d) \to \mathcal{H}_K, \qquad \mu \mapsto \int_{\mathbb{R}^d} K(x, \cdot) \, \mathrm{d}\mu(x).$$

$$\begin{tikzcd}
 & \mathcal{H}_K & \\
\mathbb{R}^d \arrow[ur, "x \mapsto K(x, \cdot)"] \arrow[rr, hook, "x \mapsto \delta_x"] & \circlearrowleft & \mathcal{M}(\mathbb{R}^d) \arrow[ul, hook, "m"']
\end{tikzcd}$$

We require m to be injective ($\mathcal{H}_K$ "characteristic") $\iff$ $\mathcal{H}_K \subset \mathcal{C}_0(\mathbb{R}^d)$ dense.

$\rightsquigarrow$ Instead of measures, compare their embeddings in $\mathcal{H}_K$: **maximum mean discrepancy (MMD)**

$$d_K\colon \mathcal{M}(\mathbb{R}^d) \times \mathcal{M}(\mathbb{R}^d) \to [0, \infty), \qquad (\mu, \nu) \mapsto \|m(\mu - \nu)\|_{\mathcal{H}_K}.$$

m injective $\implies$ d_K **is a metric**, but $(\mathcal{M}(\mathbb{R}^d), d_K)$ is *not* complete.

Easy to evaluate, e.g. for discrete measures since

$$d_K(\mu, \nu)^2 = \int_{\mathbb{R}^d \times \mathbb{R}^d} K(x, y) \, \mathrm{d}(\mu - \nu)(x) \, \mathrm{d}(\mu - \nu)(y) \qquad \forall \mu, \nu \in \mathcal{M}(\mathbb{R}^d).$$

1. RKHS & MMD

2. Moreau envelopes

3. f-divergences

4. MMD-Moreau envelopes of f-divergences

5. Wasserstein gradient flow

6. WGF of MMD-Moreau envelopes of f-divergences

Let $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ Hilbert space, $f \in \Gamma_0(H)$, i.e. $f\colon H \to (-\infty, \infty]$ **convex**, lower semicontinuous, $\mathrm{dom}(f) \coloneqq \{x \in H : f(x) < \infty\} \neq \emptyset$.

For $\varepsilon > 0$, the **ε-Moreau envelope** of f,

$$^{\varepsilon}f\colon H \to \mathbb{R}, \qquad x \mapsto \min\left\{f(x') + \frac{1}{2\varepsilon}\|x - x'\|^2 : x' \in H\right\}$$

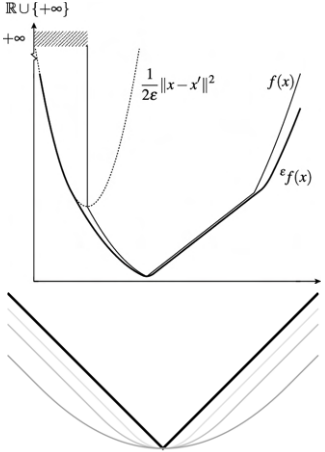
is convex, **differentiable** regularization of f **preserving its minimizers**.

Asymptotics: $^{\varepsilon}f(x) \nearrow f(x)$ for $\varepsilon \searrow 0$ and $^{\varepsilon}f(x) \searrow \inf(f)$ for $\varepsilon \to \infty$.

$(\varepsilon, x) \mapsto {}^{\varepsilon}f(x)$ is viscosity solution of Hamilton-Jacobi equation:

$$\begin{cases} \partial_\varepsilon({}^{\varepsilon}f)(x) + \frac{1}{2}\|\nabla({}^{\varepsilon}f)(x)\|_2^2 = 0, \\ {}^{0}f(x) \to f(x). \end{cases}$$

[Osher, Heaton, Fung, PNAS 120, **14**, 2023].



Moreau envelope of an extended-real-valued non-differentiable function (top) and of $|\cdot|$ for different ε (bottom).

