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

Sebastian Neumayer, Viktor Stein, Gabriele Steidl, Nicolaj Rux

Goal: minimize **f-divergence loss** $D_{f,\nu}$ with target measure $\nu \in \mathcal{M}_+(\mathbb{R}^d)$ (e.g. generative adversarial networks, variational inference).
Often **only samples** are available $\rightsquigarrow$ empirical measures.
BUT: D_f between empirical measures is $\infty \rightsquigarrow$ **regularize f-divergence**.

- **Contribution.** Prove identification of MMD-regularized f-divergence functional as Moreau envelope in RKHS. Existence and uniqueness of its Wasserstein gradient flow. Flow starting at empirical measure is particle flow.
- **Prior work.** Regularize MMD with f-divergence [5], MMD-Pasch-Hausdorff envelope of f-divergences [7], W_1-Moreau envelope of f-divergences [8].
- **Method.** Euler forward discretize particle flow (= gradient descent on the positions).
- **Result.** We can simulate particle flows for divergences with finite and infinite recession constant f'_∞. Tsallis-α divergence with moderately large α outperforms KL-divergence ($\alpha = 1$): faster target recovery and more stable.

Reproducing Kernel Hilbert Space, KME, Maximum Mean Discrepancy

$K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ **symmetric, positive definite, bounded** kernel with $K(x,\cdot) \in \mathcal{C}_0(\mathbb{R}^d)$.
We focus on **radial** kernels $K(x,y) = \phi(\|x-y\|_2^2)$ with $\phi \in \mathcal{C}^2([0,\infty))$ completely monotone.
Examples. Gaussian $\phi(r) = \exp\left(-\frac{1}{2s}r\right)$, IMQ $\phi(r) := (s+r)^{-\frac{1}{2}}$, spline $\phi(r) = (1-\sqrt{r})_+^{q+2}$.
$\rightsquigarrow$ **reproducing kernel Hilbert space (RKHS)** $\mathcal{H}_K := \overline{\operatorname{span}}(\{K(x,\cdot) : x \in \mathbb{R}^d\})$.
The **kernel mean embedding (KME)** of finite signed measures, $\mathcal{M}(\mathbb{R}^d)$, into $\mathcal{H}_K$ is

$$m\colon \mathcal{M}(\mathbb{R}^d) \to \mathcal{H}_K, \qquad \mu \mapsto m_\mu := \int_{\mathbb{R}^d} K(x,\cdot)\,\mathrm{d}\mu(x). \tag{1}$$

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

We require m to be injective ($\mathcal{H}_K$ "characteristic") $\iff$ $\mathcal{H}_K \subset \mathcal{C}_0(\mathbb{R}^d)$ dense.
Then the **maximum mean discrepancy (MMD)**

$$d_K\colon \mathcal{M}(\mathbb{R}^d) \times \mathcal{M}(\mathbb{R}^d) \to [0,\infty), \qquad (\mu,\nu) \mapsto \|m_\mu - m_\nu\|_{\mathcal{H}_K}. \tag{2}$$

is an **incomplete** metric. We have for all $\mu,\nu \in \mathcal{M}(\mathbb{R}^d)$

