Wasserstein gradient flows for Riesz MMD regularized f-divergences

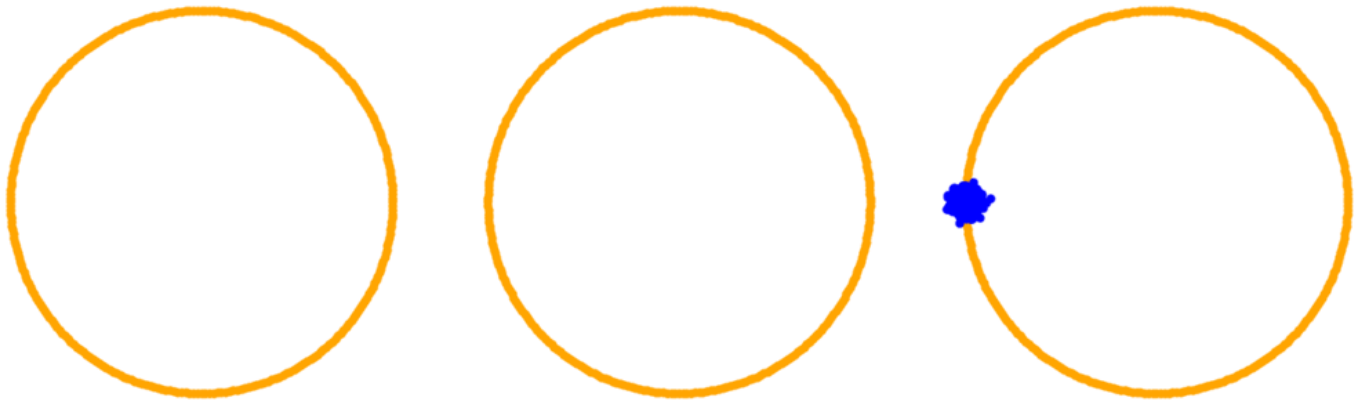
Nicolaj Rux

Technische Universität Berlin
Applied Mathematics

Wednesday 1st April, 2026



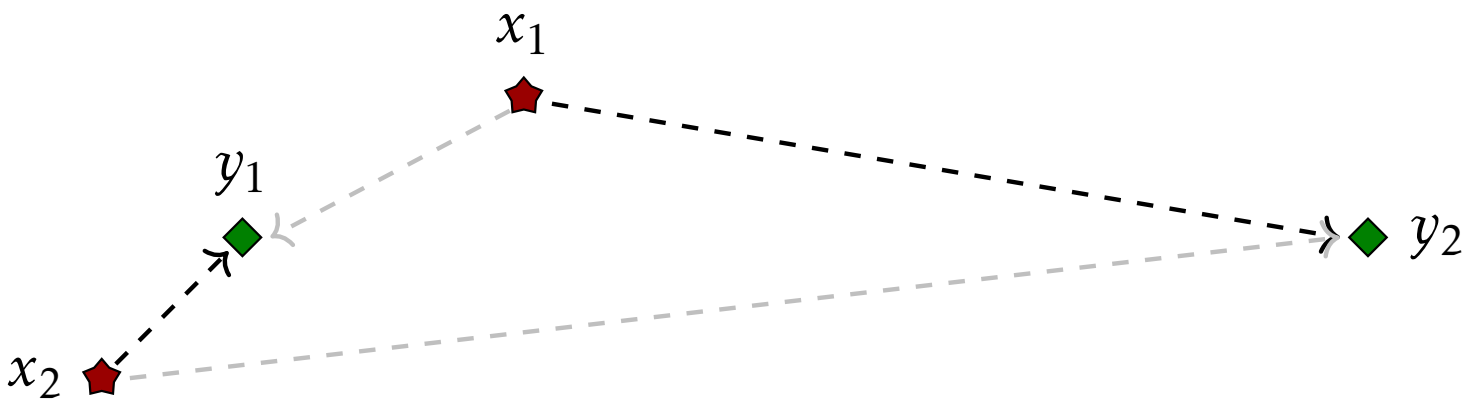
Motivation



- S. Neumayer, V. Stein, G. Steidl and N. Rux, Wasserstein Gradient Flows for Moreau Envelopes of f-Divergences in Reproducing Kernel Hilbert Spaces, https://arxiv.org/abs/2402.04613, https://github.com/ViktorAJStein/Regularized_f_Divergence_Particle_Flows/

