# Smoothed Distance Kernels for MMDs and Applications in Wasserstein Gradient Flows

Nicolaj Rux[1,2*] Michael Quellmalz[2] Gabriele Steidl[2]
nicolaj.rux@math.tu-chemnitz.de quellmalz@math.tu-berlin.de steidl@math.tu-berlin.de

April 14, 2025

[1]TU Chemnitz
Faculty of Mathematics
Reichenhainer Straße 39
D-09111 Chemnitz, Germany

[2]TU Berlin
Institute of Mathematics
Straße des 17. Juni 136
D-10623 Berlin, Germany

**Abstract**

Negative distance kernels $K(x, y) := -\|x - y\|$ were used in the definition of maximum mean discrepancies (MMDs) in statistics and lead to favorable numerical results in various applications. In particular, so-called slicing techniques for handling high-dimensional kernel summations profit from the simple parameter-free structure of the distance kernel. However, due to its non-smoothness in $x = y$, most of the classical theoretical results, e.g. on Wasserstein gradient flows of the corresponding MMD functional do not longer hold true. In this paper, we propose a new kernel which keeps the favorable properties of the negative distance kernel as being conditionally positive definite of order one with a nearly linear increase towards infinity and a simple slicing structure, but is Lipschitz differentiable now. Our construction is based on a simple 1D smoothing procedure of the absolute value function followed by a Riemann-Liouville fractional integral transform. Numerical results demonstrate that the new kernel performs similarly well as the negative distance kernel in gradient descent methods, but now with theoretical guarantees.

*Keywords:* Negative distance kernel, Maximum Mean Discrepancy, Conditionally positive definite functions, Wasserstein gradient flows, Optimal transport, Fourier transform

*Mathematics Subject Classification:* 46E22 49Q22 42B10 44A12 65D12

## Declarations

**Acknowledgements:** We thank Sebastian Neumayer for his valuable suggestions, especially in the numerical aspects of this work.

**Conflict of Interest:** On behalf of all authors, the corresponding author states that there is no conflict of interest.

**Competing Interests:** The authors have no competing interests to declare that are relevant to the content of this article.

**Funding Information:** MQs research was funded the German Research Foundation (DFG): STE 571/19-1, project number 495365311, within the Austrian Science Fund (FWF) SFB 10.55776/F68 "Tomography Across the Scales". GS acknowledges the funding support by the DFG within the Excellence Cluster MATH+. NR gratefully acknowledges the funding support from the European Union and the Free State of Saxony (ESF).

**Author contribution:** All three authors have contributed equally to the manuscript.

**Data Availability Statement:** Data availability is not applicable to this article as no new data were created or analyzed in this study.

**Research Involving Human and /or Animals:** Not applicable because no research involving humans or animals has been conducted for this article.

**Informed Consent:** Not applicable because no research involving humans has been conducted for this article.

# 1 Introduction

Symmetric, positive definite functions have been playing a role in kernel-based learning for a long time [14, 52]. While mostly Gaussian kernels are used, recently, also the conditionally positive definite negative distance kernel $K(x,y) := -\|x-y\|$ has attained interest, e.g. in statistics [50], image dithering/halftoning [16, 21, 33], sampling [39] and generative modeling [25, 30]. Indeed, more general Riesz kernels $K(x,y) := -\|x-y\|^s$, $s \in [0,2)$, were examined in optimization equilibrium problems, see, e.g. [12, 24, 18]. Let us also mention that gradient flows with respect to the Coulomb kernel $K(x,y) := \|x-y\|^{2-d}$ were quite recently examined in [7], see also [9], and $K(x,y) := \|x-y\|^{-1}$ was applied in image halftoning in [48]. For interesting translation invariance properties of MMDs and connections with Wasserstein distances, we refer to [37].

Depending on the kernel, the maximum mean discrepancy (MMD) between two measures can be defined as the sum of an interaction energy and potential energy. Fixing one of the measures, in generative learning called target measure, Wasserstein gradient flows of the corresponding functional on the Wasserstein-2 space starting in a simple (latent) measure can be applied to sample from that target distribution. While such gradient flows together with numerical forward and backward schemes for their computation are well understood for Lipschitz differentiable kernels, see, e.g. [2, 3], the convergence behavior of forward steepest descent [27] and Euler backward (JKO) schemes [31] are not clear for the negative distance kernel due to its nondifferentiability in $x = y$. One exception is the one-dimensional case, where the MMD functional becomes, in contrast to higher dimensions, (geodesically) $\lambda$-convex, see [15] and the references therein.

Gradient flows of the MMD functional or just the interaction energy with the negative distance kernel or Riesz kernels show a mathematically richer structure than those for smooth kernels and were the object of numerous examinations, see e.g. [8, 10, 11]. In particular, singular measures can become absolutely continuous along the flow curve and conversely [4, 27], so that these flows are no longer just particle flows when starting in an empirical measure. Finally, let us mention flows in the MMD dissipation geometry [58] which differ from the setting considered in this paper.

If applied in a straightforward way, MMD flows suffer from high computational costs in large scale computations, since each gradient step requires the computation of kernel sums (or their derivatives) with a large number of summands. For positive definite kernels, a remedy is to apply random Fourier feature techniques [45] based on Bochner's theorem. Unfortunately, the negative distance kernel does not fit into the setting of Bochner's theorem, but here efficient so-called slicing techniques, which project the high-dimensional problem in a bunch of one-dimensional ones, can be used [26, 28]. For an interesting quite general fast summation approach using deep learning, we refer to [29].

In this paper, we construct a smoothed negative distance kernel such that its MMD functional fulfills the classical assumptions on its Wasserstein gradient flow and ensures in particular that empirical measures evolve as particle flows with proofed convergence of Euler forward and backward schemes. On the other hand, these kernels are still conditionally positive definite of order one and behave in applications similarly as the negative distance kernel, but now with theoretical convergence guarantees.

Our paper is organized as follows: in Section 2, we provide some notation and recall several results on (generalized) Fourier transforms. For readers not familiar with the topic, more material on Fourier transforms of tempered distributions including enlightening examples are added in Appendix A. The next three sections contain the steps for defining our smoothed distance kernels: Section 3 starts with appropriate smoothings of the absolute value function in $\mathbb{R}^1$. Although not directly relevant for our construction, a relation to the often applied Huber function is addressed in Appendix B. Then, Section 4 establishes smoothed Euclidean norm functions in $\mathbb{R}^d$, $d \geq 2$ based on Riemann-Liouville integral transforms, which finally lead to our smoothed kernels in Section 5. Using these kernels, we define reproducing kernel Hilbert spaces and MMDs based on kernel mean embeddings in Section 6. Wasserstein gradient flows of our MMDs are considered in Section 7. We add considerations on the geodesic convexity of the MMDs in Appendix D. Finally, we demonstrate the very good performance of Wasserstein gradient flows of the MMD with our new kernel by numerical examples in Section 8.

All proofs, which are not indicated to be taken directly from the literature, are given in Appendix C.

## 2 Preliminaries

Let $\mathcal{C}_b(\mathbb{R}^d)$ denote the space of continuous bounded functions $f\colon \mathbb{R}^d \to \mathbb{C}$ with norm

$$\|f\|_\infty := \sup_{x\in\mathbb{R}^d} |f(x)|,$$

$\mathcal{C}_0(\mathbb{R}^d)$ the subspace of functions $f\colon \mathbb{R}^d \to \mathbb{C}$ vanishing as $\|x\| \to \infty$, $\mathcal{C}_c(\mathbb{R}^d)$ the subspace of continuous functions with compact support, $\mathcal{C}^n(\mathbb{R}^d)$, $n \in \mathbb{N}$ the space of $n$-times continuously differentiable functions and $\mathcal{C}_c^n(\mathbb{R}^d)$ the space of $n$-times continuously differentiable functions with compact support. For $1 \le p \le \infty$, let $L^p(\mathbb{R}^d)$ be the Banach space of all (equivalence class of) measurable functions $f\colon \mathbb{R}^d \to \mathbb{C}$ with finite norm $\|f\|_{L^p}$ and $L^p_{\mathrm{loc}}(\mathbb{R}^d)$ the corresponding locally integrable functions.

Further, we denote by $\mathcal{S}(\mathbb{R}^d)$ the space of complex-valued *Schwartz functions*. The *Fourier transform* $\mathcal{F}\colon \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ is the bijective mapping defined by

$$\hat{\varphi}(\omega) = \mathcal{F}[\varphi](\omega) := \int_{\mathbb{R}^d} \mathrm{e}^{-2\pi \mathrm{i}\langle x,\omega\rangle} \varphi(x)\,\mathrm{d}x, \qquad \omega \in \mathbb{R}^d. \tag{1}$$

The *Fourier transform* can be extended as a mapping $\mathcal{F}\colon L^1(\mathbb{R}^d) \to \mathcal{C}_0(\mathbb{R}^d)$. The *convolution function* $f * g$ of two functions $f, g$ on $\mathbb{R}^d$ is defined, if it exists, by

$$(f*g)(x) := \int_{y\in\mathbb{R}^d} f(x-y)g(y)\ \mathrm{d}y = \int_{y\in\mathbb{R}^d} f(y)g(x-y)\ \mathrm{d}y, \qquad x \in \mathbb{R}^d.$$

In particular, if $f, g \in L^1(\mathbb{R}^d)$, then $f * g$ is defined almost everywhere and it holds the Fourier convolution theorem

$$\mathcal{F}[f*g] = \hat{f}\,\hat{g}.$$

For $r \in \mathbb{N}_0$, we define the space

$$\mathcal{S}_r(\mathbb{R}^d) := \{\varphi \in \mathcal{S}(\mathbb{R}^d) : \varphi(x) \in \mathcal{O}(\|x\|^r) \ \text{ as } \ \|x\| \to 0\}.$$

A measurable function $\hat{f} \in L^2_{\mathrm{loc}}(\mathbb{R}^d \setminus \{0\})$ is called *generalized Fourier transform* of a slowly increasing function $f \in \mathcal{C}(\mathbb{R}^d)$, if there exists an integer $r \in \mathbb{N}_0$ such that

$$\int_{\mathbb{R}^d} f(x)\hat{\varphi}(x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \hat{f}(\omega)\varphi(\omega)\,\mathrm{d}\omega \quad \text{for all} \quad \varphi \in \mathcal{S}_{2r}(\mathbb{R}^d), \tag{2}$$

see [57, Def. 8.9]. If $f$ fulfills (2) for some $r \in \mathbb{N}_0$, then it fulfills this relation also for all integers larger than $r$. In particular, if $f \in \mathcal{S}(\mathbb{R}^d)$, then (2) holds for all $r \ge 0$. The smallest $r \in \mathbb{N}_0$ such that (2) is fulfilled is called *order of the generalized Fourier transform*. We have that $\hat{f}$ is uniquely determined. The generalized Fourier transform differs from the Fourier transform of so-called tempered distributions, in particular of continuous, slowly increasing functions, but coincides with it if restricted to test functions in $\mathcal{S}_{2r}(\mathbb{R}^d)$. This is briefly explained in Appendix A.

In this paper, we are mainly concerned with powers of the Euclidean norm.

**Theorem 2.1** ([57, Thm. 8.16]). *The function* $f(x) := \|x\|^\beta$, $x \in \mathbb{R}^d$, *with* $\beta > 0$, $\beta \notin 2\mathbb{N}$ *has the generalized Fourier transform*

$$\hat{f}(\omega) = \frac{\Gamma(\frac{d+\beta}{2})}{\pi^{\beta+\frac{d}{2}}\Gamma(-\frac{\beta}{2})}\|\omega\|^{-\beta-d}, \qquad \omega \in \mathbb{R}^d$$

*of order* $r = \lceil \frac{\beta}{2} \rceil$. *In particular, we have for* $\mathrm{abs}(x) = |x|$, $x \in \mathbb{R}$, *that*

$$\widehat{\mathrm{abs}}(\omega) = -\frac{1}{2\pi^2\omega^2}, \qquad \omega \in \mathbb{R}. \tag{3}$$

For the generalized Fourier transform, we have the following convolution property.

**Proposition 2.2.** *Let* $f \in C(\mathbb{R}^d)$ *be a slowly increasing function with generalized Fourier transform* $\hat{f}$ *of order* $r$ *and* $u \in C_c(\mathbb{R}^d)$. *Then the convolution* $f * u \in C(\mathbb{R}^d)$ *is slowly increasing and has a generalized Fourier transform of order* $r$ *which fulfills* $\mathcal{F}[f * u] = \hat{f}\,\hat{u}$.

Further, the notation of conditionally positive definiteness will be central in our paper. A continuous, even function $f\colon \mathbb{R}^d \to \mathbb{C}$ is *conditionally positive definite of order* $r \in \mathbb{N}_0$, if for all $N \in \mathbb{N}$, all $x_1, \dots, x_N \in \mathbb{R}^d$, and all $a \in \mathbb{C}^N \setminus \{0\}$ satisfying

$$\sum_{j=1}^N a_j p(x_j) = 0$$

for all $d$-dimensional polynomials $p$ of degree $\leq r - 1$, we have

$$\sum_{j,k=1}^N a_j \bar{a}_k f(x_j - x_k) \geq 0,$$

see [36, 54]. We denote the space of conditionally positive definite functions of order $r$ by $\mathrm{CP}_r(\mathbb{R}^d)$. In particular, $-\|\cdot\|^\beta \in \mathrm{CP}_1(\mathbb{R}^d)$, $\beta \in (0, 2)$. If $r = 0$, we just speak about positive definite functions.

Bochner's theorem characterizes positive definite functions as Fourier transform of positive measures, see Theorem A.3 in Appendix A. There are different ways to modify Bochner's theorem for conditionally positive definite functions. We will use the following one [57, Thm. 8.12].

**Theorem 2.3** (Bochner's Theorem for Generalized Fourier Transform). *Let* $f\colon \mathbb{R}^d \to \mathbb{C}$ *be continuous, slowly increasing, and possess a generalized Fourier transform* $\hat{f}$ *of order* $r$, *which is continuous on* $\mathbb{R}^d \setminus \{0\}$. *Then* $f$ *is conditionally positive definite of order* $r$ *if and only if* $\hat{f}$ *is nonnegative.*

# 3 Smoothed Absolute Value Function

In this section, we propose to embellish $\operatorname{abs}(x) = |x|$ by convolving it with functions from the set

$$\mathcal{U}^n(\mathbb{R}) := \left\{ u \in \mathcal{C}_c^n(\mathbb{R}) : u, \hat{u} \geq 0,\, u \text{ even}, \int_{\mathbb{R}} u \,\mathrm{d}x = 1 \right\}, \quad n \in \mathbb{N}_0. \tag{4}$$

These functions have the following nice properties.

**Proposition 3.1.** *Let* $u \in \mathcal{U}^n(\mathbb{R})$ *and* $u_\varepsilon(x) := \frac{1}{\varepsilon} u\left(\frac{x}{\varepsilon}\right)$ *for* $\varepsilon > 0$. *Then* $f := \operatorname{abs} * u$ *fulfills:*

*i)* $f > 0$ *and* $f$ *is even,*

*ii)* $f(x) = \operatorname{abs}(x)$ *for* $|x| \geq \operatorname{diam}(\operatorname{supp}(u))/2$,

*iii)* $f'' = 2u$ *so that* $f$ *is convex and* $f \in \mathcal{C}^{n+2}(\mathbb{R})$,

*iv)* $-f$ *is conditionally positive definite of order* $r = 1$, *but not positive definite,*

*v)* $(\operatorname{abs} * u_\varepsilon)(x) = \varepsilon f\left(\frac{x}{\varepsilon}\right)$, $\quad (\operatorname{abs} * u_\varepsilon)'(x) = f'\left(\frac{x}{\varepsilon}\right)$, $\quad (\operatorname{abs} * u_\varepsilon)''(x) = \frac{2}{\varepsilon} u\left(\frac{x}{\varepsilon}\right)$,

*vi)* $\operatorname{abs} * u_\varepsilon \to \operatorname{abs}$ *uniformly as* $\varepsilon \to 0$.