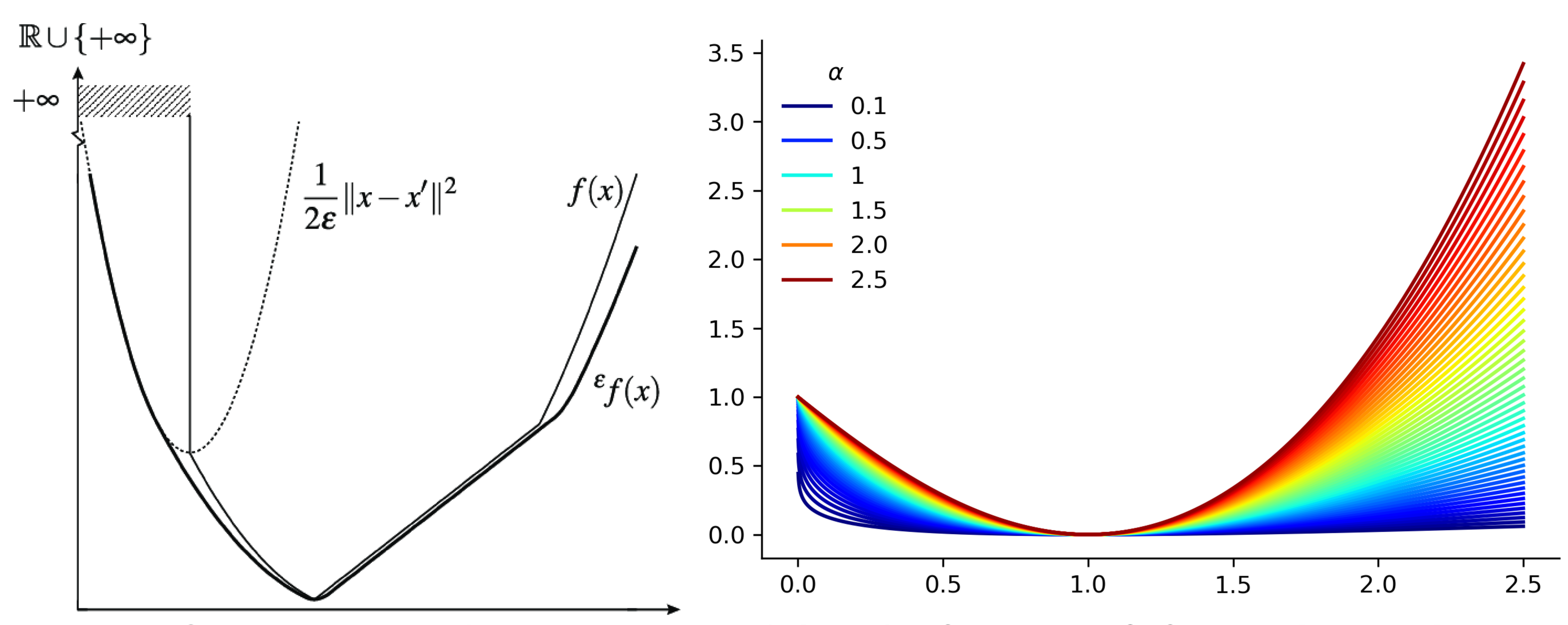
$$d_K(\mu,\nu)^2 = \int_{\mathbb{R}^d\times\mathbb{R}^d} K(x,y)\,\mathrm{d}(\mu-\nu)(x)\,\mathrm{d}(\mu-\nu)(y). \tag{3}$$

Regularization in Convex Analysis

$(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ Hilbert space, $f \in \Gamma_0(H)$, i.e. $f\colon H \to (-\infty,\infty]$ **convex** lower semicontinuous, with $\operatorname{dom}(f) := \{x \in H : f(x) < \infty\} \neq \emptyset$.
For $\varepsilon > 0$, the **ε-Moreau envelope** of f,

$$^{\varepsilon}f\colon H \to \mathbb{R}, \qquad x \mapsto \min\left\{ f(x') + \frac{1}{2\varepsilon}\|x-x'\|^2 : x' \in H \right\} \tag{4}$$

is convex, **differentiable** regularization of f **preserving its minimizers**.
Asymptotic regimes: $^{\varepsilon}f(x) \nearrow f(x)$ for $\varepsilon \searrow 0$ and $^{\varepsilon}f(x) \searrow \inf(f)$ for $\varepsilon \to \infty$.



Left: Moreau envelope ©Trygve U. Helgaker. **Right:** The functions f_α for $\alpha \in [0.1, 2.5]$.

f-divergence

We consider $f \in \Gamma_0(\mathbb{R})$ with $f|_{(-\infty,0)} \equiv \infty$ and with unique minimizer at 1: $f(1) = 0$ and $f'_\infty := \lim_{t\to\infty} \frac{1}{t}f(t) > 0$. Its convex conjugate is

$$f^*\colon \mathbb{R} \to (-\infty,\infty], \qquad s \mapsto \sup\{st - f(t) : t \geq 0\}.$$