Optimal Transport

"The Wasserstein distance W_2 quantifies the minimal effort to transport measure $\mu, \in \mathcal{P}^2(\mathbb{R})$ to $\nu \in \mathcal{P}^2(\mathbb{R})$, w.r.t $\|\cdot\|^2$. "



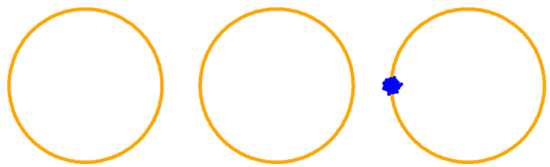
$$W_2(\mu,\nu)^2 = \min_{\gamma\in\Gamma(\mu,\nu)} \iint_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \mathrm{d}\gamma(x,y)$$

Wasserstein Gradient Flows

Definition 1

Given an energy functional $\mathcal{D}\colon (\mathcal{P}^2(\mathbb{R}^d), W_2) \to [0, \infty)$, *a Wasserstein gradient flow starting at* $\mu_0 \in \mathcal{P}^2(\mathbb{R}^d)$ *is a curve* $\omega\colon (0, \infty) \to \mathcal{P}^2(\mathbb{R}^d)$ *with* $\lim_{t \to 0} \omega(t) = \mu_0$, *following the steepest descent w.r.t* W_2.

$$\begin{cases} \omega'(t) = -\nabla \mathcal{D}(\omega(t)) \\ \omega(0) = \mu_0 \end{cases}$$



$$\mathcal{D}(\mu) := D(\mu \| \nu) \quad D(\mu \| \nu) = 0 \Leftrightarrow \mu = \nu$$

Reproducing Kernel Hilbert Spaces

Theorem 2 (Moore-Aronszajn)

*For every **symmetric, positive definite** kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, there exists a unique **RKHS** $(\mathcal{H}_K, \langle\cdot,\cdot\rangle_K)$, such that $K(x,\cdot) \in \mathcal{H}_K$ and*

$$h(x) = \langle h, K(x,\cdot)\rangle_K \quad \text{for all } h \in \mathcal{H}_K \text{ and } x \in \mathbb{R}^d.$$

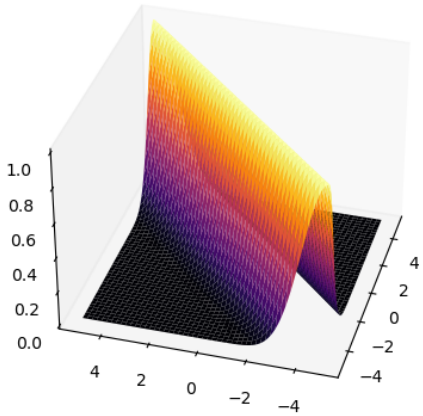
$$\text{SPD-Kernel } K \xleftarrow{\quad 1:1 \quad} \text{RKHS } \mathcal{H}_K$$

Definition 3 (Reproducing Kernel Hilbert Space)

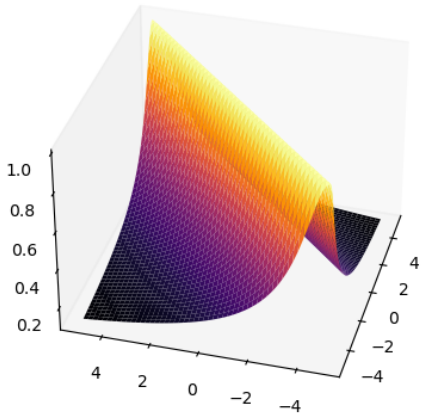
Hilbert space $\mathcal{H} \subseteq \{f : \mathbb{R}^d \to \mathbb{R}\}$ such that for all $x \in \mathbb{R}^d$, the evaluation maps $\hat{x}\colon \mathcal{H} \to \mathbb{R}, h \mapsto h(x)$ are continuous.