The most important functions $u \in \mathcal{U}^n(\mathbb{R})$ in our numerical part will be centered cardinal $B$-splines. The *centered cardinal B-spline of order* $m \in \mathbb{N}$ is recursively defined by

$$M_1 := \mathbf{1}_{[-\frac{1}{2}, \frac{1}{2}]}, \quad M_m := M_1 * M_{m-1}, \quad m = 2, 3, \ldots$$

$B$-splines have many useful properties, see [35, 49].

**Proposition 3.2.** *For the centered cardinal B-splines with* $m \in \mathbb{N}$, *the following holds true:*

*i)* $M_m \geq 0$ *and* $\int_{\mathbb{R}} M_m(x) \,\mathrm{d}x = 1$,

*ii)* $\operatorname{supp} M_m = \left[-\frac{m}{2}, \frac{m}{2}\right]$ *and* $M_m$ *is even,*

*iii)* $M_m \in \mathcal{C}^{m-2}(\mathbb{R})$, $m \geq 2$,

*iv)* $\widehat{M}_m(\omega) = \operatorname{sinc}^m(\omega)$, *where* $\operatorname{sinc}(\omega) := \frac{\sin(\pi\omega)}{\pi\omega}$. *This is a nonnegative function exactly for even* $m \in \mathbb{N}$.

*v)* *For* $m \geq 2$, *we have*

$$M_m(x) = \frac{1}{(m-1)!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} \left(x - k + \frac{m}{2}\right)_+^{m-1}, \tag{5}$$

*where* $x_+ := \max(x, 0)$, *and*

$$\begin{aligned} M_m(0) &= \frac{2}{\pi} \int_0^\infty \left(\frac{\sin(x)}{x}\right)^m \mathrm{d}x = \frac{m}{2^{m-1}} \sum_{k=0}^{\lfloor \frac{m}{2} \rfloor} \frac{(-1)^k (m-2k)^{m-1}}{m!\,(m-k)!} \\ &= \sqrt{\frac{6}{\pi m}} \left(1 + \mathcal{O}(m^{-1})\right). \end{aligned}$$

*vi) Clearly, it holds $M_{2m} \in \mathcal{U}^{2m-2}(\mathbb{R})$ for all $m \in \mathbb{N}$.*

The convolution of abs with the centered cardinal $B$-splines is given in the following proposition.

**Corollary 3.3.** *For $f := \mathrm{abs} * M_m$, it holds*

$$f(x) = \frac{2}{(m+1)!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} \left(x - k + \frac{m}{2}\right)_+^{m+1} - x.$$

Here are two examples.

**Example 3.4.** *From*

$$M_2 = \begin{cases} 1 - |x|, & |x| \le 1, \\ 0, & \text{otherwise}, \end{cases} \qquad M_4 = \frac{1}{6} \begin{cases} 3|x|^3 - 6x^2 + 4, & |x| \le 1, \\ (2 - |x|)^3, & 1 < |x| \le 2, \\ 0, & \text{otherwise}, \end{cases}$$

*we get*

$$(\mathrm{abs} * M_2)(x) = \begin{cases} \frac{1}{3}(-|x|^3 + 3|x|^2 + 1), & |x| \le 1, \\ |x|, & \text{otherwise}, \end{cases} \tag{6}$$

*and*

$$(\mathrm{abs} * M_4)\,(x) = \begin{cases} \frac{1}{20}|x|^5 - \frac{1}{6}x^4 + \frac{2}{3}x^2 + \frac{7}{15}, & 0 \le |x| < 1, \\ \frac{1}{60}(2 - |x|)^5 + |x|, & 1 \le |x| < 2, \\ |x|, & \text{otherwise}. \end{cases}$$

*For a plot of* abs $*M_2$ *with its first and second order derivatives see Figure 1.*

Asking for smoothed absolute value functions, the Huber function may first come into one's mind. Unfortunately, by the following corollary, the negative Huber function is not conditionally positive definite.

**Corollary 3.5.** *The Huber function*

$$f(x) := \begin{cases} \frac{1}{2}x^2, & |x| \le \lambda, \\ \lambda(|x| - \frac{\lambda}{2}), & \text{otherwise}, \end{cases}$$

*can be rewritten as $f = \lambda\,(\mathrm{abs} * M_{1,2\lambda}) - \frac{\lambda^2}{2}$ and has the generalized Fourier transform*

$$\hat{f}(\omega) = -\frac{\lambda}{2\pi^2\omega^2}\,\mathrm{sinc}(2\lambda\omega),$$

*which takes positive and negative values, so that $-f$ is not conditionally positive definite.*

The proof follows from formula (30) in Appendix B. The Huber function is the so-called Moreau envelope of the absolute value function. Moreau envelopes play an important role in convex analysis. Appendix B contains more results on the relation of abs $*M_m$ to Moreau envelopes, which are interesting on their own.

## 4 Smoothed Euclidean Norm

Our aim is to approximate the Euclidean norm on $\mathbb{R}^d$ by a function which on the one hand keeps its desirable properties, in particular radial symmetry, simple computation and conditional positive definiteness of order 1, and on the other hand gives rise to Lipschitz differentiable kernels in the next section. First ideas could be the following two:

- Convolve the Euclidean norm in $\mathbb{R}^d$ with some smooth filter. Unfortunately, this is numerically expensive in high dimensions.
- Use $f(\|\cdot\|)$ with $f = \mathrm{abs} * u$ and $u \in \mathcal{U}^n(\mathbb{R})$. Unfortunately, this function is in general not conditionally positive definite, as the following lemma shows.

**Lemma 4.1.** *For $f = \mathrm{abs} * M_2$, it holds that $-f(\|\cdot\|) \notin \mathrm{CP}_r(\mathbb{R}^d)$ for any $d \geq 2$ and $r \in \mathbb{N}$.*

Since the above approaches do not provide the desired functions, we propose to use the Riemann-Liouville fractional integral transform, which we consider next.

### 4.1 Riemann-Liouville Fractional Integral and Slicing in $\mathbb{R}^d$

For $d \in \mathbb{N}$, $d \geq 2$, the *Riemann-Liouville fractional integral* $\mathcal{I}_d \colon L^\infty_{\mathrm{loc}}(\mathbb{R}) \to C^n(\mathbb{R})$, $n := \lfloor \frac{(d-2)}{2} \rfloor$ is defined by

$$F(s) = \mathcal{I}_d[f](s) := c_d \int_0^1 f(ts)(1-t^2)^{\frac{d-3}{2}} \, \mathrm{d}t \quad \text{for all} \quad s \in \mathbb{R}, \tag{7}$$

where $c_d := \frac{2w_{d-2}}{w_{d-1}}$ and $w_{d-1} := \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$ denotes the surface area of the sphere $\mathbb{S}^{d-1}$. For $d = 2$, the term $(1-t)^{\frac{d-3}{2}}$ is not bounded, but integrabable, so that we require $f$ to be locally bounded in order for (7) to exist. For $d \geq 3$, the term $(1-t^2)^{\frac{d-3}{2}}$ is bounded and we can define $\mathcal{I}_d$ on $L^1_{\mathrm{loc}}(\mathbb{R})$.

Our approach is motivated by the slicing techniques for fast kernel summation in [26, 28]. In particular, the Riemann-Liouville fractional integral has the following useful property, which relates a high-dimensional radial function to a function on one-dimensional projections of its inputs, see [46].

**Theorem 4.2.** *Let $d \in \mathbb{N}$, $d \geq 2$ and $f \in L^\infty_{\mathrm{loc}}(\mathbb{R})$ be even. Then the even function $F \colon \mathbb{R} \to \mathbb{R}$ defined by the Riemann-Liouville fractional integral* (7) *fulfills the projection/slicing condition*

$$F(\|x\|) = \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} f(\langle \xi, x \rangle) \, \mathrm{d}x = \mathbb{E}_{\xi \sim \mathcal{U}_{\mathbb{S}^{d-1}}} \left[ f\left( \langle x, \xi \rangle \right) \right], \tag{8}$$

*where $\mathcal{U}_{\mathbb{S}^{d-1}}$ denotes the uniform distribution on the sphere. Further, if $f$ is positive definite, then $F(\|\cdot\|)$ is also positive definite for all $d \geq 2$. Conversely, if $F(\|\cdot\|)$ is positive definite for some $d \geq 2$, then there exists an even positive definite function $f$ on $\mathbb{R}$ such that* (7) *is fulfilled.*

The slicing formula (8) is a special case of the adjoint Radon transform, see [46]. The following two propositions extend the last property of Theorem 4.2 to conditionally positive functions.

**Proposition 4.3.** *Let $d \in \mathbb{N}$, $d \geq 2$ and $f \in L^\infty_{\mathrm{loc}}(\mathbb{R})$ be even. Further, let $f \in \mathrm{CP}_r(\mathbb{R})$ for $r \in \mathbb{N}_0$ and $F = \mathcal{I}_d[f]$. Then $F(\|\cdot\|) \in \mathrm{CP}_r(\mathbb{R}^d)$.*

**Proposition 4.4.** *Let $d \geq 3$ and let the $\lfloor \frac{d}{2} \rfloor$-th derivative of $F \in \mathcal{C}^{\lfloor \frac{d}{2} \rfloor}([0,\infty))$ be slowly increasing. Moreover, assume that $F(\|\cdot\|) \in \mathrm{CP}_r(\mathbb{R}^d)$ has a generalized Fourier transform $\rho(\|\cdot\|) \in \mathcal{C}(\mathbb{R}^d \setminus \{0\})$. Then the function $f \in \mathrm{CP}_r(\mathbb{R})$ with generalized Fourier transform*

$$\hat{f} \in \mathcal{C}(\mathbb{R} \setminus \{0\}), \qquad \hat{f}(\omega) = \frac{w_{d-1}}{2}\rho(\omega)|\omega|^{d-1},$$

*fulfills* (7).

## 4.2 Riemann-Liouville Fractional Integral of Smoothed Absolute Value

Next, we are interested in the Riemann-Liouville fractional integral of the smoothed absolute value function $f := \mathrm{abs} * u$, $u \in \mathcal{U}^n(\mathbb{R})$. First of all, the absolute value function is an eigenfunction of $\mathcal{I}_d$, see, e.g. [28].

**Lemma 4.5.** *The functions $\mathrm{abs}^\beta$, $\beta > -1$ are eigenfunctions of $\mathcal{I}_d$ with eigenvalues $\frac{\Gamma(\frac{d}{2})\Gamma(\frac{\beta+1}{2})}{\sqrt{\pi}\Gamma(\frac{d+\beta}{2})}$.*

The Riemann-Liouville fractional integral of $\mathrm{abs} * u$ has the following properties.

**Proposition 4.6.** *Let $n, d \in \mathbb{N}$ with $d \geq 2$ and $u \in \mathcal{U}^n(\mathbb{R})$. Then the function $F := \mathcal{I}_d[\mathrm{abs} * u]$ is even, convex, positive and $(n+2)$-times continuously differentiable. Further, it satisfies for $s \to \infty$ the relation*

$$F(s) = C_d\,|s| + \mathcal{O}\left(\frac{1}{s}\right), \quad C_d := \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d+1}{2})}.$$

*In particular, $F - C_d\,\mathrm{abs} \in \mathcal{C}_0(\mathbb{R}) \cap L^2(\mathbb{R})$ and $F' \in \mathcal{C}_b(\mathbb{R})$ with $F'(0) = 0$.*
*The function $F^\varepsilon := \mathcal{I}_d[\mathrm{abs} * u_\varepsilon]$ converges in $L^2(\mathbb{R})$ and also pointwise to $F$ as $\varepsilon \to 0$.*

For the special case of $B$-splines $u := M_m$, we have the following result.

**Proposition 4.7.** *For $m \in \mathbb{N}$ with $m \geq 2$, let $f := abs * M_m$. Then we have for $d \geq 2$*

$$\mathcal{I}_d[f](s) = c_d \sum_{k=0}^{m} (-1)^k \binom{m}{k} \sum_{n=0}^{m+1} \frac{(\frac{m}{2}-2)^{m+1-n}}{n!(m+1-n)!} s^n q_d(n, k-\tfrac{m}{2}; s) - \frac{\pi c_{d+1}}{2} s, \qquad s > 0,$$

*where*

$$q_d(n,a;s) := \begin{cases} \frac{\Gamma(\frac{d-1}{2})\Gamma(\frac{n+1}{2})}{\Gamma(\frac{d+n}{2})}, & a \leq 0, \\ \frac{\Gamma(\frac{d-1}{2})\Gamma(\frac{n+1}{2})}{\Gamma(\frac{d+n}{2})} - B_{a^2/s^2}(\frac{n+1}{2}, \frac{d-1}{2}), & 0 < a < s, \\ 0, & a \geq s \end{cases}$$

*with the incomplete Beta function $B_x(a,b) := \int_0^x t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$ for $a, b > -1$ and $x \in [0,1]$.*

In particular, we obtain for $u = M_2$ and $u = M_4$ the following functions $F$.

**Example 4.8.** *For $d = 3$, it holds*

$$\mathcal{I}_3[\text{abs} * M_2](s) = \frac{1}{12} \begin{cases} -|s|^3 + 4s^2 + 4, & |s| \le 1, \\ 6|s| + \frac{1}{|s|}, & \textit{otherwise}, \end{cases}$$

*and*

$$\mathcal{I}_3[\text{abs} * M_4](s) = \frac{1}{360} \begin{cases} 3|s|^5 - 12s^4 + 80s^2 + 168, & |s| \le 1, \\ -|s|^5 + 12s^4 - 60|s|^3 + 160s^2 - 60|s| + 192 - \frac{4}{|s|}, & 1 \le |x| \le 2, \\ 180|s| + \frac{60}{|s|}, & \textit{otherwise}. \end{cases}$$

*For an illustration of the first function, see Figure 1. We have $\mathcal{I}_3[\text{abs} * M_2] \in \mathcal{C}^3(\mathbb{R})$ and $\mathcal{I}_3[\text{abs} * M_4] \in \mathcal{C}^5(\mathbb{R})$.*

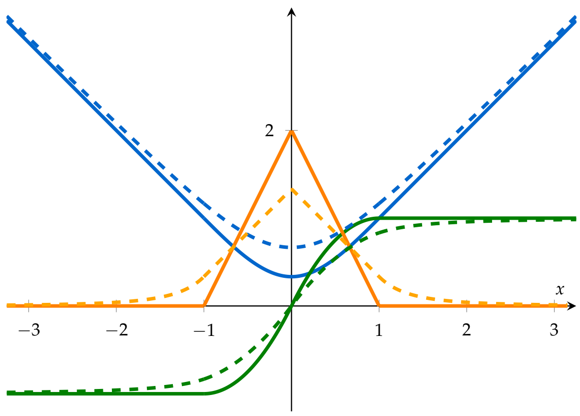



Figure 1: Smoothed absolute value $f = \text{abs} * M_2$ (solid, blue) with its first (solid, green) and second (solid, orange) derivatives, and $F = 2\mathcal{I}_3[f]$ (dashed blue) with its first (dashed, green) and second (dashed, orange) derivatives.

Based on the previous results, we propose to approximate the negative Euclidean norm on $\mathbb{R}^d$ by

$$\Phi = F(\|\cdot\|) := \mathcal{I}_d[f](\|\cdot\|), \quad f := -\text{abs} * u, \quad u \in \mathcal{U}^n(\mathbb{R}), \quad n \in \mathbb{N}_0. \tag{9}$$

Summarizing Propositions 3.1 and 4.3, this function has the following properties.