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

$$D_{f,\nu}(\rho\nu + \mu_s) := \int_{\mathbb{R}^d} f\circ\rho\,\mathrm{d}\nu + f'_\infty \cdot \mu_s(\mathbb{R}^d) \qquad (\infty\cdot 0 := 0) \tag{5}$$

$$= \sup_{h\in\mathcal{C}_b(\mathbb{R}^d;\operatorname{dom}(f^*))} \mathbb{E}_\mu[h] - \mathbb{E}_\nu[f^*\circ h], \qquad \mathbb{E}_\mu[h] := \int_{\mathbb{R}^d} h(x)\,\mathrm{d}\mu(x) \tag{6}$$

$D_{f,\nu}$ is convex and weak* lower semicontinuous.
Examples. $f_{\mathrm{KL}}(x) := x\ln(x) - x + 1$ for $x \geq 0$ yields the **Kullback-Leibler divergence** and $f_\alpha(x) := \frac{1}{\alpha-1}(x^\alpha - \alpha x + \alpha - 1)$ the **Tsallis-α divergence** T_α for $\alpha > 0$. We have $T_1 = \mathrm{KL}$.

MMD-Regularized f-divergence

The **MMD-regularized f-divergence** functional is

$$\boxed{D_{f,\nu}^\lambda(\mu) := \min\left\{ D_{f,\nu}(\sigma) + \frac{1}{2\lambda}d_K(\mu,\sigma)^2 : \sigma \in \mathcal{M}(\mathbb{R}^d)\right\}, \qquad \mu \in \mathcal{M}(\mathbb{R}^d).} \tag{7}$$

Generalizes the KALE-functional [4], which is recovered for $f = f_{\mathrm{KL}}$.

Theorem. (Moreau envelope interpretation)
The $\mathcal{H}$-extension of $D_{f,\nu}$,

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

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

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

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

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

1. **Dual formulation**
$$D_{f,\nu}^\lambda(\mu) = \max\left\{ \mathbb{E}_\mu[h] - \mathbb{E}_\nu[f^*\circ h] - \frac{\lambda}{2}\|h\|_{\mathcal{H}_K}^2 : h \in \mathcal{H}_K,\ h \leq f'_\infty \right\}. \tag{8}$$
2. $D_{f,\nu}^\lambda$ is Fréchet diff'able with $\frac{1}{\lambda}$**-Lipschitz gradient** with respect to d_K:
$$\nabla D_{f,\nu}^\lambda(\mu) = \operatorname{argmax}(8).$$
3. **Asymptotic** regimes: Mosco resp. pointwise convergence
$$D_{f,\nu}^\lambda \to D_{f,\nu} \quad \lambda \searrow 0 \qquad \text{and} \qquad (1+\lambda)D_{f,\nu}^\lambda \to \frac{1}{2}d_K(\cdot,\nu)^2 \quad \lambda \to \infty$$
4. **Divergence property**: $D_{f,\nu}^\lambda(\mu) = 0 \iff \mu = \nu$.
5. $(\mu,\nu) \mapsto D_{f,\nu}^\lambda(\mu)$ **metrizes weak convergence** on $\mathcal{M}_+(\mathbb{R}^d)$-balls.

Wasserstein Gradient Flow with respect to $D_{f,\nu}^\lambda$

$D_{f,\nu}^\lambda$ is $(-M)$**-convex along generalized geodesics** with $M := \frac{8}{\lambda}\sqrt{(d+2)\phi''(0)\phi(0)}$.
(reduced) Fréchet **subdifferential**: $\partial D_{f,\nu}^\lambda(\mu) = \{\nabla \operatorname{argmax}(8)\}$.
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 $\partial_t\gamma_t = \operatorname{div}\left(\gamma_t\nabla\left(\partial D_{f,\nu}^\lambda(\gamma_t)\right)\right)$, $\gamma_0 = \mu_0$.
If μ_0 is empirical, then so is μ_t for all $t > 0$ (particle flows are W_2 gradient flows).

Numerical Experiments - Particle Descent Algorithm

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

$$\boxed{x_j^{(n+1)} = x_j^{(n)} - \tau\nabla\hat{p}_n\big(x_j^{(n)}\big), \qquad \hat{p}_n = \operatorname{argmax} \text{ in } D_{f,\nu}^\lambda(\mu_n) \qquad j \in \{1,\dots,N\},\ n \in \mathbb{N}.}$$

Representer-type theorem. If $f'_\infty = \infty$ or if $\lambda > 2d_K(\mu_n,\nu)\sqrt{\phi(0)}\frac{1}{f'_\infty}$, then finding $\hat{p}_n$ is a **finite-dimensional strongly convex** problem (we solve it with **L-BFGS-B**).