1. RKHS & MMD

2. Moreau envelopes

3. f-divergences

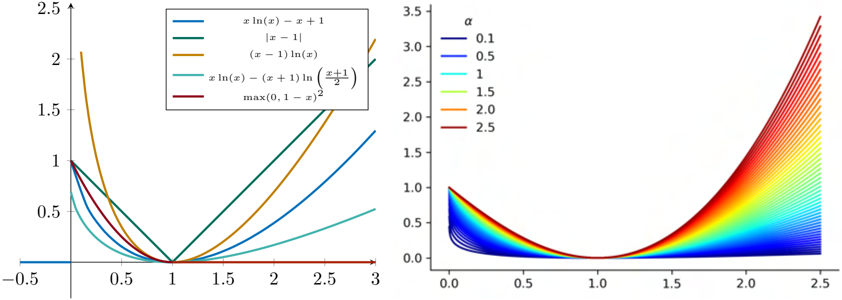
4. MMD-Moreau envelopes of f-divergences

5. Wasserstein gradient flow

6. WGF of MMD-Moreau envelopes of f-divergences

We consider $f \in \Gamma_0(\mathbb{R})$ with $f|_{(-\infty,0)} \equiv \infty$ and with **unique minimizer at 1**: $f(1) = 0$ and positive recession constant $f'_\infty \coloneqq \lim_{t\to\infty} \frac{1}{t} f(t) > 0$.

Examples. $f_{\mathrm{KL}}(x) \coloneqq x\ln(x) - x + 1$ for $x \geq 0$ yields the **Kullback-Leibler divergence** and $f_\alpha(x) \coloneqq \frac{1}{\alpha-1}(x^\alpha - \alpha x + \alpha - 1)$ the **Tsallis-α divergence** T_α for $\alpha > 0$. In the limit: $T_1 = \mathrm{KL}$.



Left: Examples of entropy functions, except the red. Right: The functions f_α for $\alpha \in [0.1, 2.5]$.

f**-divergence** of $\mu = \rho\nu + \mu_s \in \mathcal{M}_+(\mathbb{R}^d)$ (unique Lebesgue decomposition) to $\nu \in \mathcal{M}_+(\mathbb{R}^d)$

$$\boxed{\begin{aligned} D_{f,\nu}(\rho\nu + \mu_s) &\coloneqq \int_{\mathbb{R}^d} f \circ \rho \, \mathrm{d}\nu + f'_\infty \cdot \mu_s(\mathbb{R}^d) && (\infty \cdot 0 \coloneqq 0) \\ &= \sup_{h \in C_b(\mathbb{R}^d; \mathrm{dom}(f^*))} \mathbb{E}_\mu[h] - \mathbb{E}_\nu[f^* \circ h], && \mathbb{E}_\sigma[h] \coloneqq \int_{\mathbb{R}^d} h(x) \, \mathrm{d}\sigma(x) \end{aligned}}$$

The **convex conjugate** of f is $f^* \colon \mathbb{R} \to (-\infty, \infty]$, $\quad s \mapsto \sup\{st - f(t) : t \geq 0\}$.

Theorem (Properties of $D_{f,\nu}$)

$D_{f,\nu} \colon \mathcal{M}_+(\mathbb{R}^d) \to [0, \infty]$ *is convex, weak* lower semicontinuous. We have:* $D_{f,\nu}(\mu) = 0 \iff \mu = \nu$.

MMD-Regularized f-divergence - Moreau envelope interpretation

We define the **MMD-regularized f-divergence** functional

$$\boxed{D_{f,\nu}^{\lambda}(\mu) \coloneqq \min\left\{D_{f,\nu}(\sigma) + \frac{1}{2\lambda} d_K(\mu,\sigma)^2 : \sigma \in \mathcal{M}(\mathbb{R}^d)\right\}, \qquad \lambda > 0,\ \mu \in \mathcal{M}(\mathbb{R}^d).} \tag{1}$$

Theorem (Moreau envelope interpretation of $D_{f,\nu}^{\lambda}$ [NSSR24])

The $\mathcal{H}_K$-extension of $D_{f,\nu}$,

$$G_{f,\nu} \colon \mathcal{H}_K \to [0,\infty], \quad h \mapsto \begin{cases} D_{f,\nu}(\mu), & \text{if } \exists \mu \in \mathcal{M}_+(\mathbb{R}^d) \text{ s.t. } h = m(\mu), \\ \infty, & \text{else.} \end{cases}$$