Examples of Kernels

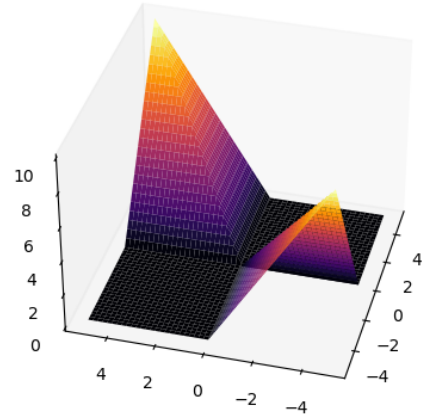
- Gaussian $x, y \mapsto \exp(-\frac{1}{2\sigma^2}\|x-y\|^2) \quad \in \mathcal{C}_0$
- Inverse Multi Quadric $x, y \mapsto (\sigma + \|x-y\|^2)^{-\frac{1}{2}} \quad \in \mathcal{C}_0$
- Negative Distance (Riesz)
  $x, y \mapsto -\|x-y\| + \|x\| + \|y\| + 1 \quad \notin \mathcal{C}_0$



Gaussian



IMQ



Negative Distance

Kernel Mean Embedding (KME)

- Isomorphism $\Phi_K : \mathcal{H}_K \to \mathcal{H}_K^*, \Phi_K(h) = \langle h, \cdot \rangle_K$
- Isomorphism $\Psi : \mathcal{M}(\mathbb{R}^d) \to \mathcal{C}_0(\mathbb{R}^d)^*, \Psi(\mu)(h) = \int h \mathrm{d}\mu$
- Bounded embedding $i_K : (\mathcal{H}_K, \|\cdot\|_K) \to (\mathcal{C}_0(\mathbb{R}^d), \|\cdot\|_\infty), h \mapsto h$

$$\begin{array}{ccc} \mathcal{C}_0(\mathbb{R}^d)^* & \xrightarrow{i_K^*} & \mathcal{H}_K^* \\ \uparrow \Psi & & \uparrow \Phi_K \\ \mathcal{M}(\mathbb{R}^d) & \xrightarrow{m_K} & \mathcal{H}_K \end{array}$$

$$m_K(\mu)(x) = \int_{\mathbb{R}^d} K(x, y) \mathrm{d}\mu(y)$$

$$\mathcal{M}(\mathbb{R}^d) \xrightarrow{m_K} \mathcal{H}_K \xrightarrow{\|\cdot\|_{\mathcal{H}_K}} [0, \infty)$$

Evolution of Negative Distance Kernel

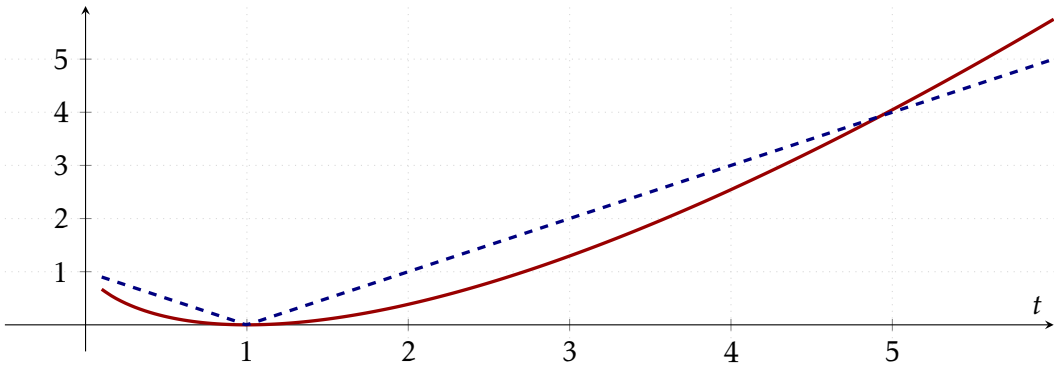
- $x, y \mapsto -\|x - y\|$ symmetric, but not positive definite.
- $R(x, y) = -\|x - y\| + \|\mathbf{x}\| + \|\mathbf{y}\|$ symmetric positive definite, but KME $m_R(\mathcal{M}(\mathbb{R}^d)) \not\subseteq \mathcal{H}_R$.
- Restrict KME to