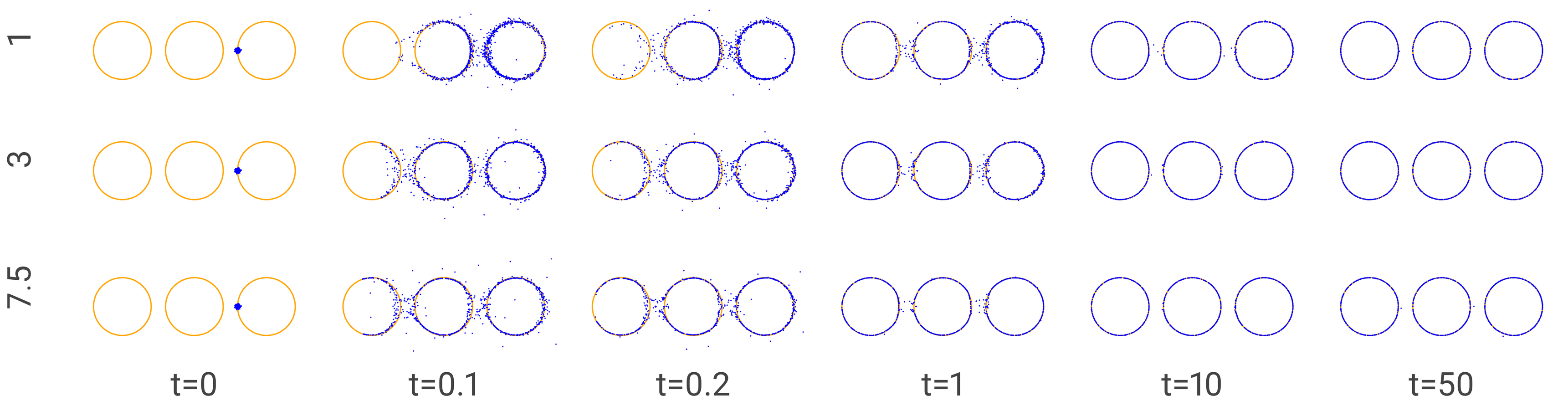


Figure 1. Wasserstein gradient flow of the regularized Tsallis-α divergence $D_{f_\alpha,\nu}^\lambda$ for $\alpha \in \{1,3,7.5\}$, where ν are the three rings. Code: `https://github.com/ViktorAJStein/Regularized_f_Divergence_Particle_Flows`

Further work. Non-differentiable (e.g. Laplace = $\frac{1}{2}$-Matérn) and unbounded (e.g. Riesz, Coulomb) kernels. Other divergences, e.g. Rényi. Different time discretizations. Prove consistency bounds [1] and convergence rates.

[1] H. Leclerc, Q. Mérigot, F. Santambrogio and F. Stra (2020) Lagrangian discretization of crowd motion and linear diffusion. SIAM J. Numer. Anal. **58** (4).
[2] L. Ambrosio, N. Gigli and G. Savare (2008). Gradient flows in metric spaces and in the space of probability measures, 2nd edition. Lectures in Mathematics ETH Zürich. Birkhäuser.
[3] J. Hertrich, C. Wald, F. Altekrüger and P. Hagemann (2024). Generative sliced MMD flows with Riesz kernels. ICLR'24.
[4] P. Glaser, M. Arbel and A. Gretton (2021) KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. NeurIPS'21.
[5] H. Kremer, Y., Nemmour, B. Schölkopf and J.-J. Zhu (2023) Estimation beyond data reweighting: kernel methods of moments. ICML'23.
[6] M. Liero, A. Mielke and G. Savaré (2017) Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. Invent. Math. 211 **3**.
[7] J. Birrell et. al. (2022) (f,Γ)-Divergences: Interpolating between f-Divergences and Integral Probability Metrics. J. Mach. Learn. Res. 23 **39**.
[8] D. Terjék (2021) Moreau-Yosida f-divergences. ICML'21.

TECHNISCHE UNIVERSITÄT BERLIN
TECHNISCHE UNIVERSITÄT CHEMNITZ