*is convex, **lower semicontinuous** and its Moreau envelope concatenated with m is the MMD-regularized f-divergence:*

$$\boxed{{}^{\lambda}G_{f,\nu} \circ m = D_{f,\nu}^{\lambda}}$$

$$\begin{tikzcd}
[0,\infty) & \mathcal{H}_K \arrow[l, "{}^{\lambda}G_{f,\nu}"'] \arrow[d, "G_{f,\nu}"] \\
\mathcal{M}(\mathbb{R}^d) \arrow[u, "D_{f,\nu}^{\lambda}"] \arrow[ur, hook, "m"] \arrow[r, "D_{f,\nu}"'] & [0,\infty]
\end{tikzcd}$$

- **Dual formulation**

$$D_{f,\nu}^{\lambda}(\mu) = \max\left\{\mathbb{E}_{\mu}[p] - \mathbb{E}_{\nu}[f^* \circ p] - \frac{\lambda}{2}\|p\|_{\mathcal{H}_K}^2 : p \in \mathcal{H}_K,\ p \le f'_{\infty}\right\}. \tag{2}$$

$\hat{p} \in \mathcal{H}_K$ maximizes (2) $\iff$ $\hat{g} = m(\mu) - \lambda\hat{p}$ is primal solution.

$$\frac{\lambda}{2}\|\hat{p}\|_{\mathcal{H}_K}^2 \le D_{f,\nu}^{\lambda}(\mu) \le \|\hat{p}\|_{\mathcal{H}_K}(\|m_{\mu}\|_{\mathcal{H}_K} + \|m_{\nu}\|_{\mathcal{H}_K}) \quad \text{and} \quad \|\hat{p}\|_{\mathcal{H}_K} \le \frac{2}{\lambda} d_K(\mu, \nu).$$

- $D_{f,\nu}^{\lambda}$ is Fréchet differentiable on $\mathcal{M}(\mathbb{R}^d)$ and its gradient is **λ-Lipschitz** with respect to d_K:

$$\nabla D_{f,\nu}^{\lambda}(\mu) = \operatorname{argmax}(2).$$

Theorem. (Properties of $D_{f,\nu}^{\lambda}$) [NSSR24]

- **Asymptotic** regimes: Mosco resp. pointwise convergence (if $0 \in \operatorname{int}(\operatorname{dom}(f^*))$ resp. f^* differentiable in 0)

$$D_{f,\nu}^{\lambda} \to D_{f,\nu} \quad \lambda \searrow 0 \qquad \text{and} \qquad (1+\lambda) D_{f,\nu}^{\lambda} \to \frac{1}{2} d_K(\cdot,\nu)^2 \quad \lambda \to \infty$$

	MMD metric d_K	f-divergence D_f
ingredients	characteristic kernel K	convex, lsc f, $f(1) = 0$
examples	Gaussian, IMQ, Matérn	KL, JSD, Hellinger, χ^2
positive definite	👍	👍
symmetric	👍	👎
triangle inequality	👍	👎
variational formulation	👍	👍
topology on measures	🚫	💪
geometry on measures	😐	🥴

- **Divergence property**: $D_{f,\nu}^{\lambda}(\mu) = 0 \iff \mu = \nu$.
- If f^* is differentiable in 0, then $(\mu, \nu) \mapsto D_{f,\nu}^{\lambda}(\mu)$ **metrizes weak convergence** on $\mathcal{M}_+(\mathbb{R}^d)$-balls.

1. RKHS & MMD

2. Moreau envelopes

3. f-divergences

4. MMD-Moreau envelopes of f-divergences

5. Wasserstein gradient flow

6. WGF of MMD-Moreau envelopes of f-divergences

Wasserstein space and generalized geodesics

$\mathcal{P}_2(\mathbb{R}^d) \coloneqq \{\mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|x\|_2^2 \,\mathrm{d}\mu(x) < \infty\}$, $\|\cdot\|_2$ Eucl. norm.

$$W_2(\mu,\nu)^2 = \min_{\pi \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\pi(x,y), \quad \mu,\nu \in \mathcal{P}_2(\mathbb{R}^d).$$