$$\mathcal{M}_{1/2}(\mathbb{R}^d) := \left\{ \mu \in \mathcal{M}(\mathbb{R}^d) \colon \int_{\mathbb{R}^d} \|x\|^{\frac{1}{2}} \mathrm{d}|\mu|(x) < \infty \right\}$$

- KME: $\mathcal{M}_{1/2}(\mathbb{R}^d) \to \mathcal{H}_R$, but m_R **not** injective:

$$m_R(\delta_0)(x) = \int_{\mathbb{R}^d} K(x, y) \mathrm{d}\delta_0(y) = R(x, 0) = 0$$

- Kernel $x, y \mapsto -\|x - y\| + \|x\| + \|y\| + \mathbf{1}$ is symmetric positive definite and KME on $\mathcal{M}_{1/2}(\mathbb{R}^d)$ is injective.



- $\mathrm{KL}(t) = t\ln(t) - t + 1$
- $\chi(t) = |t-1|$

Definition 4 (f-Divergence)

Given two non negative measures $\mu, \nu \in \mathcal{M}(\mathbb{R}^d)$ there is a unique decomposition $\mu = \mu^c + \mu^s$, where $\mu^c \ll \nu$ and $\mu^s \perp \nu$.

$$D_f(\mu\|\nu) := \int_{\mathbb{R}^d} f \circ \frac{\mathrm{d}\mu^c}{\mathrm{d}\nu}\mathrm{d}\nu + f'_\infty \mu^s(\mathbb{R}^d).$$

MMD-regularised f-Divergence

- Entropy function f
- Target measure ν
- regularizer $\lambda > 0$
- Kernel K

Definition 5

$$D^\lambda := D^\lambda_{f,K,\nu}(\mu) := \inf_{\sigma \in \mathcal{M}_{1/2}{}^+(\mathbb{R}^d)} \left\{ D_f(\sigma \| \nu) + \frac{1}{2\lambda} \|\mu - \sigma\|_K^2 \right\}$$

- $\|\cdot\|_K = \|\cdot\|_{\mathcal{H}_K} \circ m_K$

Extended f-Divergence

- $E\colon \mathcal{H}_K \to [0,\infty]$ given

$$E(h) := \begin{cases} D_f(\mu\|\nu) & \text{if } \exists \mu \in \mathcal{M}_{1/2}{}^{+}(\mathbb{R}^d)\colon\ m_K(\mu) = h \\ \infty & \text{else.} \end{cases}$$

- $E \circ m_K = D_f(\cdot\|\nu)$

Definition 6 (Moreau Envelope)

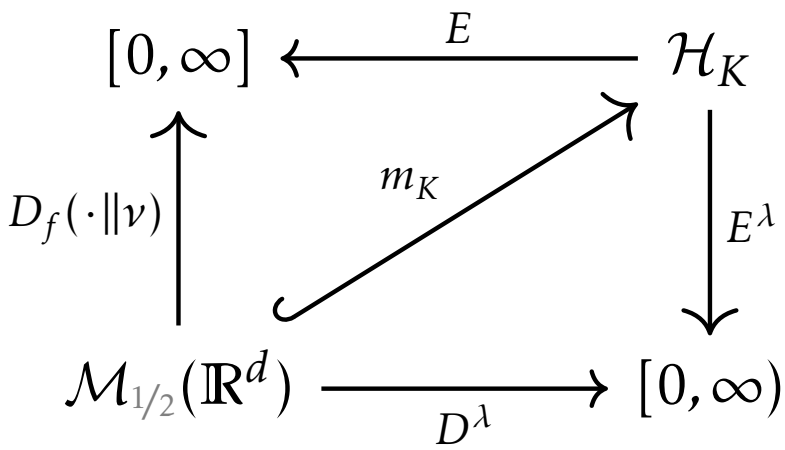
For $E \in \Gamma_0(\mathcal{H}_K)$ the Moreau Envelope is defined as