**Theorem 4.9.** *The function $\Phi$ in (9) has the following properties:*

*i) $\Phi$ is conditionally positive definite of order one on $\mathbb{R}^d$.*

*ii) $\Phi(x) < 0$ for all $x \in \mathbb{R}^d$.*

*iii) $\Phi(x) = -C_d \|x\| + \varphi(\|x\|)$ with $\varphi \in \mathcal{C}_0(\mathbb{R})$ and $\varphi(s) \in \mathcal{O}(\frac{1}{s})$ as $s \to \infty$.*

*iv)* $\Phi$ *is* $n+2$ *times continuously differentiable.*

*v)* $\nabla\Phi$ *is Lipschitz-L continuous with* $L := 2\sqrt{d}\|u\|_\infty$

*vi)* $\Phi$ *is concave and* $(-L)$*-convex, i.e., for all* $\lambda \in [0,1]$ *and all* $x, y \in \mathbb{R}^d$*, we have*

$$\Phi(\lambda x + (1-\lambda)y) \le \lambda\Phi(x) + (1-\lambda)\Phi(y) + \tfrac{L}{2}\lambda(1-\lambda)\|x-y\|^2.$$

# 5 Smoothed Distance Kernels

In this section, we show how the above functions $\Phi$ induce characteristic kernels with nice Lipschitz properties. These kernels can be used to define MMDs between measures and the MMDs can then serve as functionals for Wasserstein gradient flows.

We call a symmetric function $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a *kernel*. A kernel is *positive definite*, if for all $N \in \mathbb{N}$, all $x_1, \ldots, x_N \in \mathbb{R}^d$, and all $a \in \mathbb{C}^N$ it holds

$$\sum_{j,k=1}^{N} a_j a_k K(x_j, x_k) \ge 0.$$

Unfortunately, the kernel $K(x,y) := F(\|x-y\|)$ with $F$ in (9) is not positive definite, since $F(\|\cdot\|)$ is only conditionally positive definite of order $r=1$. However, we have the following proposition, see [57, Thm 10.18]. Here $\Pi_{r-1}(\mathbb{R}^d)$ denotes the linear space of $d$-variate polynomials of degree $\le r-1$ which has dimension $N := \binom{d+r-1}{r-1}$.

**Proposition 5.1.** *Let* $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ *be a conditionally positive definite function of order* $r \in \mathbb{N}$*. Let* $\Xi := \{\xi_k : k = 1, \ldots, N\}$ *be a set of points such that* $p(\xi_k) = 0$ *for all* $k = 1, \ldots, N$ *and any* $p \in \Pi_{r-1}(\mathbb{R}^d)$ *implies that* $p$ *is the the zero polynomial. Denote by* $p_j$*,* $j = 1, \ldots, N$ *the set of Lagrangian basis polynomials with respect to* $\Xi$*, i.e.,* $p_j(\xi_k) = \delta_{j,k}$*. Then*

$$\begin{aligned} K(x,y) := \Phi(x-y) - \sum_{j=1}^{N} p_j(x)\Phi(\xi_j - y) - \sum_{k=1}^{N} p_k(y)\Phi(x-\xi_j) \\ + \sum_{j,k=1}^{N} p_j(x)p_k(y)\Phi(\xi_j - \xi_k) \end{aligned} \tag{10}$$

*is a positive definite kernel. In particular, we have in case* $r=1$ *that*

$$\Phi(x-y) - \Phi(x) - \Phi(y) + \Phi(0)$$

*is positive definite, where we can skip the constant third term if* $\Phi(0) \le 0$*.*

For our kernel from the function in (9), we obtain directly by Theorem 4.9 v) and Proposition 5.1 the following corollary.

**Corollary 5.2.** *Let $F(\|\cdot\|)$ be defined by* (9). *Then*

$$K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}, \qquad K(x,y) := F(\|x-y\|) - F(\|x\|) - F(\|y\|) \tag{11}$$

*is a positive definite kernel and*

$$K(x,x) = F(0) - 2F(\|x\|) \in \mathcal{O}(\|x\|). \tag{12}$$

*Moreover, $K$ is continuously differentiable with Lipschitz continuous gradient, i.e.,*

$$\|\nabla K(x,x') - \nabla K(y,y')\| \leq L(\|x-y\| + \|x'-y'\|) \quad \text{for all} \quad x,x',y,y' \in \mathbb{R}^d. \tag{13}$$

# 6 Maximum Mean Discrepancy with respect to $K$

A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathbb{R}^d$ is called a *reproducing kernel Hilbert space* (RKHS), if the point evaluations $h \mapsto h(x)$, $h \in \mathcal{H}$, are continuous for all $x \in \mathbb{R}^d$. There exist various textbooks on RKHS from different points of view, see, e.g., [14, 51, 52]. By [52, Thm. 4.20], every RKHS admits a unique positive definite kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, which is determined by the reproducing property

$$h(x) = \langle h, K(x,\cdot)\rangle_{\mathcal{H}} \quad \text{for all} \quad h \in \mathcal{H}. \tag{14}$$

In particular, we have $K(x,\cdot) \in \mathcal{H}$ for all $x \in \mathbb{R}^d$ and

$$|h(x)| \leq \|h\|_{\mathcal{H}} \|K(x,\cdot)\|_{\mathcal{H}} = \|h\|_{\mathcal{H}} \sqrt{K(x,x)}. \tag{15}$$

Conversely, for any positive definite kernel $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, there exists a unique RKHS with reproducing kernel $K$, denoted by $\mathcal{H}_K$ [52, Thm. 4.21].

RKHSs are closely related to measure spaces. Let $\mathcal{M}(\mathbb{R}^d)$ denote the space of finite, real-valued Radon measures and $\mathcal{P}(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$. Further, let

$$\mathcal{M}_\alpha(\mathbb{R}^d) := \left\{ \mu \in \mathcal{M}(\mathbb{R}^d)\colon \int_{\mathbb{R}^d} \|x\|^\alpha \,\mathrm{d}\mu(x) < \infty \right\}, \quad 0 < \alpha < \infty$$

and similarly

$$\mathcal{P}_\alpha(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d)\colon \int_{\mathbb{R}^d} \|x\|^\alpha \,\mathrm{d}\mu(x) < \infty \right\}, \quad 0 < \alpha < \infty.$$

Let $K(x,x) \in \mathcal{O}(\|x\|^\alpha)$. For example, we have by (12) for our kernel in (11) that $\alpha = 1$. Then, it can be seen by (15) that $\mathcal{H}_K \subset L^1(\mu)$ for all $\mu \in \mathcal{M}_{\alpha/2}(\mathbb{R}^d)$ and the so-called *kernel mean embedding* (KME) $m\colon \mathcal{M}_{\alpha/2}(\mathbb{R}^d) \to \mathcal{H}_K$, $\mu \mapsto m_\mu$ given by

$$\langle h, m_\mu\rangle_{\mathcal{H}_K} = \int_{\mathbb{R}^d} h \,\mathrm{d}\mu \quad \text{for all} \quad h \in \mathcal{H}_K \tag{16}$$

is well-defined, meaning that for every $\mu \in \mathcal{M}_{\alpha/2}(\mathbb{R}^d)$ there exists a unique $m_\mu \in \mathcal{H}_K$ such that (16) is fulfilled [52, Lemma 4.24]. In particular, we have by (14) that

$$m_\mu(x) = \int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\mu(y). \tag{17}$$

The KME is not surjective [53]. For a positive definite kernel $K$ with $K(x,x) \in \mathcal{O}(\|x\|^\alpha)$, the *maximum mean discrepancy* (MMD) $\mathcal{D}_K \colon \mathcal{M}_{\alpha/2}(\mathbb{R}^d) \times \mathcal{M}_{\alpha/2}(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ is by (15) well-defined by

$$\begin{aligned}
\mathcal{D}_K^2(\mu,\nu) &:= \int_{\mathbb{R}^d \times \mathbb{R}^d} K(x,y)\,\mathrm{d}(\mu(x)-\nu(x))\,\mathrm{d}(\mu(y)-\nu(y)) \qquad (18)\\
&= \int_{\mathbb{R}^d \times \mathbb{R}^d} K(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) - 2\int_{\mathbb{R}^d \times \mathbb{R}^d} K(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\nu(y)\\
&\quad + \int_{\mathbb{R}^d \times \mathbb{R}^d} K(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y)\\
&= \|m_\mu - m_\nu\|_{\mathcal{H}_K}^2,
\end{aligned}$$

see [6, 23], where the last equality follows directly from the KME (17). If the KME is injective, then $K$ is called a *characteristic kernel*. In this case, the MMD $\mathcal{D}_K$ is a distance on $\mathcal{M}_{\alpha/2}(\mathbb{R}^d)$. Kernels induced by Gaussians are typical characteristic kernels. By the following proposition, also our kernel (11) is characteristic, so that $\mathcal{D}_K$ is a distance on $\mathcal{M}_{1/2}(\mathbb{R}^d)$.

**Proposition 6.1.** *Let $K$ be defined by* (11). *Then the kernel mean embedding $m \colon \mathcal{M}_{1/2}(\mathbb{R}^d) \to \mathcal{H}_K$ in* (16) *is injective, i.e. $K$ is a characteristic kernel. More precisely, for all $\mu \in \mathcal{M}_{1/2}(\mathbb{R}^d)$, it holds*

$$\|m_\mu\|_{\mathcal{H}_K}^2 = \frac{1}{w_{d-1}\pi^2}\int_{\mathbb{R}^d} |\mu(\mathbb{R}^d) - \hat{\mu}(s)|^2 \frac{\hat{u}(\|s\|)}{\|s\|^{d+1}}\,\mathrm{d}s - F(0)\,\mu(\mathbb{R}^d)^2,$$

*where $\hat{\mu}$ denotes the Fourier transform of $\mu$, see* (28).

Fortunately, by the following theorem, when dealing with MMDs it is not necessary to work with the clumsy kernels (10), but instead we can directly use the conditionally positive definite kernels. Note that the MMD with respect to the negative distance kernel is also known as energy distances in statistics [55].

**Theorem 6.2.** *Let $\Phi \in \mathrm{CP}_r(\mathbb{R}^d)$ with $r \in \mathbb{N}$ fulfill $\Phi \in \mathcal{O}(\|\cdot\|^\alpha)$, and let $\tilde{K}(x,y) := \Phi(x-y)$. Define the associate positive definite kernel $K$ by* (10). *Then $\mathcal{D}_{\tilde{K}}$ in* (18) *is well-defined for $\mu,\nu \in \mathcal{M}_\alpha(\mathbb{R}^d)$ and $\mathcal{D}_K$ for $\mu,\nu \in \mathcal{M}_\beta(\mathbb{R}^d)$, where $\beta := \max\{r-1, (r-1+\alpha)/2\}$. If $\mu,\nu \in \mathcal{M}_\alpha(\mathbb{R}^d) \cap \mathcal{M}_\beta(\mathbb{R}^d)$ have the same first $r-1$ moments, i.e.*

$$\int_{\mathbb{R}^d} p(x)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} p(x)\,\mathrm{d}\nu(x) \quad \text{for all } p \in \Pi_{r-1}(\mathbb{R}^d),$$

*then*

$$\mathcal{D}_{\tilde{K}}(\mu,\nu) = \mathcal{D}_K(\mu,\nu).$$

For our function $\Phi(x) := F(\|x\|)$ with $F$ in (9), we know already that $\mathcal{D}_K$ is well-defined for measures in $\mathcal{M}_{1/2}(\mathbb{R}^d)$ which is in agreement with the proposition. However, by the proposition, $\mathcal{D}_{\tilde{K}}$ is only well-defined for measures in $\mathcal{M}_1(\mathbb{R}^d)$. If in addition $\int_{\mathbb{R}} \mathrm{d}\mu = \int_{\mathbb{R}} \mathrm{d}\nu$, then their distances $\mathcal{D}_K$ and $\mathcal{D}_{\tilde{K}}$ are the same. In particular, both distances are well-defined and coincide for measures in $\mathcal{P}_1(\mathbb{R}^d) \supset \mathcal{P}_2(\mathbb{R}^d)$.

By the following remark, there is a relation between the degree of conditional positive definiteness and the growth of a function $\Phi$ towards infinity.

**Remark 6.3.** *By [34, Cor 2.3], we have*

$$\Phi \in \mathrm{CP}_r(\mathbb{R}^d) \quad \Longrightarrow \quad \Phi \in \mathcal{O}(\|\cdot\|^{2r}),$$

*which implies that $\alpha \le 2r$ in the assumption of Theorem 6.2. In general, this bound cannot be improved, since $(-1)^r\|\cdot\|^{2r-\varepsilon} \in \mathrm{CP}_r(\mathbb{R}^d)$ for any $r \in \mathbb{N}$ and $\varepsilon \in [0,2)$ by [57, Cor 8.18] and [54, Lem 3.3]. However, for our function $\Phi(x) := F(\|x\|)$ with $F$ in* (9)*, the above result says that $\Phi \in \mathcal{O}(\|\cdot\|^2)$, but we know already that $\Phi \in \mathcal{O}(\|\cdot\|)$.*

Finally, smoothness properties of the kernel transfer to the corresponding RKHS.

**Proposition 6.4.** *For $d \ge 3$ and $n \ge 0$, let $u \in \mathcal{U}^n(\mathbb{R}^d)$. Let the kernel $K$ be given by* (11)*. Then every $h \in \mathcal{H}_K$ is $\lfloor \frac{n+2}{2} \rfloor$-times continuously differentiable. If $n \ge 2$ is even, then the gradient $\nabla h$ is $\sqrt{2d\|u''\|_\infty}\|h\|_{\mathcal{H}_K}$ Lipschitz continuous.*

# 7 Wasserstein Gradient Flows of MMDs

The behavior of Wasserstein gradient flows of MMDs depends on the kernel in their definition. While there exist many results for smooth kernels like the Gaussian, see, e.g., [3], gradient flows of MMDs with Riesz kernels and in particular with the negative distance kernel have completely different properties, see, e.g., [27]. In contrast to smooth kernels, empirical measures do in general not remain empirical ones along the flow. Even if a steepest descent scheme, resp. the implicit Euler scheme exists, a convergence theory is still missing in dimensions larger than one.

Let us briefly recall basic facts on Wasserstein gradient flows, see [2, 47] and show that our new kernels fulfill all assumptions which are required to ensure the existence of its MMD gradient flow and the convergence of a forward and backward schemes.

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by

$$\Pi(\mu,\nu) := \{\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : (P_1)_\#\pi = \mu,\ (P_2)_\#\pi = \nu\}$$

the set of couplings with marginals $\mu$ and $\nu$, and by $(P_i)_\#\mu := \mu \circ P_i^{-1} \in \mathcal{P}_2(\mathbb{R}^d)$ the *pushforward* of $\mu$ with respect to the projection $P_i(x_1, x_2) := x_i$, $i = 1, 2$. Together with the Wasserstein distance

$$W_2(\mu,\nu)^2 := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \, \mathrm{d}\pi(x,y), \qquad \mu,\nu \in \mathcal{P}_2(\mathbb{R}^d), \tag{19}$$

the set $\mathcal{P}_2(\mathbb{R}^d)$ becomes a complete metric space. The set of optimal couplings in (19) is denoted by $\Pi_{\mathrm{opt}}(\mu,\nu)$. A curve $\gamma\colon I \to \mathcal{P}_2(\mathbb{R}^d)$ on an interval $I = [a,b]$, $a < b$ is called *absolutely continuous*, if there exists a Borel velocity field $v\colon I \times \mathbb{R}^d \to \mathbb{R}^d$ with $\|v_t\|_{L^2(\mathbb{R}^d;\gamma_t)} \in L^1(I)$ such that the *continuity equation*

$$\partial_t \gamma_t + \nabla_x \cdot (v_t \gamma_t) = 0$$

is fulfilled on $I \times \mathbb{R}^d$ in a weak sense, i.e., for all $\varphi \in C_c^\infty\big((a,b)\times\mathbb{R}^d\big)$ it holds

$$\int_0^\infty \int_{\mathbb{R}^d} \partial_t \varphi(t,x) + \langle \nabla_x \varphi(t,x), v_t(x)\rangle \,\mathrm{d}\gamma_t(x)\,\mathrm{d}t = 0.$$

There are many velocity fields corresponding to the same absolutely continuous curve, but only one with minimal $\|v_t\|_{L^2(\mathbb{R}^d;\gamma_t)}$ for a.e. $t \in I$. For a lower semi-continuous function $G\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, the reduced Fréchet subdifferential $\partial G$ consists of all $v \in L^2(\mathbb{R}^d,\mu;\mathbb{R}^d)$ such that for all $\eta \in \mathcal{P}_2(\mathbb{R}^d)$,

$$G(\eta) - G(\mu) \geq \inf_{\pi \in \Pi_{\mathrm{opt}}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \langle v(x), y - x\rangle \,\mathrm{d}\pi(x,y) + o(W_2(\mu,\nu)).$$

If the minimal velocity field in the continuity equation is determined by

$$v_t \in -\partial G(\gamma_t), \quad \text{for a.e.} \quad t > 0, \tag{20}$$

then $\gamma_t$ is called *Wasserstein gradient flow of $G$*.