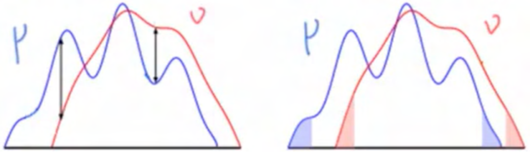


Fig. 2: Vertical (L_2) vs. horizontal (W_2) mass displacement. ©A. Korba

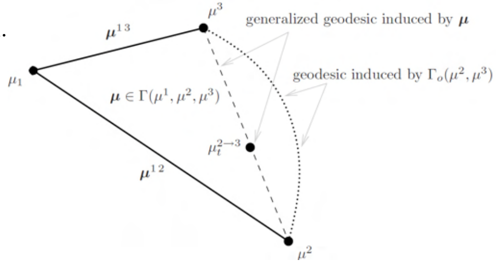


Fig. 3: Generalized geodesic from μ_2 to μ_3 with base μ_1 [AGS08].

Definition (Generalized geodesic convexity)

A function $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ is *M-convex along generalized geodesics* with $M \in \mathbb{R}$ if, for every $\sigma,\mu,\nu \in \operatorname{dom}(\mathcal{F})$, there exists a $\boldsymbol{\alpha} \in \mathcal{P}_2(\mathbb{R}^{3d})$with $(P_{1,2})_\#\boldsymbol{\alpha} \in \Gamma^{\mathrm{opt}}(\sigma,\mu)$ and $(P_{1,3})_\#\boldsymbol{\alpha} \in \Gamma^{\mathrm{opt}}(\sigma,\nu)$ s.t.

$$\mathcal{F}\left(((1-t)P_2+tP_3)_\#\boldsymbol{\alpha}\right) \le (1-t)\,\mathcal{F}(\mu)+t\,\mathcal{F}(\nu)-\frac{M}{2}t(1-t)\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}\|y-z\|_2^2 \,\mathrm{d}\boldsymbol{\alpha}(x,y,z), \qquad \forall t \in [0,1].$$

Definition (Fréchet subdifferential in Wasserstein space)

The (reduced) Fréchet subdifferential of $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ at $\mu \in \operatorname{dom}(\mathcal{F})$ is

$$\partial \mathcal{F}(\mu) := \left\{ \xi \in L^2(\mathbb{R}^d; \mu) : \mathcal{F}(\nu) - \mathcal{F}(\mu) \geq \inf_{\pi \in \Gamma^{\mathrm{opt}}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle \xi(x_1), x_2 - x_1 \rangle \, \mathrm{d}\pi(x,y) + o(W_2(\mu, \nu)) \right\}$$

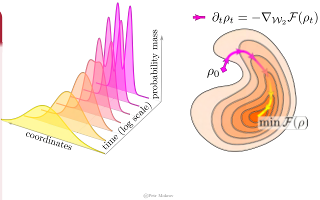
A curve $\gamma\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is *absolutely continuous* if $\exists$ L^2-Borel velocity field $v\colon \mathbb{R}^d \times (0,\infty) \to \mathbb{R}^d$ s.t.

$$\partial_t \gamma_t + \nabla \cdot (v_t \gamma_t) = 0, \qquad (t,x) \in (0,\infty) \times \mathbb{R}^d, \text{ weakly.} \qquad \text{(Continuity Eq.)}$$

Definition (Wasserstein gradient flow)

A locally absolutely continuous curve $\gamma\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ with velocity field $v_t \in \mathrm{T}_{\gamma_t} \mathcal{P}_2(\mathbb{R}^d)$ is a *Wasserstein gradient flow with respect to* $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, \infty]$ if

$$v_t \in -\partial \mathcal{F}(\gamma_t), \quad \text{for a.e. } t > 0.$$



Theorem (Convexity and gradient of $D_{f,\nu}^{\lambda}$ [NSSR24])

*Since K is radial and smooth, $D_{f,\nu}^{\lambda}$ is M**-convex along generalized geodesics** with $M := -8\lambda^{-1}\sqrt{(d+2)\phi''(0)\phi(0)}$ and its (reduced) Fréchet **subdifferential** is $\partial D_{f,\nu}^{\lambda}(\mu) = \{\nabla \operatorname{argmax}(2)\}$.*

Remark. M seems non-optimal, since for $\lambda \to 0$, $D_{f,\nu}^{\lambda} \to D_{f,\nu}$ and $D_{f,\nu}$ is 0-convex, but $M \to -\infty$.