$$E^\lambda(h) := \inf_{g\in\mathcal{H}_K} \left\{ E(g) + \frac{1}{2\lambda}\|h-g\|_K^2 \right\} \in \Gamma_0(\mathcal{H}_K).$$

From Measures to RKHS

Theorem 7

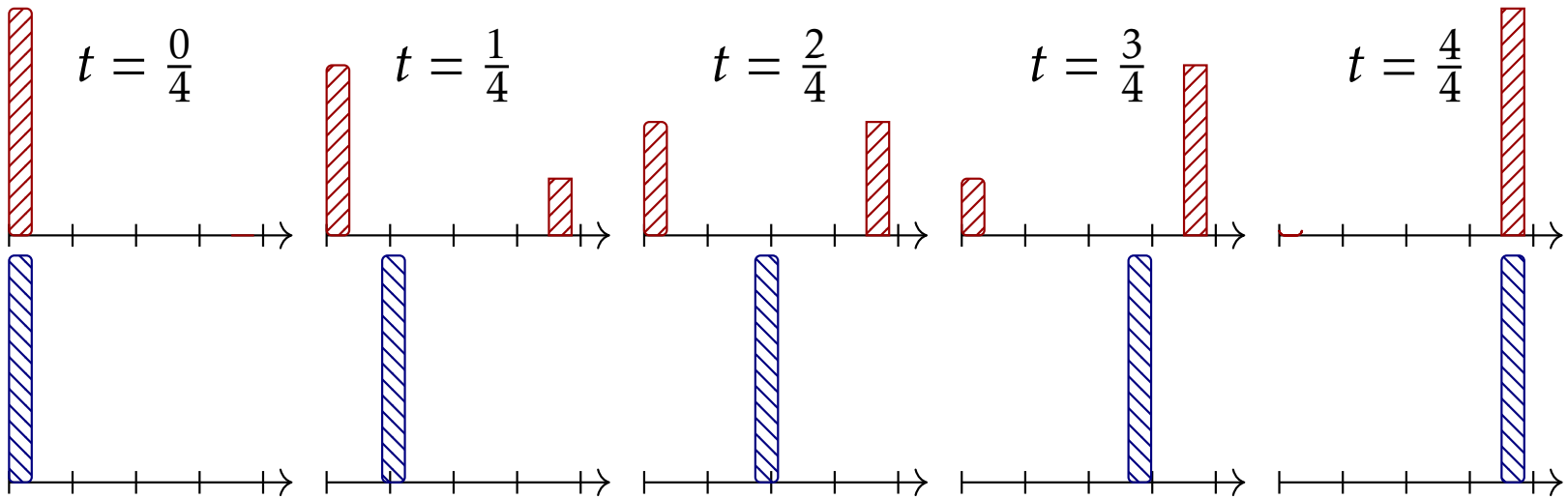
For symmetric, positive definite, characteristic C_0-kernels, the extended f-Divergence E is in $\Gamma_0(\mathcal{H}_K)$.



- For Negative Distance kernel E is in $\Gamma_0(\mathcal{H}_R)$ under certain growth conditions on f and moment constrains on ν.