Let $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a characteristic kernel such that its MMD is well-defined for measures in $\mathcal{P}_2(\mathbb{R}^d)$. Examples are Gaussian kernels, the negative distance kernel, as well as our smoothed negative distance kernels in (11). For a fixed target measure $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, we consider gradient flows of the squared MMD functional

$$G\colon \mathcal{P}_2(\mathbb{R}^d) \to [0,\infty), \quad G(\mu) := \tfrac{1}{2}\mathcal{D}_K^2(\mu,\nu). \tag{21}$$

If $K$ is continuously differentiable, the velocity field in (20) becomes

$$\begin{aligned} v_t = -\nabla_x \frac{\delta G}{\delta \gamma_t} &= -\nabla_x \int_{\mathbb{R}^d} K(\cdot,y)\ (\mathrm{d}\gamma_t(y) - \mathrm{d}\nu(y)) \\ &= -\int_{\mathbb{R}^d} \nabla_x K(\cdot,y)\ (\mathrm{d}\gamma_t(y) - \mathrm{d}\nu(y)), \end{aligned} \tag{22}$$

see, e.g., [47]. Here, $\frac{\delta G}{\delta \gamma}$ denotes the functional derivative defined, if it exists, by the function with $\frac{\mathrm{d}}{\mathrm{d}\epsilon} G(\gamma + \epsilon(\eta - \gamma))\big|_{\epsilon=0} = \int \frac{\delta G}{\delta \gamma}(\gamma)(\mathrm{d}\eta - \mathrm{d}\gamma)$ for any $\eta \in \mathcal{P}_2(\mathbb{R}^d)$. Note that $v_t = -\nabla_x m_{\gamma_t - \nu}$ for a positive definite kernel. For the negative distance kernel, we can compute $\frac{\delta G}{\delta \gamma_t}$ as above, but the gradient $\nabla_x$ does not exist in $x = y$, i.e., (22) is not well-defined, which causes the different behavior of those flows. The following result guarantees the existence of Wasserstein gradient flows of MMDs with sufficiently smooth kernels and its approximation by a Euler forward scheme.

**Proposition 7.1.** *[3, Prop 1&3] Let $K \in \mathcal{C}^1(\mathbb{R}^d \times \mathbb{R}^d)$ be a positive definite, characteristic kernel that has a Lipschitz-continuous gradient in the sense of* (13). *Then, for any $\nu, \mu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique Wasserstein gradient flow $\gamma\colon [0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ of the MMD functional* (21) *starting in $\gamma^{(0)} = \mu$. For a step size $\tau > 0$, we define the Euler forward iteration by*

$$\gamma^{(k+1)} := (I - \tau v^{(k)})_{\#}\gamma^{(k)} \tag{23}$$

*where $v^{(k)}$ is related to $\gamma^{(k)}$, $k \in \mathbb{N}_0$ by* (22). *The approximated interpolation path*

$$\gamma_t^\tau := (I - (t - k\tau)v^{(k)})_{\#}\gamma^{(k)}, \qquad t \in [k\tau, (k+1)\tau),$$

*satisfies $W_2(\gamma_t^\tau, \gamma_t) \le \tau\, C_T$ for all $t \in [0, T]$, where the constant $C_T$ depends only on $T > 0$.*

By Proposition 6.1 and Corollary 5.2, we obtain the following for our smoothed norm kernel.

**Corollary 7.2.** *The kernel $K(x,y) = F(\|x-y\|) - F(\|x\|) - F(\|y\|)$ with $F$ from* (9) *fulfills the conditions of Proposition 7.1. There exists a Wasserstein gradient flow of the corresponding MMD functional* (21) *and it can be approximated by the Euler forward scheme* (23). *It holds*

$$v_t = -\int_{\mathbb{R}^d} \nabla_x K(\cdot, y)\, \mathrm{d}(\gamma_t - \nu)(y) = -\int_{\mathbb{R}^d} \nabla_x F(\|x-y\|)\, \mathrm{d}(\gamma_t - \nu)(y),$$

*so that $K$ can be replaced by $\tilde{K}(x,y) = F(\|x-y\|)$ without changing the flow results.*

**Remark 7.3.** *Let $d' \ge d$ and $F = \mathcal{I}_{d'}[f]$ for $f \in \mathrm{CP}_r(\mathbb{R})$. By Proposition 4.3, we have $F(\|\cdot\|) \in \mathrm{CP}_r(\mathbb{R}^{d'})$ and, hence, also $F(\|\cdot\|) \in \mathrm{CP}_r(\mathbb{R}^d)$. Therefore, we can also use $F = \mathcal{I}_{d'}[f]$ to smooth the negative distance kernel in Corollary 7.2.*

There is a more general theory on Wasserstein gradient flows of $\lambda$-convex functionals, $\lambda \in \mathbb{R}$, along generalized geodesics, see [2, Thm. 11.2.1]. In Appendix D, we show that the functional $G$ in (21) with our smoothed negative distance kernel fulfills this $\lambda$-convexity with $\lambda < 0$ and establish an analogue to Corollary 7.2 for the Euler backward scheme. In particular, note that it is only ensured for $\lambda > 0$ that the gradient flow converges to the (global) minimizer of $G$ as $t \to \infty$. Example D.5 in Appendix D shows that convergence to the global minimizer $\nu$ in (21) is in general not ensured, and the iteration may become stuck in another extreme point.

Finally, let us mention that our new kernel can also be used in the definition of other functionals $G$, e.g., MMD-regularized $f$-divergences, where so far only bounded positive definite, characteristic kernels were applied.

**Remark 7.4** (MMD-regularized $f$-Divergence)**.** *In [41], inspired by [20], Wasserstein gradient flows of MMD-regularized $f$-divergences were considered. Unfortunately, the approach requires differentiability of the kernel and therefore does not work for negative distance kernels. In contrast, using Proposition 6.4, it can be shown that our new smoothed distance kernel fits into the setting of the above papers.*

## 8 Numerical Results

In this section, we consider empirical probability measures

$$\mu := \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}, \qquad \nu := \frac{1}{M}\sum_{m=1}^{M}\delta_{y_m}, \qquad x_n, y_m \in \mathbb{R}^d,$$

where $\delta_x$ is the Dirac measure at $x \in \mathbb{R}^d$. The MMD (18) between these measures is

$$\mathcal{D}_K^2(\mu,\nu) = \frac{1}{N^2}\sum_{n,n'=1}^{N} K(x_n, x_{n'}) - \frac{2}{MN}\sum_{n,m=1}^{N,M} K(x_n, y_m) + \frac{1}{M^2}\sum_{m,m'=1}^{M} K(y_m, y_{m'}).$$

We compare the gradient flows of $G = \frac{1}{2}\mathcal{D}_K^2(\cdot,\nu)$ for $K(x,y) = F(\|x-y\|)$ with the

a) **Gaussian** $F(s) := \exp(\frac{-s^2}{2\sigma^2})$ for $\sigma > 0$,

b) **SND**: smoothed negative distance

$$F := -\mathcal{I}_3[\mathrm{abs} * u_\varepsilon] \text{ with } u_\varepsilon(x) := \tfrac{1}{\varepsilon} M_2(\tfrac{x}{\varepsilon}), \tag{24}$$

c) **ND**: negative distance $F := -\frac{1}{2}\mathrm{abs}$.

Remark 7.3 entitles us to use $\mathcal{I}_3$ instead of $\mathcal{I}_2$ in the definition of SND.

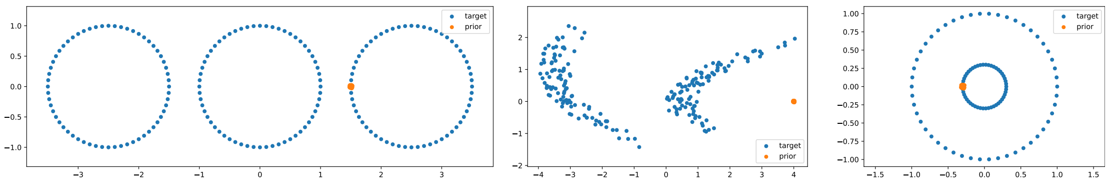
The Wasserstein gradient flow of the MMD with a kernel fulfilling the assumptions of Proposition 7.1 keeps the empirical measure structure and moves just the positions of the Dirac measures. This holds true for the Gaussian and for our SND kernel. For the ND kernel, we have to replace $G(\mu)$ by $+\infty$ if $\mu$ is not an empirical measure, see, e.g. [30]. The forwards Euler scheme (23) for smooth kernels reads as

$$\begin{aligned} x_i^{(k+1)} = x_i^{(k)} - \tau\Big(&\frac{1}{2N}\sum_{n=1}^{N}(x_i^{(k)} - x_n^{(k)})\frac{F'(\|x_i^{(k)} - x_n^{(k)}\|)}{\|x_i^{(k)} - x_n^{(k)}\|} \\ &- \frac{1}{M}\sum_{m=1}^{M}(x_i^{(k)} - y_m)\frac{F'(\|x_i^{(k)} - y_m\|)}{\|x_i^{(k)} - y_m\|}\Big). \end{aligned} \tag{25}$$

For the Gaussian as well as for the SND kernel, we have $F'(0) = 0$ and by L'Hôpital's rule

$$F''(0) = \lim_{s\to 0}\frac{F'(s)}{s}$$

is well-defined. For the ND kernel, the summands in (25) have just the form $\frac{x}{\|x\|}$ with $x \in \{x_i^{(k)} - x_n^{(k)}, x_i^{(k)} - y_m : i, n = 1, \ldots, N; m = 1, \ldots, M\}$ if $x \neq 0$, and we set $\frac{x}{\|x\|} := 0$ for $x = 0$.




(a) Three-rings (b) Bananas (c) Annulus

Figure 2: Target measures $\nu$ (blue) and initialization $\gamma^{(0)}$ (orange).

**Fast Summation by Slicing.** The computation of (25) includes the summation of kernel values of the form $s_m = \sum_{n=1}^N w_n F'(\|x_n - y_m\|)$ with some weights $w_n \in \mathbb{C}$. If $F = \mathcal{I}_d[f]$, then we have by (8) that $F'(x) = \mathbb{E}_{\xi \sim \mathcal{U}_{\mathbb{S}^{d-1}}}[\xi f'(\langle x, \xi \rangle)]$. Now the summation can be approximated by slicing [26, 28, 30] via

$$s_m = \mathbb{E}_{\xi \sim \mathcal{U}_{\mathbb{S}^{d-1}}}\Big[\sum_{n=1}^N w_n \xi f'(\langle x_n - y_m, \xi \rangle)\Big] \approx \frac{1}{P} \sum_{p=1}^P \xi_p \sum_{n=1}^N w_n f'(\langle x_n - y_m, \xi_p \rangle), \qquad (26)$$

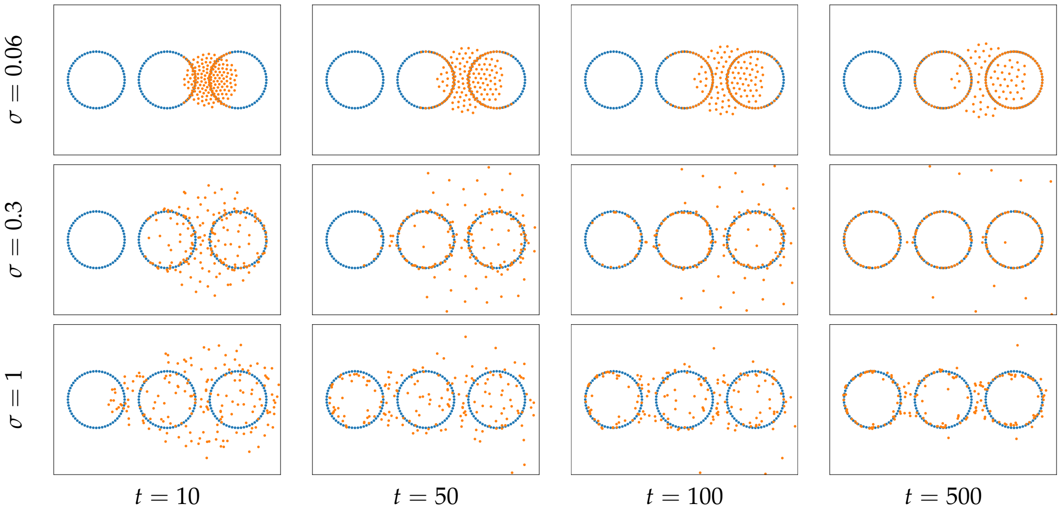
where $(\xi_p)_{p=1}^P \in (\mathbb{S}^{d-1})^P$ are equidistributed quadrature nodes on $\mathbb{S}^{d-1}$. This is a collection of $P$ one-dimensional kernel sums, which can be computed efficiently, e.g. via fast Fourier summation [43] or, if $F$ is the ND kernel just by sorting [26].

### 8.1 Three-Rings Target

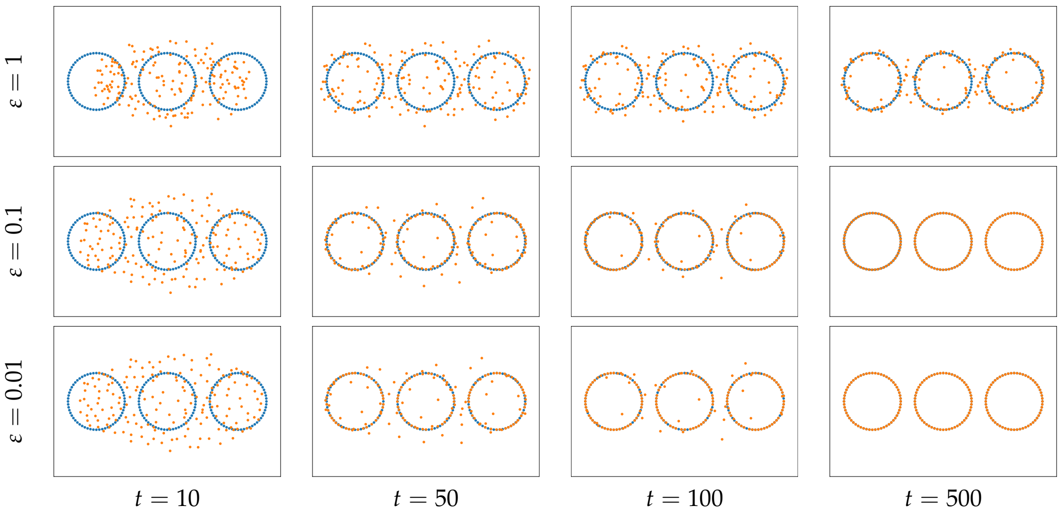
The three-rings target $\nu$ in Figure 2a from [20, Fig. 1] consists of three circles in $\mathbb{R}^2$ with radius 1 and midpoints $(-2.5, 0)$, $(0, 0)$ and $(2.5, 0)$ discretized with $M = 3 \cdot 40 = 120$ points. The initialization $\gamma^{(0)}$ is a highly localized Gaussian with standard deviation $10^{-4}$, see Figure 2a.