Corollary

*There **exists a unique Wasserstein gradient flow** $(\gamma_t)_{t>0}$ of $D_{f,\nu}^{\lambda}$ starting at $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, fulfilling the continuity equation $\partial_t \gamma_t = \nabla \cdot \big(\gamma_t \big(\partial D_{f,\nu}^{\lambda}(\gamma_t)\big)\big)$, $\gamma_0 = \mu_0$.*

Lemma (Particle flows are W_2 gradient flows)

If μ_0 is empirical, then so is γ_t for all $t > 0$.

Take i.i.d. samples $(x_j^{(0)})_{j=1}^N \sim \mu_0$ and $(y_j)_{j=1}^M \sim \nu$. Forward Euler discretization in time with step size $\tau > 0$ yields

$$\gamma_{n+1} := (\mathrm{id} - \tau \nabla \hat{p}_n)_\# \gamma_n, \qquad \hat{p}_n = \operatorname{argmax} \text{ in } D_{f,\nu}^{\lambda}(\gamma_n)$$

so $(\gamma_n)_{n\in\mathbb{N}} = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_j^{(n)}}$ with gradient step

$$\boxed{x_j^{(n+1)} = x_j^{(n)} - \tau \nabla \hat{p}_n\big(x_j^{(n)}\big), \qquad j \in \{1, \dots, N\},\ n \in \mathbb{N}.}$$

THEOREM (REPRESENTER-TYPE THEOREM [NSSR24])

*If $f'_\infty = \infty$ or if $\lambda > 2 d_K(\gamma_n, \nu)\sqrt{\phi(0)}\frac{1}{f'_\infty}$, then finding $\hat{p}_n$ is a **finite-dimensional strongly convex** problem.*

To find $\hat{p}_n$, we use **L-BFGS-B**, a quasi-Newton method. We use annealing strategy for λ if $f'_\infty < \infty$.

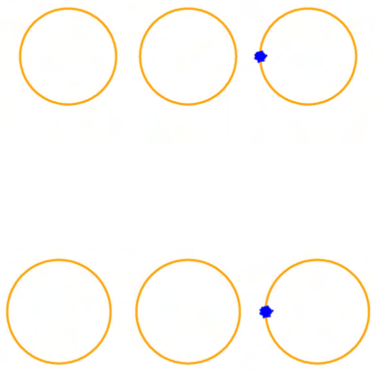


FIG. 4: IMQ kernel, $\lambda = \frac{1}{100}$ $\tau = \frac{1}{1000}$, Top: Tsallis-3 divergence, Bottom: Tsallis-$\frac{1}{2}$ divergence, with annealing.

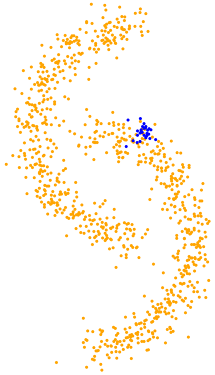


FIG. 5: Number of starting particles N, less than number of samples of target, $M \rightsquigarrow$ quantization

- **Non-differentiable** (e.g. Laplace $= \frac{1}{2}$-Matérn) and unbounded (e.g. Riesz, Coulomb) **kernels**.
- **Convergence rates** in suitable metric.
- Prove consistency bounds [Leclerc, Mérigot, Santambrogio, Stra. 2020] and **better** M**-convexity estimates**.
- Convergence for annealing strategy?
- Different domains, e.g. compact subsets of $\mathbb{R}^d$ (manifolds like sphere, torus), groups, infinite-dimensional spaces.
- Regularize other divergences, e.g. Rényi divergences, Bregman divergences.
- Gradient flow of $D_{f,\nu}^{\lambda}$ with respect to other metrics, like Kantorovich-Hellinger (related to unbalanced OT), MMD, Fisher-Rao or Wasserstein-p for $p \in [1, \infty]$.
- More elaborate time discretizations, variable step sizes.