Theorem 8 (Existence of Wasserstein Gradient Flows)

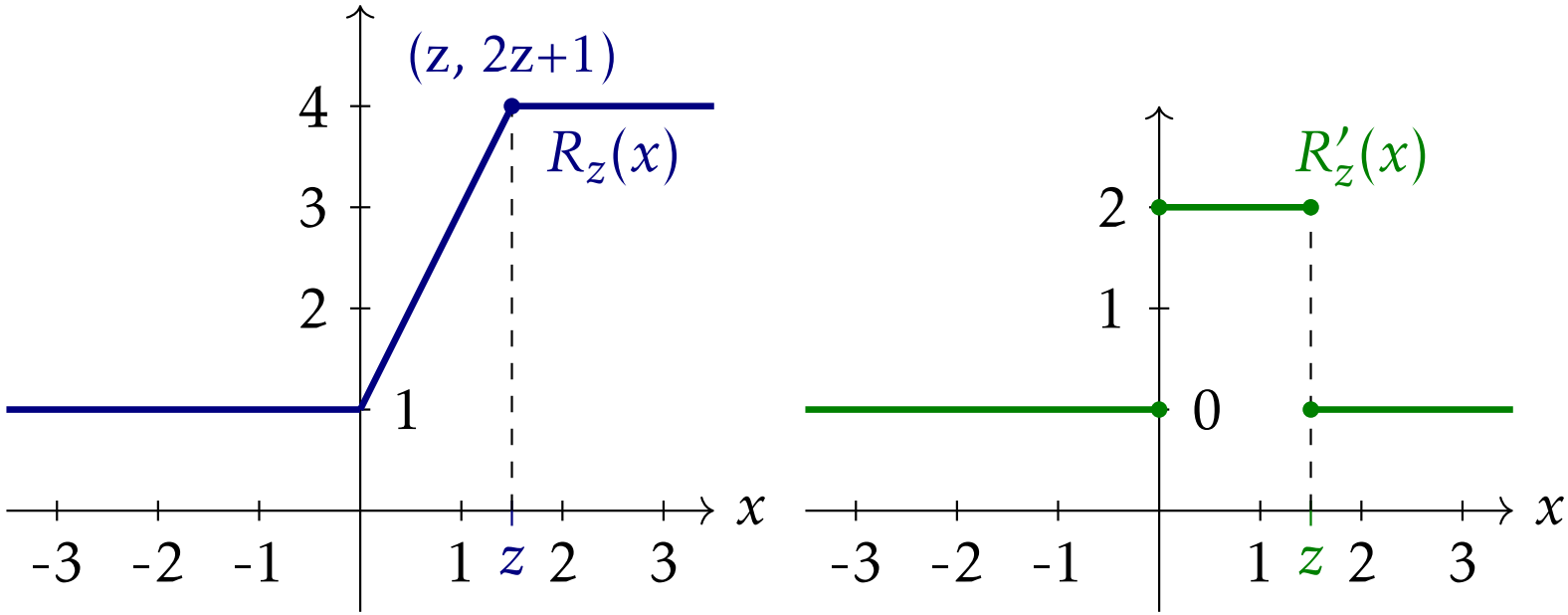
Let $\mathcal{D} : (\mathcal{P}_2(\mathbb{R}^d), W_2) \to [0, \infty]$ be proper, ***lower semi-continuous*** *and M-****convex along generalized geodesics****. Given $\mu_0 \in \mathrm{dom}(\mathcal{D})$, there is a unique Wasserstein gradient flow ω of $\mathcal{D}$ starting at μ_0.*