We computed the iteration (23) with step size $\tau = 0.01$ in double precision and display the flow after $k \in \{1\,000, 5\,000, 10\,000, 50\,000\}$ iterations or equivalently after time $t = \tau k$. Figure 3a shows the flows for the Gaussian kernel with standard deviation $\sigma \in \{0.06, 0.3, 1\}$. Here, the quality of the result heavily depends on the choice of $\sigma$. If $\sigma$ is too small, the points cover only two circles; if too large, the points do not lie on the circles. The sweet spot is around $\sigma = 0.3$, but even then some particles get stuck far from the target. Figure 3b shows the flows for the SND kernel (24) for $\varepsilon \in \{1, 0.1, 0.01\}$. Here, it is preferable to choose a small $\varepsilon$, then all three circles are recovered well. Figure 3c depicts the results with the ND kernel. The flows in Figure 3c and Figure 3b for $\varepsilon = 0.01$ are almost identical.

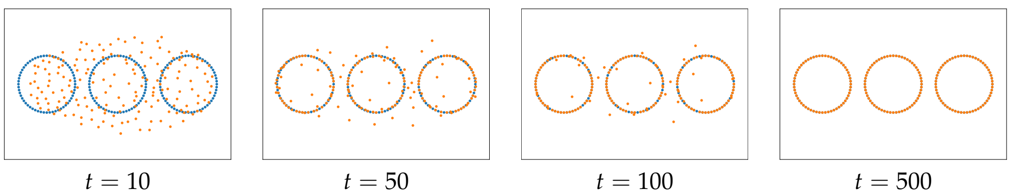
In Figure 4, we plot the Wasserstein error $W_2(\gamma_t^\tau, \nu)$ between the three-rings target measure $\nu$ and the discretized Wasserstein gradient flow $\gamma_t^\tau$ at time $t$ computed with PythonOT [17]. The first plot in Figure 4 corresponds to the flow $\gamma_t^\tau$ shown in Figures 3a, 3b, and 3c. The remaining plots in Figure 4 depict the same experiment with different step sizes $\tau$ and machine precision, where we always used the same random seed. Regardless of precision, step size, or bandwidth $\sigma$, the Gauss kernel stagnates away from the target measure $\nu$.




(a) Gaussian




(b) SND




(c) ND

Figure 3: MMD flow (25) with step size $\tau = 0.01$. For the Gaussian kernel, the result depends heavily on the choice of the parameter $\sigma$. For our SND kernel with small $\varepsilon$, the performance is as good as for the ND kernel, which is better than for the Gaussians.

For the behavior of the ND and SND kernel close to the target we first consider the following proposition, whose proof is given in Appendix D.

**Proposition 8.1.** *Let $F \in \mathcal{C}^1((0,\infty))$. For the target measure $\nu = \delta_y$ and the initial measure $\gamma^{(0)} = \delta_{x^{(0)}}$ with $y, x^{(0)} \in \mathbb{R}^d$, the sequence $x^{(k)}$ from* (25) *simplifies to*

$$x^{(k+1)} = x^{(k)} + \tau(x^{(k)} - y)\frac{F'(\|x^{(k)} - y\|)}{\|x^{(k)} - y\|}, \tag{27}$$

*and we have the following:*

*i) If $F = -\frac{1}{2}\,\mathrm{abs}$ with step size $\tau > 0$ and $0 < \|y - x^{(0)}\| < \frac{\tau}{2}$, then $x^{(k)} = x^{(0)}$ for even $k$ and $x^{(k)} = x^{(1)}$ for odd $k$. In particular, $(x^{(k)})_k$ does not converge to $y$.*

*ii) If $F = -\mathcal{I}_d[\mathrm{abs} * u]$ with $u \in \mathcal{U}^0(\mathbb{R})$, then for sufficiently small $\tau$ and $\|x^{(0)} - y\| < \tau$, the sequence $(x^{(k)})_k$ converges exponentially to $y$.*

Proposition 8.1 is for a single Dirac flow, but gives an intuition when $x_i^{(k)} \approx y_i$. In (25), the repulsion term of $x_i$ and $x_j$ approximately cancels with the attraction term of $x_i$ and $y_j$, leaving only the attraction of $x_i$ and $y_i$ as in (27). For the ND kernel, Proposition 8.1 i) indicates that $\mu_t^\tau$ oscillates around the target for any $\tau > 0$ without convergence. In contrast, Proposition 8.1 ii) states that for the SND kernel, $W_2(\mu_t^\tau, \nu)$ decays exponentially if $\tau$ is sufficiently small. Numerically, Figure 4 confirms this behavior. In single precision, ND and SND with $\varepsilon = 0.01$ plateau at $\approx 10^{-3}$. With double precision, ND oscillates at the same error, while SND drops to $\approx 10^{-7}$. Thus, SND matches ND globally but exhibits better local convergence for fixed $\tau > 0$ due to its smoothness.

An additional numerical example with two concentric circles is given in Appendix E.

## 8.2 Bananas Target

The Bananas target $\nu$ in Figure 2b is inspired by the talk from Aude Genevay[1] and its implementation by Viktor Stein[2]. The target consists of two banana shaped clusters in $\mathbb{R}^2$, where each banana consists of 100 points, so $M = 200$.

We compute the flows with step size $\tau = 0.02$ in double precision for the Gauss kernel with $\sigma \in \{0.06, 0.3, 1\}$, the SND kernel with $\varepsilon \in \{0.1, 0.01, 0.001\}$, and the ND kernel, see Figure 5. For small $\sigma = 0.06$ the Gauss kernel struggles to reach the bananas. When $\sigma = 0.3$, the right banana is reached, but some particles blow up and leave the frame. For $\sigma = 1$, the flows reaches the bananas, but collapses in the modes and do not recover the structure of the target. In contrast, the SND flow always manages to reach both bananas without blowing up, while a smaller $\varepsilon$ again gives more desirable results. The respective Wasserstein errors in Figure 6 show a similar behavior as for the rings.

[1] MIFODS Workshop on Learning with Complex Structure 2020, see `https://youtu.be/TFdIJib_zEA`.
[2] https://github.com/ViktorAJStein/Regularized_f_Divergence_Particle_Flows.

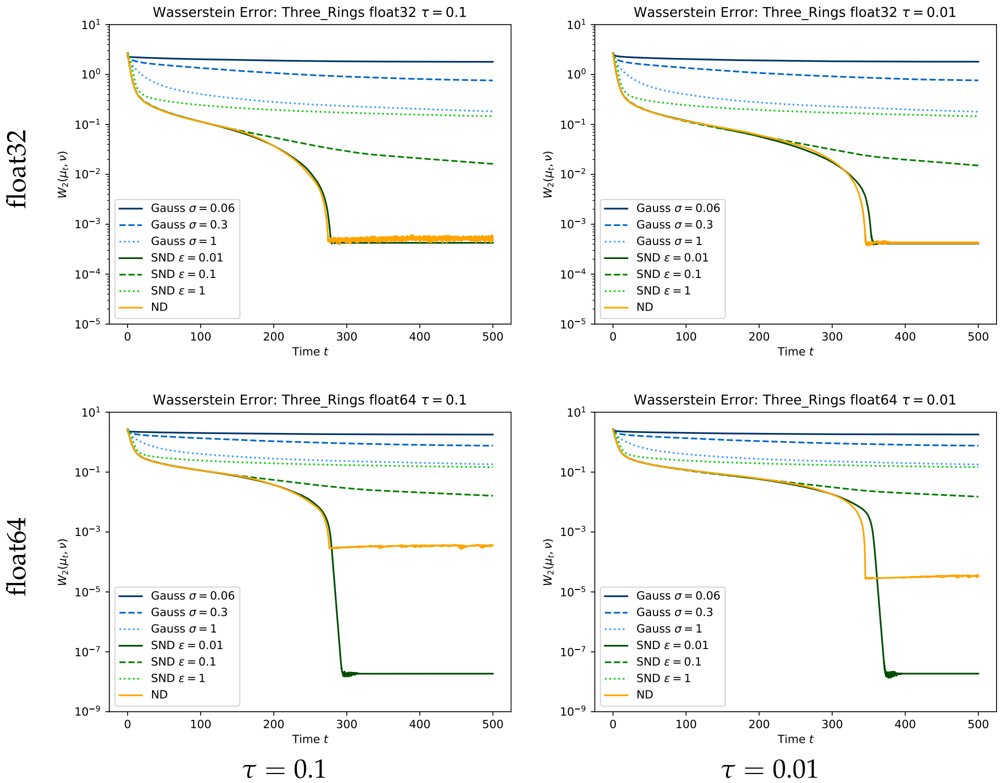
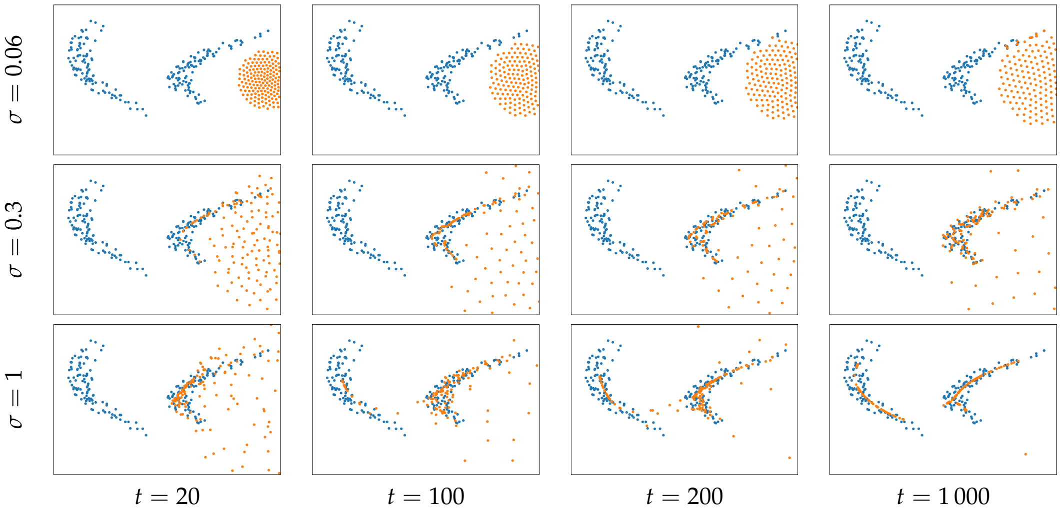
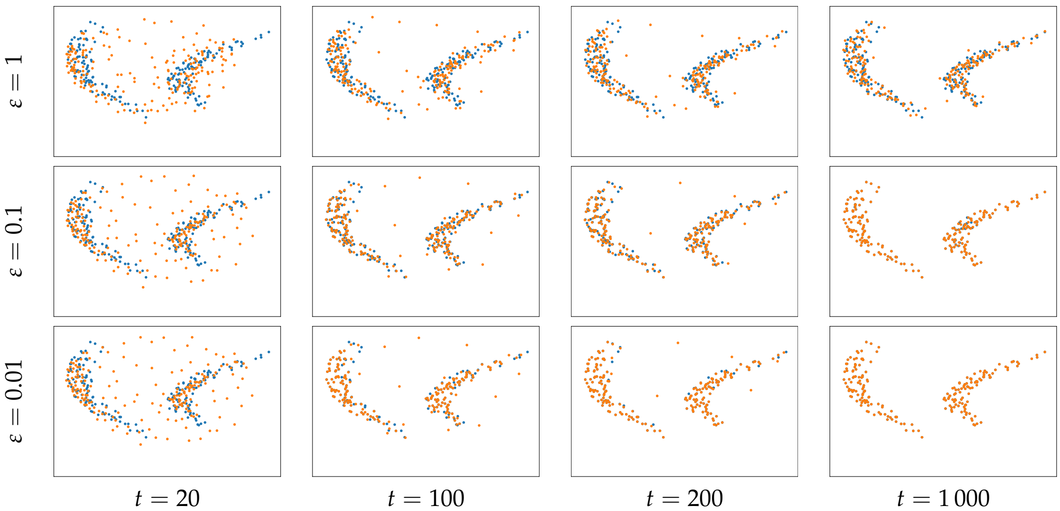


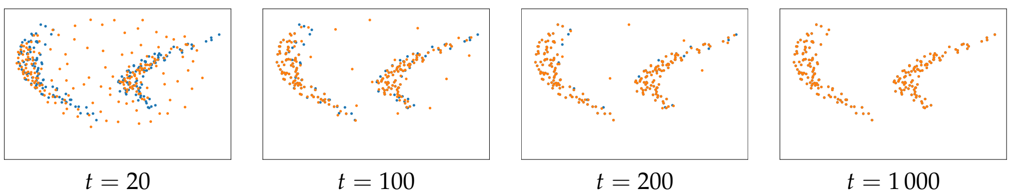

Figure 4: $W_2$ error between three-rings target $\nu$ and flow $\gamma_t^\tau$ after $k$ with time $t = \tau k$. We compare single precision (first row) and double precision (second row) for step sizes $\tau = 0.1$ (left) and $\tau = 0.01$ (right). In single precision, SND with $\varepsilon = 0.01$ and ND have the smallest error which gets stuck in $\approx 10^{-3}$. In double precision, SND with $\varepsilon = 0.01$ even outperforms ND. For some explanation, see Proposition 8.1.




(a) Gaussian




(b) SND




(c) ND

Figure 5: MMD flow (25) with step size $\tau = 0.02$. For the Gaussian kernel, the result depends heavily on the choice of the parameter $\sigma$. For our SND kernel with small $\varepsilon$, the performance is as good as for the ND kernel, which is better than for the Gaussians.

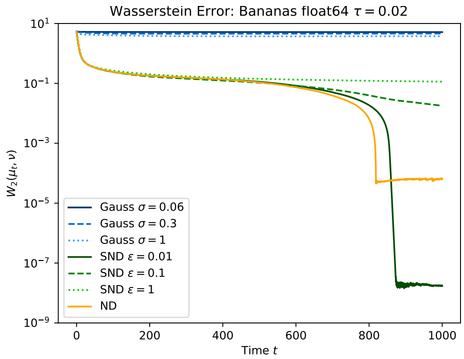



Figure 6: $W_2$ error between bananas target $\nu$ and flow $\gamma_k^\tau$ after $k$ iterations with time $t = \tau k$. Computation with double precision with step size $\tau = 0.02$ shows a similar behavior as in Figure 4.

### 8.3 MNIST Dataset

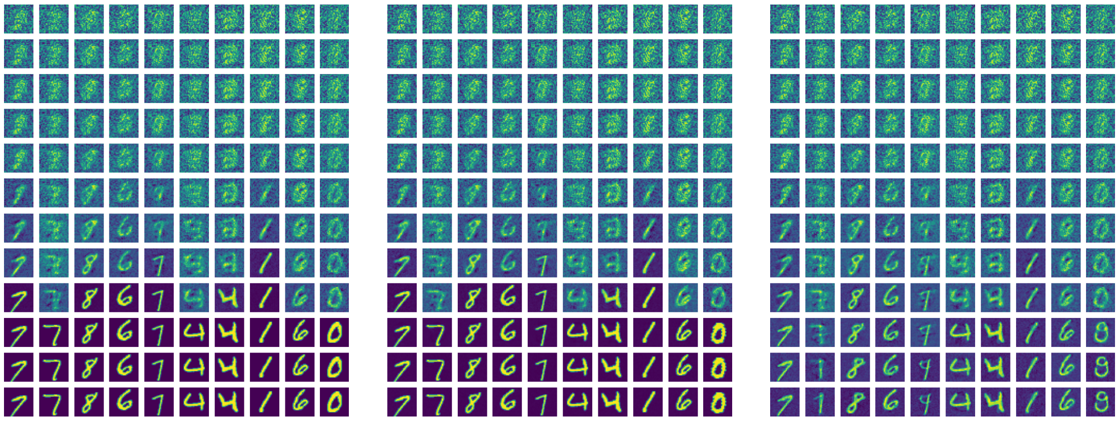
We consider as target the MNIST dataset, where each $28 \times 28$ image is considered a point $y \in \mathbb{R}^d$ with $d = 784 = 28^2$. We use $N = M = 100$ images as flow and target. The initialization $x_n^{(0)}$, $n = 1, \ldots, N$, are iid samples from a uniform distribution on $[0,1]^d$. We compute the MMD flows (25) with $2^{15} = 32768$ iterations for the SND kernel $F = -C_{784}\mathcal{I}_{784}[\mathrm{abs} * M_{2,\varepsilon}]$ with $\varepsilon \in \{0.001, 0.01, 0.1\}$ and the ND kernel $F = -\mathrm{abs}$. The step size is $\tau = 1$ and the computations are preformed in single precision. We use slicing summation (26) with $P = 785$ directions $\xi_p$ that are the vertices of the centrally symmetric simplex, to which we apply a random rotation in each iteration step, cf. [28]. The slicing summation requires only the sliced kernel $\mathrm{abs} * M_{2,\varepsilon}$ given in (6), but not the representation of $F$, which becomes quite clumsy in general, see Proposition 4.7.