- We created novel objective. Minimizing it allows sampling from a target measure of which only samples are known.
- Clear, rigorous interpretation using Convex Analysis and RKHS.
- Theory covers (almost) all f-divergences.
- Best of both worlds: $D_{f,\nu}^{\lambda}$ interpolates between $D_{f,\nu}$ and $d_K(\cdot, \nu)^2$.
- Effective algorithms due to (modified) representer theorem & GPU / PyTorch.

Thank you for your attention!

I am happy to take any questions.

Paper link: `arxiv.org/abs/2402.04613`

My website: `viktorajstein.github.io`

[AGS08] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, 2 ed., Springer Science & Business Media, 2008.

[BDK+22] Jeremiah Birrell, Paul Dupuis, Markos A. Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet, (f,Γ)*-divergences: Interpolating between* f*-divergences and integral probability metrics*, J. Mach. Learn. Res. **23** (2022), no. 39, 1–70.

[GAG21] Pierre Glaser, Michael Arbel, and Arthur Gretton, *KALE flow: A relaxed KL gradient flow for probabilities with disjoint support*, Advances in Neural Information Processing Systems (Virtual event), vol. 34, 6–14 Dec 2021, pp. 8018–8031.

[HWAH24] J. Hertrich, C. Wald, F. Altekrüger, and P. Hagemann, *Generative sliced MMD flows with Riesz kernels*, International Conference on Learning Representations (ICLR) (Vienna, Austria), 7 – 11 May 2024.

[KYSZ23] H. Kremer, Nemmour Y., B. Schölkopf, and J.-J. Zhu, *Estimation beyond data reweighting: kernel methods of moments*, ICML'23: Proceedings of the 40th International Conference on Machine Learning (Honolulu, Hawaii, USA), vol. 202, July 23 - 29 2023, p. 17745–17783.

[LMS17] Matthias Liero, Alexander Mielke, and Giuseppe Savaré, *Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures*, Invent. Math. **211** (2017), no. 3, 969–1117.

[LMSS20] Hugo Leclerc, Quentin Mérigot, Filippo Santambrogio, and Federico Stra, *Lagrangian discretization of crowd motion and linear diffusion*, SIAM J. Numer. Anal. **58** (2020), no. 4, 2093–2118. MR 4123686

[Ter21] Dávid Terjék, *Moreau-Yosida f-divergences*, International Conference on Machine Learning (ICML) (Virtual event), PMLR, Jul 18–24 2021, pp. 10214–10224.

Shameless plug: other works

Interpolating between OT and KL regularized OT using Rényi Divergences

Rényi divergence $\notin \{f\text{-div.}, \text{Bregman div.}\}$, $\alpha \in (0,1)$

$$R_\alpha(\mu \mid \nu) := \frac{1}{\alpha - 1} \ln \left[\int_X \left(\frac{\mathrm{d}\mu}{\mathrm{d}\tau} \right)^\alpha \left(\frac{\mathrm{d}\nu}{\mathrm{d}\tau} \right)^{1-\alpha} \mathrm{d}\tau \right],$$

$$\mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) := \min_{\pi \in \Pi(\mu,\nu)} \langle c, \pi \rangle + \varepsilon R_\alpha(\pi \mid \mu \otimes \nu)$$

is a metric, where $\varepsilon > 0$, $\mu, \nu \in \mathcal{P}(X)$, X compact.

$$\boxed{\mathrm{OT}(\mu,\nu) \xleftarrow[\text{or } \varepsilon \to 0]{\alpha \searrow 0} \mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) \xrightarrow{\alpha \nearrow 1} \mathrm{OT}_\varepsilon^{\mathrm{KL}}(\mu,\nu).}$$

In the works: **debiased** Rényi-Sinkhorn divergence

$$\mathrm{OT}_{\varepsilon,\alpha}(\mu,\nu) - \frac{1}{2}\mathrm{OT}_{\varepsilon,\alpha}(\mu,\mu) - \frac{1}{2}\mathrm{OT}_{\varepsilon,\alpha}(\nu,\nu).$$

W_2 gradient flows of $d_K(\cdot,\nu)^2$ with $K(x,y) := -|x-y|$ in 1D.

Reformulation as **maximal monotone** inclusion Cauchy problem in $L_2(0,1)$ via **quantile functions**.

Comprehensive description of solutions' behavior, **instantaneous measure-to-L^∞ regularization**, implicit Euler is simple.