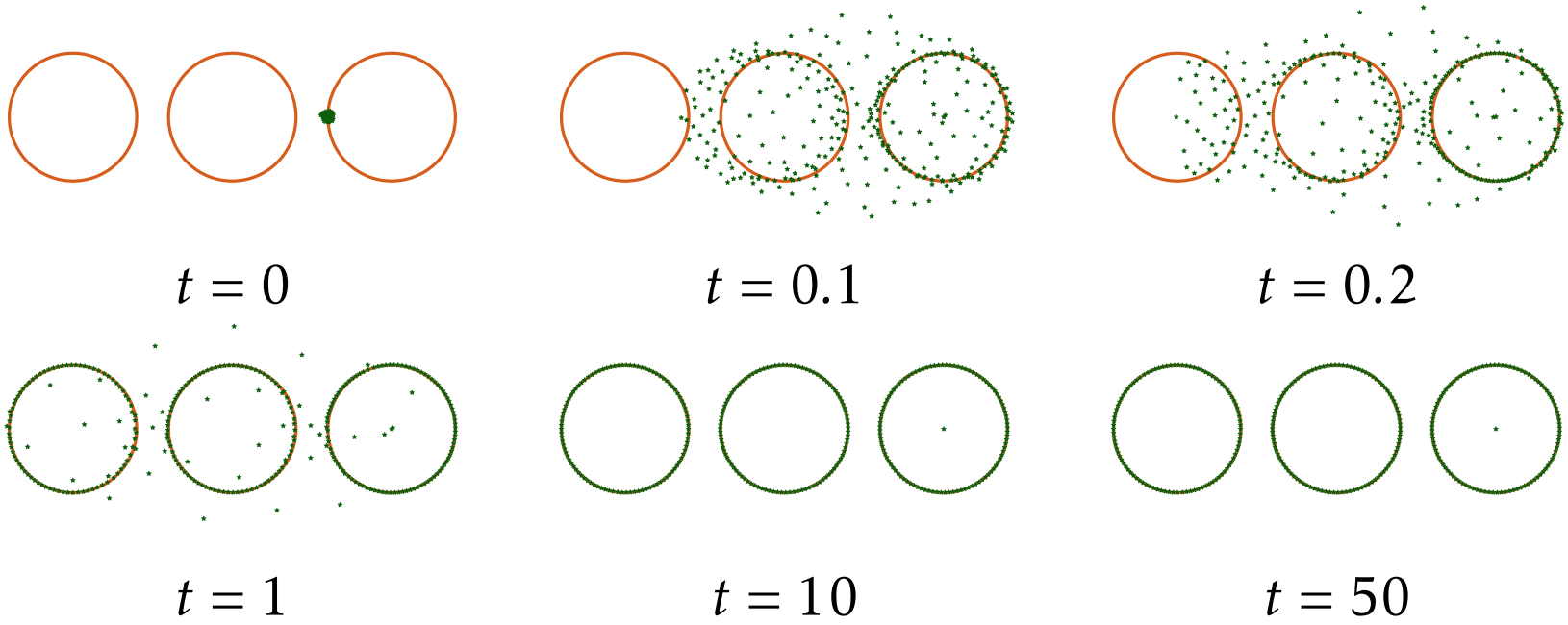
$$\tilde{\gamma}_t = (1-t)\delta_x + t\delta_y \quad \text{and} \quad \gamma_t = \delta_{(1-t)x+ty}$$

[0] Ambrosio Et al., Gradient Flows in Metric Spaces and in the Space of Probability Measures, [1, Thm. 11.2.1 & Eq. 11.2.1b]