The resulting images are shown in Figure 7, where we see the MMDs for the SND with small $\varepsilon$ and the Riesz kernel work comparably well and converge to the target measure $\nu$. For larger smoothing parameter $\varepsilon$, the flow needs more iterations to converge.

The distance to the target measure in the Wasserstein and MMD metrics is shown in Figure 8. Here we use the sliced approximation of the MMD. Note that the MMD, which is the objective we minimize, still depends on the kernel $K$. We see a similar behavior as for the previous low-dimensional examples with the error plateauing at some level, which becomes better for smaller $\varepsilon$ even slightly beating the ND kernel.

## 9 Conclusions

We introduced a smoothed negative distance kernel as an alternative to the negative distance kernel in MMDs. The novel kernel retains desired numerical properties of the negative distance kernel, but comes with well-defined gradient expressions and theoretical convergence guarantees of the corresponding gradient flow schemes. Therefore



(a) ND (b) SND ($\varepsilon = 0.01$). (c) SND ($\varepsilon = 0.1$).

Figure 7: MMD flow for MNIST target with different kernels. Each row shows the first 10 images $x_n \in \mathbb{R}^{28\times 28}$, the $\ell$-th row corresponds to the iteration $k = 2^{3+\ell}$, $\ell = 1, \dots, 12$.

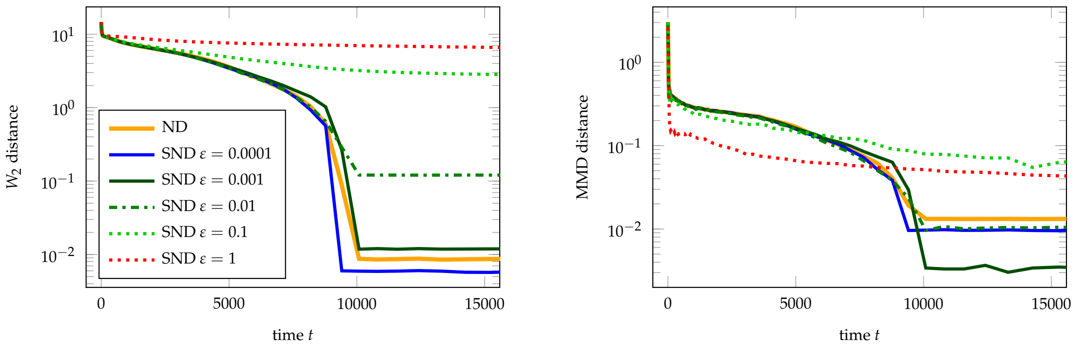



Figure 8: MMD flow for MNIST target. Left: Wasserstein distance $W_2(\gamma^{(k)}, \nu)$. Right: MMD distance $\frac{1}{2}\mathcal{D}_K^2(\gamma^{(k)}, \nu)$ for the respective kernels $K$.

our novel kernel appears to be well suited for various applications.

Concerning our future work, it may be interesting to examine if our kernel can be also used in Stein variational gradient descent [32, 42], where negative distance kernels do neither theoretically nor practically work.

# References


[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 10th edition, 1964.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, second edition, 2008.

[3] M. Arbel, A. Korba, A. Salim, and A. Gretton. Maximum mean discrepancy gradient flow. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, 2019.

[4] D. Balagué, J. A. Carrillo, T. Laurent, and G. Raoul. Dimensionality of local minimizers of the interaction energy. *Archive for Rational Mechanics and Analysis*, 209:1055–1088, 2013.

[5] H. H. Bauschke and P. L. Combettes. *Correction to: Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, Cham, 2017.

[6] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 07 2006.

[7] S. Boufadène and F.-X. Vialard. On the global convergence of Wasserstein gradient flow of the Coulomb discrepancy. HAL preprint hal-04282762, 2023.

[8] J. Carrillo, M. Delgadino, and A. Mellet. Regularity of local minimizers of the interaction energy via obstacle problems. *Communications on Mathematical Physics*, 343(3):747–781, 2016.

[9] J. Carrillo, M. Di Francesco, A. Esposito, S. Fagioli, and M. Schmidtchen. Measure solutions to a system of continuity equations driven by newtonian nonlocal interactions. *Discrete and Continuous Dynamical Systems*, 40(2):1191–1231, 2020.

[10] J. Carrillo and Y. Huang. Explicit equilibrium solutions for the aggregation equation with power-law potentials. *Kinetic and Related Models*, 10(1):171–192, 2017.

[11] J. Carrillo and R. Shu. From radial symmetry to fractal behavior of aggregation equilibria for repulsive-attractive potentials. *Calculus of Variations and Partial Differential Equations*, 62(1), 2023.

[12] D. Chafai, E. B. Saff, and R. S. Womersley. On the solution of a Riesz equilibrium problem and integral identities for special functions. *Journal of Mathematical Analysis and Applications*, 515:126367, 2022.

[13] G. Criscuolo. A new algorithm for Cauchy principal value and Hadamard finite-part integrals. *Journal of Computational and Applied Mathematics*, 78:255–275, 1997.

[14] F. Cucker and D. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

[15] R. Duong, V. Stein, R. Beinert, J. Hertrich, and G. Steidl. Wasserstein gradient flows of MMD functionals with distance kernel and Cauchy problems on quantile functions, 2024.

[16] M. Ehler, M. Gräf, S. Neumayer, and G. Steidl. Curve based approximation of measures on manifolds by discrepancy minimization. *Foundations of Computational Mathematics*, 21(6):1595–1642, 2021.

[17] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[18] R. L. Frank and R. M. Matzke. Minimizers for an aggregation model with attractive-repulsive interaction. *Archive for Rational Mechanics and Analysis*, 249(15), 2025.

[19] I. Gelfand and G. Shilov. *Generalized Functions, Vol I*. Academic Press, New York, 1964.

[20] P. Glaser, M. Arbel, and A. Gretton. KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. In *Advances in Neural Information Processing Systems*, volume 34, pages 8018–8031, 2021.

[21] M. Gräf, D. Potts, and G. Steidl. Quadrature rules, discrepancies and their relations to halftoning on the torus and the sphere. *SIAM Journal on Scientific Computing*, 34(5):2760–2791, 2012.

[22] L. Grafakos and G. Teschl. On Fourier transforms of radial functions and distributions. *Journal of Fourier Analysis and Applications*, 19(1):167–179, 2012.

[23] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(25):723–773, 2012.

[24] T. S. Gutleb, J. A. Carrillo, and S. Olver. Computation of power law equilibrium measures on balls of arbitrary dimension. *arXiv:2109.00843v1*, 2021.

[25] P. Hagemann, J. Hertrich, F. Altekrüger, R. Beinert, J. Chemseddine, and G. Steidl. Posterior sampling based on gradient flows of the MMD with negative distance kernel. *International Conference on Learning Representations (ICLR)*, 2024.

[26] J. Hertrich. Fast kernel summation in high dimensions via slicing and Fourier transforms. *SIAM Journal on Mathematics of Data Science*, 6:1109–1137, 2024.

[27] J. Hertrich, M. Gräf, R. Beinert, and G. Steidl. Wasserstein steepest descent flows of discrepancies with Riesz kernels. *Journal of Mathematical Analysis and Applications*, 531(1, Part 1):127829, 2024.

[28] J. Hertrich, T. Jahn, and M. Quellmalz. Fast summation of radial kernels via QMC slicing. *International Conference on Learning Representations (ICLR)*, 2025.

[29] J. Hertrich and S. Neumayer. Generative feature training of thin 2-layer networks. *Transactions on Machine Learning*, accepted 2025.

[30] J. Hertrich, C. Wald, F. Altekrüger, and P. Hagemann. Generative sliced MMD flows with riesz kernels. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[31] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[32] A. Korba, P. Aubin-Frankowski, S. Majewski, and P. Ablin. Kernel Stein discrepancy descent. In *Proceedings of the 38-th International Conference on Machine Learning (ICML)*, pages 5719–5730. PMLR, 2021.

[33] F. Krahmer and A. Veselovska. Enhanced digital halftoning via weighted sigma-delta modulation. *SIAM Journal on Imaging Sciences*, 16(3):1727–1761, 2023.

[34] W. R. Madych and S. A. Nelson. Multivariate interpolation and conditionally positive definite functions. II. *Mathematics of Computation*, 54(189):211–230, 1990.

[35] R. G. Medhurst and J. H. Roberts. Evaluation of the integral $I_n(b) = \frac{2}{\pi}\int_0^\infty (\frac{\sin x}{x})^n \cos(bx)dx$. *Mathematics of Computation*, 19(90):123–126, 1965.

[36] C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[37] T. Modeste and C. Dombry. Characterization of translation invariant MMD on $R^d$ and connections with Wasserstein distances. *Journal of Machine Learning Research*, 25(237):1–39, 2024.

[38] J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[39] Y. Nemmour, H. Kremer, B. Schölkopf, and J.-J. Zhu. Maximum mean discrepancy distributionally robust nonlinear chance-constrained optimization with finite-sample guarantee. In *61st IEEE Conference on Decision and Control (CDC)*. 2022.

[40] Y. Nesterov. *Lectures on Convex Optimization*. Springer, Cham, 2018.

[41] S. Neumayer, V. Stein, G. Steidl, and N. Rux. Wasserstein gradient flows for Moreau envelopes of f-divergences in reproducing kernel Hilbert spaces. *Analysis and Applications*, 2025.

[42] N. Nüsken and D. M. Renger. Stein variational gradient descent: many-particle and long-time asymptotics. *Foundations of Data Science*, 5(3), 2023.

[43] G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier Analysis*. Springer, second edition, 2023.

[44] D. L. Ragozin. Rotation invariant measure algebras on Euclidean space. *Indiana University Mathematics Journal*, 23(12):1139–54, 1974.

[45] A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in Neural Information Processing Systems*, 20, 2007.

[46] N. Rux, M. Quellmalz, and G. Steidl. Slicing of radial functions: a dimension walk in the Fourier space. *Sampling Theory, Signal Processing, and Data Analysis*, 23(6), 2025.

[47] F. Santambrogio. *Optimal Transport for Applied Mathematicians*. Birkhäuser, Basel, 2015.

[48] C. Schmaltz, P. Gwosdek, A. Bruhn, and J. Weickert. Electrostatic halftoning. *Computer Graphics Forum*, 29(8):2313–2327, 2010.

[49] I. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics*, 4:45–99, 1946.

[50] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263 – 2291, 2013.

[51] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, fourth edition, 2009.

[52] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

[53] I. Steinwart and J. Fasciati-Ziegel. Strictly proper kernel scores and characteristic kernels on compact spaces. *Applied and Computational Harmonic Analysis*, 51:510–542, 2021.

[54] X. Sun. Conditionally positive definite functions and their application to multivariate interpolations. *Journal of Approximation Theory*, 74(2):159–180, 1993.

[55] G. Székely. E-statistics: The energy of statistical samples. *Techical Report, Bowling Green University*, 2002.

[56] C. Villani. *Optimal Transport: Old and New*. Springer, Berlin, 2009.

[57] H. Wendland. *Scattered Data Approximation*. Cambridge University Press, 2004.

[58] J.-J. Zhu and A. Mielke. Kernel approximation of Fisher–Rao gradient flows. *Arxiv prepint 2410.20622*, 2024.

# A Fourier Transform of Tempered Distributions

Let $\mathcal{S}'(\mathbb{R}^d)$ denote the space of tempered distributions, i.e., of linear functionals $T$ on $\mathcal{S}(\mathbb{R}^d)$ fulfilling

$$\varphi_k \underset{\mathcal{S}}{\to} \varphi \quad \Longrightarrow \quad \lim_{k\to\infty} \langle T, \varphi_k \rangle = \langle T, \varphi \rangle,$$

where $\underset{\mathcal{S}}{\to}$ denotes the convergence with respect to

$$\|\varphi\|_m := \max_{|\beta|\le m} \|(1 + \|x\|_2)^m D^\beta \varphi(x)\|_{\mathcal{C}_0(\mathbb{R}^d)} \quad \text{for all} \quad m \in \mathbb{N}_0.$$

In particular, $\mathcal{S}'(\mathbb{R}^d)$ contains all slowly increasing functions $f$, i.e. the functions fulfilling $|f(x)| \le C(1 + \|x\|^N)$ for some $N \in \mathbb{N}_0$ and all functions in $L^p(\mathbb{R}^d)$, $p \in [1, \infty)$. As usual, for distributions of function type, the distribution $T_f$ is identified with the function itself and the dual pairing becomes

$$\langle T_f, \varphi \rangle = \int_{\mathbb{R}^d} f\, \varphi \,\mathrm{d}x \quad \text{for all} \quad \varphi \in \mathcal{S}(\mathbb{R}^d).$$

The Fourier transform $\mathcal{F} : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$, $T \mapsto \hat{T}$ is defined by

$$\langle T, \hat{\varphi} \rangle = \langle \hat{T}, \varphi \rangle \quad \text{for all} \quad \varphi \in \mathcal{S}(\mathbb{R}^d).$$

In particular, we have for $f \in L^1(\mathbb{R}^d)$ in the above sense that $\hat{T}_f = T_{\hat{f}}$ with $\hat{f}$ given by (1).

**Example A.1** (Distributional versus Generalized Fourier transform of polynomials)**.** *Let* $p \colon \mathbb{R} \to \mathbb{R}$, $p(x) := \sum_{k=0}^{r-1} p_k x^k$. *Then*

$$\int_{\mathbb{R}^d} p(x)\hat{\varphi}(x) \,\mathrm{d}x = 0 \quad \textit{for all} \quad \varphi \in \mathcal{S}_r,$$

*so the Generalized Fourier transform of p of order r is the zero function, see [57, Prop. 8.10]. In contrast, the distributional Fourier transform of p is given by*

$$\hat{p} = \sum_{k=0}^{r-1} \left(\frac{i}{2\pi}\right)^k p_k \delta^{(k)}.$$

*If we test only against functions in* $\mathcal{S}_r$ *both approaches coincide.*

**Example A.2** (Distributional Fourier transform of abs)**.** *Since* abs *is slowly increasing, it is a tempered distribution. Its distributional Fourier transform can be written as the distributional derivative of the Cauchy principal value,*

$$\widehat{\mathrm{abs}} = \frac{1}{2\pi^2} \left(\mathrm{pv}\left(\frac{1}{\cdot}\right)\right)',$$

*where*

$$\left\langle \mathrm{pv}\left(\frac{1}{\cdot}\right), \varphi \right\rangle := \lim_{\varepsilon \searrow 0} \int_{|x|>\varepsilon} \frac{\varphi(x)}{x}\,\mathrm{d}x = \int_{\mathbb{R}} \frac{\varphi(x)-\varphi(0)}{x}\,\mathrm{d}x, \qquad \varphi \in \mathcal{S}(\mathbb{R}),$$

*see [43, Sect 4.3], [19]. This can also be represented as the so-called Hadamard finite part* $\mathrm{H}\left(\frac{-1}{2\pi^2(\cdot)^2}\right)$*, see [13], given by*

$$\left\langle \widehat{\mathrm{abs}}, \varphi \right\rangle = \left\langle \mathrm{H}\left(\frac{-1}{2\pi^2(\cdot)^2}\right), \varphi \right\rangle := -\int_{\mathbb{R}} \frac{\varphi(\omega)-\varphi(0)-\varphi'(0)\omega}{2\pi^2\omega^2}\,\mathrm{d}\omega.$$

*If we test only against functions from* $\varphi \in \mathcal{S}_2(\mathbb{R})$*, we have* $\varphi(0) = \varphi'(0) = 0$*, so that this coincides with the generalized Fourier transform* (3). □

Another special case of tempered distributions are finite Borel measures $\mathcal{M}(\mathbb{R}^d)$, see [43, Sect. 4.4]. More precisely, since $\mathcal{S}(\mathbb{R}^d)$ is a dense subspace of $\big(\mathcal{C}_0(\mathbb{R}^d), \|\cdot\|_\infty\big)$, we know by the Riesz representation theorem that $\mu \in \mathcal{M}(\mathbb{R}^d)$ can be identified with a tempered distribution $T_\mu \colon \mathcal{S}(\mathbb{R}^d) \to \mathbb{C}$ which acts on any $\varphi \in \mathcal{S}(\mathbb{R}^d)$ by

$$\langle T_\mu, \varphi \rangle := \int_{\mathbb{R}^d} \varphi\,\mathrm{d}\mu.$$

The *Fourier transform* on $\mathcal{M}(\mathbb{R}^d)$ is defined by $\mathcal{F} \colon \mathcal{M}(\mathbb{R}^d) \to \mathcal{C}_b(\mathbb{R}^d)$ with

$$\mathcal{F}\mu(\omega) = \hat{\mu}(\omega) := \int_{\mathbb{R}^d} \mathrm{e}^{-2\pi i \omega \cdot}\,\mathrm{d}\mu, \qquad \omega \in \mathbb{R}^d, \tag{28}$$

and we have $\hat{T}_\mu = T_{\hat{\mu}}$. For positive measures $\mu \in \mathcal{M}(\mathbb{R}^d)$, i.e. $\mu(B) \ge 0$ for all Borel sets $B \subseteq \mathbb{R}^d$, we obtain a one-to-one mapping to positive definite functions by Bochner's theorem.