Negative Distance Kernel

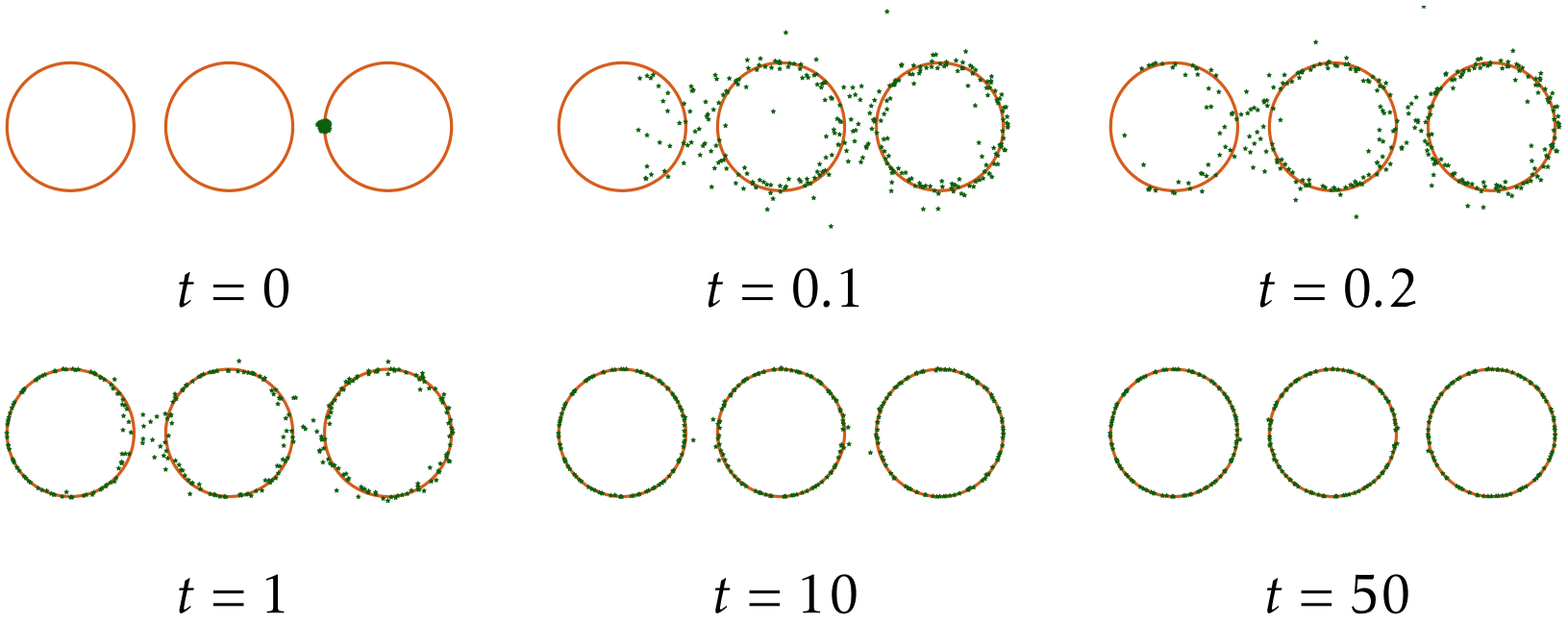


Negative Distance Kernel



[0]Tsallis $\alpha = 3$, modified negative distance, $\lambda = 0,01$, $\tau = 0.001$, $\nu - 900$, $\mu - 300$ particles.

Inverse Multi Quadric



[0]Tsallis $\alpha = 3$, inverse multi quadric, $\lambda = 0,01$, $\tau = 0.001$, $\sigma = 0.1$, $\nu - 900$, $\mu - 300$ particles.

Thank you for your attention :)

I am happy to take any questions!

References I

L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, 2nd ed., ser. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2008.

C. Villani, *Optimal Transport*, ser. Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 2009, vol. 338, old and New. [Online]. Available: https://doi.org/10.1007/978-3-540-71050-9

T. Modeste and C. Dombry, "Characterization of translation invariant MMD on R d and connections with Wasserstein distances," *HAL*, Jul. 2023, working paper or preprint. [Online]. Available: https://hal.science/hal-03855093

S. Neumayer, V. Stein, G. Steidl, and N. Rux, "Wasserstein gradient flows for Moreau envelopes of f-divergences in reproducing kernel Hilbert spaces," *arXiv preprint arXiv:2402.04613*, 2024. [Online]. Available: https://arxiv.org/abs/2402.04613

P. Glaser, M. Arbel, and A. Gretton, "KALE Flow: A Relaxed KL Gradient Flow for Probabilities with Disjoint Support," in *NeurIPS 2021 -Thirty-Fifth Annual Conference on Neural Information Processing Systems*, ser. Advances in Neural Information Processing Systems, Online, France, Dec. 2021, pp. 1–29. [Online]. Available: https://hal.science/hal-03455473