**Theorem A.3** (Bochner)**.** *Any positive definite function* $f \colon \mathbb{R}^d \to \mathbb{C}$ *is the Fourier transform of a positive measure and conversely. If in addition* $f(0) = 1$*, then it is the Fourier transform of a probability measure.*

# B Relation with Moreau Envelopes

For a proper, convex, lower semi-continuous function $g \colon \mathbb{R}^d \to \mathbb{R}$ and $\lambda > 0$, the *proximal function* $\mathrm{prox} \colon \mathbb{R}^d \to \mathbb{R}^d$ is defined by

$$\mathrm{prox}_{\lambda g}(x) = \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x-y\|^2 + \lambda g(y) \right\}$$

and its *Moreau envelope* by

$$H_{\lambda g}(x) = \min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x-y\|^2 + \lambda g(y) \right\}.$$

The Moreau envelope is differentiable and

$$\nabla H_{\lambda g}(x) = x - \mathrm{prox}_{\lambda g}(x),$$

so that

$$\mathrm{prox}_{\lambda g}(x) = x - \nabla H_{\lambda g}(x) = \nabla \underbrace{\left(\frac{1}{2}\|x\|^2 - H_{\lambda g}(x)\right)}_{\psi(x)}. \tag{29}$$

Conversely, we have the following result of Moreau [38, Cor 10c].

**Proposition B.1.** *A function $G\colon \mathbb{R}^d \to \mathbb{R}^d$ is the proximal function of a proper, convex, lower semi-continuous function if and only if i) there exists a convex differentiable function $\psi$ such that $G = \nabla\psi$, and ii) $G$ is nonexpansive, i.e., $\|G(x) - G(y)\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^d$.*

In particular, we obtain for $g = \mathrm{abs}$ that

$$\mathrm{prox}_{\lambda\,\mathrm{abs}}(x) = \begin{cases} x - \lambda & x > \lambda, \\ 0 & x \in [-\lambda, \lambda], \\ x + \lambda & x < -\lambda, \end{cases}$$

and the Moreau envelope

$$H_{\lambda\,\mathrm{abs}}(x) = \begin{cases} \frac{1}{2}x^2 & |x| \leq \lambda \\ \lambda(|x| - \frac{\lambda}{2}) & \text{otherwise} \end{cases}$$

is known as the *Huber function*.

On the other hand, we have

$$(\mathrm{abs} * M_1)(x) = \begin{cases} x^2 + \frac{1}{4} & |x| \leq \frac{1}{2}, \\ |x| & |x| > \frac{1}{2}, \end{cases}$$

so that by Proposition 3.1, for $\varepsilon = 2\lambda$,

$$(\mathrm{abs} * M_{1,2\lambda})(x) = \frac{1}{\lambda} H_{\lambda\,\mathrm{abs}}(x) + \frac{\lambda}{2} = \begin{cases} \frac{x^2}{2\lambda} + \frac{\lambda}{2} & |x| \leq \lambda, \\ |x| & |x| > \lambda. \end{cases}$$

Thus, by (3) and Propositions 2.2 and 3.2, the Generalized Fourier transform of the Huber function is given by

$$\hat{H}_{\lambda\,\mathrm{abs}}(\omega) = -\frac{\lambda}{2\pi^2\omega^2}\,\mathrm{sinc}(2\lambda\omega). \tag{30}$$

Since this function has positive and negative values, we conclude by Proposition 2.3 that the negative Huber function is not conditionally positive definite of any order.

Let us see if abs $*M_{1,\varepsilon}$ is the Moreau envelope of some function. Regarding (29), we consider

$$\psi(x) := \frac{1}{2}x^2 - (\text{abs} * M_{1,2\lambda})(x) = \begin{cases} \frac{1}{2}\left(1 - \frac{1}{\lambda}\right)x^2 - \frac{\lambda}{2} & |x| \le \lambda, \\ \frac{x^2}{2} - |x| & |x| > \lambda, \end{cases}$$

which is convex for $\lambda \ge 1$ and has a nonexpansive derivative

$$\psi'(x) = x - (\text{abs} * M_{1,\varepsilon})'(x) = \begin{cases} \left(1 - \frac{1}{\lambda}\right)x & |x| \le \lambda, \\ x - 1 & x > \lambda, \\ x + 1 & x < -\lambda. \end{cases}$$

Thus, by Moreau's Proposition B.1, we see that abs $*M_{1,2\lambda}$ is a Moreau envelope if and only if $\varepsilon = 2\lambda \ge 2$. More general, we have the following proposition

**Proposition B.2.** *For $m \in \mathbb{N}$, the function* abs $*M_{m,\varepsilon}$ *is the Moreau envelope of a proper, convex, lower semi-continuous function if and only if $\varepsilon \ge 2M_m(0)$.*

*Proof.* By the above considerations, the assertion is true for $m = 1$. By Proposition 3.1, the function

$$\psi(x) := \frac{1}{2}x^2 - (\text{abs} * M_{m,\varepsilon})(x)$$

fulfills

$$\psi''(x) = 1 - \frac{2}{\varepsilon}M_m\left(\frac{x}{\varepsilon}\right) \ge 1 - \frac{2}{\varepsilon}M_m(0) \ge 0$$

if and only if $\varepsilon \ge 2M_m(0)$, and exactly in this case $\psi$ is convex. Further, because $\psi'' \le 1$, we see that $\psi'$
is nonexpansive and by Moreau's Proposition B.1, the function abs $*M_{m,\varepsilon}$ is a Moreau envelope. □

# C Proofs

## Proofs from Section 2

**Proof of Proposition 2.2.** Since $f \in \mathcal{C}(\mathbb{R}^d)$ is slowly increasing and $u \in \mathcal{C}_c(\mathbb{R}^d)$, we conclude by straightforward computations that $f * u$ is continuous and slowly increasing, too. Therefore, $\langle f * u, \hat{\varphi} \rangle$ exists for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$. Using Fubini's theorem, we obtain

$$\begin{aligned} \langle f * u, \hat{\varphi} \rangle &= \int_{\mathbb{R}^d} (f * u)(x)\hat{\varphi}(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(y)f(x-y)\,\mathrm{d}y\hat{\varphi}(x)\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} u(y)\int_{\mathbb{R}^d} f(x-y)\hat{\varphi}(x)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned}$$

By the translation-modulation theorem, we know that

$$\hat{\varphi}(x) = \mathcal{F}[\mathrm{e}^{-2\pi\mathrm{i}\langle\cdot,y\rangle}\varphi](x-y),$$

so that

$$\begin{aligned}\langle f * u, \hat{\varphi}\rangle &= \int_{\mathbb{R}^d} u(y) \int_{\mathbb{R}^d} f(x-y)\mathcal{F}[\mathrm{e}^{-2\pi\mathrm{i}\langle\cdot,y\rangle}\varphi](x-y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{\mathbb{R}^d} u(y) \int_{\mathbb{R}^d} f(x)\mathcal{F}[\mathrm{e}^{-2\pi\mathrm{i}\langle\cdot,y\rangle}\varphi](x)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Since $f$ has a generalized Fourier transform $\hat{f}$ of order $r$, this implies for $\varphi \in \mathcal{S}_{2r}(\mathbb{R}^d)$ that

$$\begin{aligned}\langle f * u, \hat{\varphi}\rangle &= \int_{\mathbb{R}^d} \hat{f}(x)\varphi(x) \int_{\mathbb{R}^d} u(y)\mathrm{e}^{-2\pi\mathrm{i}\langle x,y\rangle}\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{\mathbb{R}^d} \hat{f}(x)\varphi(x)\hat{u}(x)\,\mathrm{d}x.\end{aligned}$$

Hence, $f * u$ has a generalized Fourier transform of order $r$, namely $\hat{f}\,\hat{u}$. □

## Proofs from Section 3

**Proof of Proposition 3.1.** i) follows directly by definition of $f$ and since $u$ is even.

To show ii), let $x > R := \operatorname{diam}(\operatorname{supp} u)/2$. The case $x < -R$ follows similarly. Then we obtain

$$(\operatorname{abs} * u)(x) = \int_{-R}^{R} u(y)(x-y)\,\mathrm{d}y = x\int_{-R}^{R} u(y)\,\mathrm{d}y - \int_{-R}^{R} y\,u(y)\,\mathrm{d}y = x\cdot 1 - 0 = x.$$

In iii), we only have to show that $f'' = 2u$. Then the smoothness of $f$ follows by $u \in \mathcal{C}^n(\mathbb{R})$. Using Lebesgue's dominated convergence theorem, we conclude

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}(\operatorname{abs} * u)(x) &= \lim_{h\to 0}\int_{-R}^{R} \underbrace{\frac{|x+h-y|-|x-y|}{h}u(y)}_{|g_{x,h}|\le u}\,\mathrm{d}y = \int_{-R}^{R} \operatorname{sgn}(x-y)u(y)\,\mathrm{d}y \\ &= (\operatorname{sgn} * u)(x),\end{aligned}$$

where

$$\operatorname{sgn}(x) := \begin{cases} 1 & x \ge 0, \\ -1 & x < 0. \end{cases}$$

The right derivative of sgn is given by

$$\lim_{h\searrow 0}\frac{\operatorname{sgn}(x+h)-\operatorname{sgn}(x)}{h} = \lim_{h\searrow 0}\frac{2}{h}\mathbb{1}_{[-h,0)}(x).$$

Therefore, we have by continuity of $u$ that

$$\begin{aligned}\lim_{h\searrow 0}\frac{(\operatorname{sgn} * u)(x+h)-(\operatorname{sgn} * u)(x)}{h} &= \lim_{h\searrow 0}\int_{\mathbb{R}}\frac{2}{h}\mathbb{1}_{[-h,0)}(y)u(x-y)\,\mathrm{d}y \\ &= 2\lim_{h\searrow 0}\frac{1}{h}\int_{-h}^{0} u(x-y)\,\mathrm{d}y = 2u(x).\end{aligned}$$

We obtain the same result for the left derivative. Since $f'' = 2u \geq 0$, the function $f$ is convex.

For iv), we have by Lemma 2.2 and (3) that

$$-\hat{f} = \mathcal{F}[-\operatorname{abs} * u] = (-\widehat{\operatorname{abs}})\,\hat{u} \geq 0.$$

Therefore $-f$ is conditionally positive definite of order $r = 1$ by Theorem 2.3.

Assertion v) follows by straightforward computation.

Finally, we show vi). Note that we cannot apply the usual convergence theorems for approximate identities in vi), because abs $\notin C_0(\mathbb{R})$.

Let $\operatorname{supp} u \subseteq [-R, R]$ for some $R > 0$. Then, we have by v) and ii) that

$$(\operatorname{abs} * u_\varepsilon)(x) = |x| \quad \text{for} \quad |x| \geq \varepsilon R.$$

For $|x| \leq \varepsilon R$, we conclude using $\int_{\mathbb{R}} u_\varepsilon(y)\,\mathrm{d}y = 1$ that

$$\begin{aligned}
\big||x| - (\operatorname{abs} * u_\varepsilon)(x)\big| &= \left| \int_{\mathbb{R}} |x| u_\varepsilon(x-y)\,\mathrm{d}y - \int_{\mathbb{R}} |y| u_\varepsilon(x-y)\,\mathrm{d}y \right| \\
&\leq \int_{\mathbb{R}} |x-y| u_\varepsilon(x-y)\,\mathrm{d}y = \int_{-\varepsilon R}^{\varepsilon R} |y|\, u_\varepsilon(y)\,\mathrm{d}y \\
&\leq \varepsilon R \int_{-\varepsilon R}^{\varepsilon R} u_\varepsilon(y)\,\mathrm{d}y = \varepsilon R,
\end{aligned}$$

which shows the uniform convergence. □

**Proof of Corollary 3.3.** By (5) and since $f'' = 2M_m$, we obtain

$$f(x) = \frac{2}{(m+1)!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} \left(x - k + \frac{m}{2}\right)_+^{m+1} + ax + b$$

with some $a, b \in \mathbb{R}$. Further, for $x \leq -\frac{m}{2}$, we conclude by $f(x) = -x$ that

$$f(x) = ax + b = -x,$$

so that $a = -1$ and $b = 0$. □

## Proofs from Section 4

**Proof of Lemma 4.1.** The function $f := \operatorname{abs} * M_2$ has the generalized Fourier transform of order 1 given by $\hat{f}(r) = -\frac{\operatorname{sinc}^2(r)}{2\pi^2 r^2}$. We define

$$g(r) := -\frac{1}{2\pi} \frac{1}{r} \frac{\mathrm{d}}{\mathrm{d}r} \hat{f}(r).$$

For integrable functions it was proven in [22, Thm. 1.1] that $g(\|\cdot\|)$ is the 3-dimensional Fourier transform of $f(\|\cdot\|)$. However, since $f$ is not integrable, we use the generalized

Fourier transform to argue that $\mathcal{F}_3[f(\|\cdot\|)] = g(\|\cdot\|)$: for an even test function $\psi \in \mathcal{S}_2(\mathbb{R})$, we apply [22, Thm. 1.1] to obtain

$$\mathcal{F}_3[\psi(\|\cdot\|)](\|x\|) = -\frac{1}{2\pi}\frac{1}{\|x\|}\hat{\psi}'(\|x\|),$$

and with the surface area $\omega_2 = 4\pi$, integration by parts gives

$$\begin{aligned}\int_{\mathbb{R}^3} g(\|x\|)\psi(\|x\|)\,\mathrm{d}x &= \omega_2\int_0^\infty g(r)\psi(r)r^2\,\mathrm{d}r = -\int_0^\infty 2\hat{f}'(r)\psi(r)r\,\mathrm{d}r\\ &= -\Big[2\hat{f}(r)\psi(r)r\Big]_0^\infty + \int_0^\infty 2\hat{f}(r)\frac{\mathrm{d}}{\mathrm{d}r}(\psi(r)r)\,\mathrm{d}r.\end{aligned}$$

Since $\psi \in \mathcal{S}_2(\mathbb{R})$, the first summand $[2\hat{f}(r)\psi(r)r]_0^\infty$ vanishes. The derivative of the odd function $\psi(r)r$ is even and still in $\mathcal{S}_2(\mathbb{R})$. Thus, we get by (2) that

$$\begin{aligned}\int_{\mathbb{R}^3} g(\|x\|)\psi(\|x\|)\,\mathrm{d}x &= \int_{\mathbb{R}} \hat{f}(r)\frac{\mathrm{d}}{\mathrm{d}r}(\psi(r)r)\,\mathrm{d}r = -\int_{\mathbb{R}} f(r)r\frac{\mathrm{d}}{\mathrm{d}r}\hat{\psi}(r)\,\mathrm{d}r\\ &= 4\pi\int_0^\infty f(r)\frac{-1}{2\pi r}\frac{\mathrm{d}}{\mathrm{d}r}\hat{\psi}(r)r^2\,\mathrm{d}r = \int_{\mathbb{R}^3} f(\|x\|)\mathcal{F}_3[\psi(\|\cdot\|)](\|x\|)\,\mathrm{d}x,\end{aligned}$$

i.e. $\mathcal{F}_3[f(\|\cdot\|)] = g(\|\cdot\|)$. Next we show that for all test functions $\varphi \in \mathcal{S}_2(\mathbb{R}^3)$ it holds

$$\int_{\mathbb{R}^3} g(\|x\|)\varphi(x)\,\mathrm{d}x = \int_{\mathbb{R}^3} f(\|x\|)\hat{\varphi}(x)\,\mathrm{d}x.$$

For an arbitrary $\varphi \in \mathcal{S}(\mathbb{R}^d)$, define the radial test function

$$\operatorname{Rad}\varphi(x) := \frac{1}{\omega_{d-1}}\int_{\mathbb{S}^{d-1}} \varphi(\|x\|\xi)\,\mathrm{d}\xi.$$

In [46, Thm. 4.2 i)] it was shown, that Rad is a continuous projection of $\mathcal{S}(\mathbb{R}^d)$ to the space of radial Schwartz functions $\mathcal{S}_{\mathrm{rad}}(\mathbb{R}^d)$. By the uniqueness of rotational invariant measures on the sphere, see [44, (2.3)], we have

$$\operatorname{Rad}\varphi(x) = \int_{SO(d)} f(Rx)\,\mathrm{d}\mathcal{U}_{SO(d)}(R),$$

where $\mathcal{U}_{SO(d)}$ is the uniform measure on the set $SO(d)$ of $d \times d$ rotation matrices. Since the Fourier transform commutes with rotations, we have $\operatorname{Rad}[\mathcal{F}\varphi] = \mathcal{F}[\operatorname{Rad}\varphi]$ for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$, so the operators $\mathcal{F}$ and Rad commute. It is easy to see that for $\varphi \in \mathcal{S}_2(\mathbb{R}^3)$, we also have $\operatorname{Rad}\varphi \in \mathcal{S}_2(\mathbb{R}^3)$. The action of the test functions $\varphi$ and $\operatorname{Rad}\varphi$ on $g(\|\cdot\|)$ is the same, because $g(\|\cdot\|)$ is radial. Hence we have for all $\varphi \in \mathcal{S}_2(\mathbb{R}^3)$ that

$$\begin{aligned}\int_{\mathbb{R}^3} g(\|x\|)\varphi(x)\,\mathrm{d}x &= \langle g(\|\cdot\|),\varphi\rangle = \langle g(\|\cdot\|),\operatorname{Rad}\varphi\rangle = \langle f(\|\cdot\|),\mathcal{F}[\operatorname{Rad}\varphi]\rangle\\ &= \langle f(\|\cdot\|),\operatorname{Rad}\hat{\varphi}]\rangle = \langle f(\|\cdot\|),\hat{\varphi}\rangle = \int_{\mathbb{R}^3} f(\|x\|)\hat{\varphi}(x)\,\mathrm{d}x.\end{aligned}$$

Consequently, $f(\|\cdot\|)$ has the generalized Fourier transform $g(\|\cdot\|)$ of order 1. Theorem 2.3 shows that $-f(\|\cdot\|) \notin \mathrm{CP}_1(\mathbb{R}^3)$, because $g$ changes its sign. Since $g(\|\cdot\|)$ is the generalized Fourier transform of $f(\|\cdot\|)$ for all $r \geq 1$, we see that $-f(\|\cdot\|) \notin \mathrm{CP}_r(\mathbb{R}^3)$ for all $r \geq 1$. Since $-f(\|\cdot\|) \notin \mathrm{CP}_1(\mathbb{R}^3)$, we have by [57, Prop. 8.2] that $-f(\|\cdot\|) \notin \mathrm{CP}_r(\mathbb{R}^d)$ for all $r \geq 0$ and all $d \geq 3$. □

**Proof of Proposition 4.3.** Assume that $f \in \mathrm{CP}_r(\mathbb{R})$, then $f(x) \in \mathcal{O}(|x|^{2r})$, by [34, Cor 2.3]. The function $F$ is well-defined, because $f$ is continuous and is slowly increasing. Let $x_1, \dots, x_N \in \mathbb{R}^d$ and $a_1, \dots, a_n \in \mathbb{R}$ such that

$$\sum_{j=1}^{N} a_j P(x_j) = 0 \tag{31}$$

for all polynomials $P$ on $\mathbb{R}^d$ of degree $< r$. In particular, any polynomial $p(t) = \sum_{k=0}^{r-1} c_k t^k$ on $\mathbb{R}$ determines for an arbitrary fixed $\xi \in \mathbb{S}^{d-1}$, a polynomial on $\mathbb{R}^d$ of degree $< r$ by

$$P_\xi(x) := p(\langle \xi, x\rangle) = \sum_{k=0}^{r-1} c_k \langle \xi, x\rangle^k = \sum_{k=0}^{r-1} c_k \left( \sum_{l=1}^{d} \xi_l x_l \right)^k .$$

By (31), we have

$$\sum_{j=1}^{N} a_j P_\xi(x_j) = \sum_{j=1}^{N} a_j p(\langle \xi, x_j\rangle) = 0.$$

Since $f$ is conditionally positive definite of order $r$, we know that

$$0 \leq \sum_{j,k=1}^{N} a_j a_k f(|\langle \xi, x_j\rangle - \langle \xi, x_k\rangle|),$$

so that by Theorem 4.2 also

$$0 \leq \frac{1}{w_{d-1}} \int_{\mathbb{S}^{d-1}} \sum_{j,k=1}^{N} a_j a_k f(|\langle \xi, x_j - x_k\rangle|) \,\mathrm{d}\xi \,\mathrm{d}\xi = \sum_{j,k=1}^{N} a_j a_k F(\|x_j - x_k\|).$$

Hence $F(\|\cdot\|)$ is conditionally positive definite of order $r$. □

**Proof of Proposition 4.4.** In [46, Eq. (6) & (7)], two operators were introduced: the rotation operator $\mathcal{R}_d$ acts on a function $F\colon [0,\infty) \to \mathbb{R}$ as $\mathcal{R}_d F(x) := F(\|x\|)$, and the spherical averaging operator $\mathcal{A}_d$ assigns to a function $\Phi\colon \mathbb{R}^d \to \mathbb{R}$ integrable on the spheres $t\,\mathbb{S}^{d-1}$ for all $t > 0$ the function

$$\mathcal{A}_d \Phi(t) := \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} \Phi(t\xi) \,\mathrm{d}\xi.$$

For $d = 1$, the spherical averaging operator reduces to $\mathcal{A}_1 \Phi(t) = \frac{1}{2}(\Phi(t) + \Phi(-t))$. Moving to distributions, the operator $\mathcal{R}_d^\star$ acts on a tempered distribution $T$ as

$$\langle \mathcal{R}_d^\star T, \psi\rangle = \langle T, (\mathcal{R}_d \circ \mathcal{A}_1)\psi\rangle \quad \text{for all} \quad \psi \in \mathcal{S}(\mathbb{R}).$$

Since $F(\|\cdot\|)$ is continuous and slowly increasing, it can be identified with a tempered distribution. Let $\mathcal{F}_d$ denote the Fourier transform of tempered distributions. Since $F$ is $\lfloor \frac{d}{2} \rfloor$-times continuously differentiable, we have by [46, Cor. 4.9] that

$$f := (\mathcal{F}_1 \circ \mathcal{R}_d^\star \circ \mathcal{F}_d^{-1})[\mathcal{R}_d F]$$

is a distribution arising from a continuous, even function which satisfies $\mathcal{I}_d[f] = F$.

Let $\psi \in \mathcal{S}_{2r}(\mathbb{R})$ be an even. Then $(\mathcal{R}_d \circ \mathcal{A}_1)\psi = \psi(\|\cdot\|)$ is a radial Schwartz function in $\mathcal{S}_{2r}(\mathbb{R}^d)$. Since $\mathcal{R}_d F$ has a Generalized Fourier transform $\rho(\|\cdot\|)$ of order $r$, we obtain

$$\begin{aligned}
\int_{\mathbb{R}} f(r)\hat{\psi}(r)\,\mathrm{d}r &= \langle f, \hat{\psi}\rangle = \langle \hat{f}, \psi\rangle = \langle (\mathcal{R}_d^\star \circ \mathcal{F}_d^{-1})[\mathcal{R}_d F], \psi\rangle = \langle \mathcal{R}_d F, (\mathcal{F}_d^{-1} \circ \mathcal{R}_d \circ \mathcal{A}_1)\psi\rangle \\
&= \int_{\mathbb{R}^d} F(\|x\|)\mathcal{F}_d[\psi(\|\cdot\|)](x)\,\mathrm{d}x = \int_{\mathbb{R}^d} \rho(\|x\|)\psi(\|x\|)\,\mathrm{d}x \\
&= \frac{w_{d-1}}{2} \int_{\mathbb{R}} \rho(\omega)|\omega|^{d-1}\psi(\omega)\,\mathrm{d}\omega.
\end{aligned}$$

In particular, $f$ has the generalized Fourier transform $\frac{w_{d-1}}{2}\rho(\omega)|\omega|^{d-1} \in \mathcal{C}(\mathbb{R} \setminus \{0\})$ of order $r$, which is nonnegative, so that $f$ is conditionally positive definite of order $r$. $\square$

**Proof of Proposition 4.6.** For $d \geq 2$, the term $(1-t^2)^{\frac{d-3}{2}}$, $t \in [0,1]$ is integrable. Since $f = \operatorname{abs} * u_\varepsilon \in \mathcal{C}^n(\mathbb{R}) \subseteq L^\infty_{\mathrm{loc}}(\mathbb{R})$, $n \in \mathbb{N}_0$, the function $F = \mathcal{I}_d[f]$ is well-defined. By Proposition 3.1, we know that $f$ is nonnegative and even. Hence, also $F$ is nonnegative and even. By Leibniz's integral rule and since $f \in \mathcal{C}^{n+1}(\mathbb{R})$, we obtain for $k = 1, \dots, n+2$ that

$$\begin{aligned}
\frac{\mathrm{d}^k}{\mathrm{d}s^k} F(s) &= \frac{\mathrm{d}^k}{\mathrm{d}s^k} c_d \int_0^1 f(ts)(1-t^2)^{\frac{d-3}{2}}\,\mathrm{d}t = c_d \int_0^1 \frac{\mathrm{d}^k}{\mathrm{d}s^k} f(ts)(1-t^2)^{\frac{d-3}{2}}\,\mathrm{d}t \\
&= c_d \int_0^1 t^k f^{(k)}(ts)(1-t^2)^{\frac{d-3}{2}}\,\mathrm{d}t,
\end{aligned}$$

so that $F \in \mathcal{C}^{n+2}(\mathbb{R})$. Since $f$ is at least twice differentiable and $f(t) = \operatorname{abs}(t)$ for $|t|$ large enough, it follows, that $\|f'\|_\infty < \infty$. For the first derivative of $F$ we get

$$|F'(s)| = \left| c_d \int_0^1 t f'(ts)(1-t^2)^{\frac{d-3}{2}}\,\mathrm{d}t \right| \leq c_d \|(1-t^2)^{\frac{d-3}{2}}\|_{L^1(0,1)} \|f'\|_\infty < \infty.$$

The convexity of $F$ follows directly from the convexity of $f$.

Let $\operatorname{supp}(u) \subseteq [-R, R]$ for some $R > 0$. Then we have by Proposition 3.1 for $s \geq R$ that $f(s) = s$. Further, by Lemma 4.5, it holds $\mathcal{I}_d[\operatorname{abs}] = C_d \operatorname{abs}$. Hence, we obtain for $s > R$ that

$$\begin{aligned}
|F(s) - C_d \operatorname{abs}(s)| &= \left| c_d \int_0^1 (f(st) - \operatorname{abs}(st))(1-t^2)^{\frac{d-3}{2}}\,\mathrm{d}t \right| \\
&\leq c_d \int_0^{\frac{R}{s}} |f(st) - st|(1-t^2)^{\frac{d-3}{2}}\,\mathrm{d}t \\
&= \frac{c_d}{s} \int_0^R |f(t) - t| \left(1 - \frac{t^2}{s^2}\right)^{\frac{d-3}{2}}\,\mathrm{d}t \in \mathcal{O}\left(\frac{1}{s}\right).
\end{aligned}$$

In particular, it holds $h := F - C_d \operatorname{abs} \in \mathcal{C}_0(\mathbb{R}) \cap L^2(\mathbb{R})$.

Finally, we obtain by Proposition 3.1 that

$$\begin{aligned} F^\varepsilon(s) &= c_d \int_0^1 (\operatorname{abs} * u_\varepsilon)(ts)(1-t^2)^{\frac{d-3}{2}}\, \mathrm{d}t \\ &= c_d \varepsilon \int_0^1 f\left(\frac{st}{\varepsilon}\right)(1-t^2)^{\frac{d-3}{2}}\, \mathrm{d}t = \varepsilon F\left(\frac{s}{\varepsilon}\right). \end{aligned}$$

and further

$$\begin{aligned} \|F^\varepsilon - C_d \operatorname{abs}\|^2_{L^2(\mathbb{R})} \leq \int_{\mathbb{R}} |F^\varepsilon(s) - C_d \operatorname{abs}(s)|^2\, \mathrm{d}s &= \varepsilon^2 \int_{\mathbb{R}} \left| F\left(\frac{s}{\varepsilon}\right) - C_d \operatorname{abs}\left(\frac{s}{\varepsilon}\right)\right|^2 \mathrm{d}s \\ &= \varepsilon^2 \int_{\mathbb{R}} |h\left(\frac{s}{\varepsilon}\right)|^2\, \mathrm{d}s = \varepsilon^3 \|h\|^2_{L_2(\mathbb{R})}. \end{aligned}$$

This gives us the order of convergence in $L_2(\mathbb{R})$. The pointwise convergence of $F^\varepsilon$ directly follows from (7). $\square$

**Proof of Proposition 4.7.** By Corollary 3.3, we have

$$f(x) = \frac{2}{(m+1)!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} \left(x - k + \frac{m}{2}\right)_+^{m+1} - x, \qquad x \in \mathbb{R}.$$

Defining for $m \in \mathbb{N}$ and $a \in \mathbb{R}$ the function

$$b_{m,a}(x) = (x-a)_+^m,$$

we have

$$f(x) = \frac{2}{(m+1)!} \sum_{k=0}^{m} (-1)^k \binom{m}{k} b_{m+1,k-\frac{m}{2}}(x) - x.$$

If $a \leq 0$, we have $b_{m,a}(x) = (x-a)^m$ for $x \geq 0$. Then

$$\begin{aligned} \mathcal{I}_d[b_{m,k}](s) &= c_d \int_0^1 f_{m,a}(st)(1-t^2)^{\frac{d-3}{2}}\, \mathrm{d}t = c_d \int_0^1 (st-a)^m (1-t^2)^{\frac{d-3}{2}}\, \mathrm{d}t \\ &= c_d \sum_{n=0}^{m} \binom{m}{n} s^n (-a)^{m-n} \int_0^1 t^n (1-t^2)^{\frac{d-3}{2}}\, \mathrm{d}t \\ &= \frac{c_d}{2} \sum_{n=0}^{m} \binom{m}{n} s^n (-a)^{m-n} \int_0^1 t^{\frac{n-1}{2}} (1-t)^{\frac{d-3}{2}}\, \mathrm{d}t. \end{aligned}$$

Since the Beta function satisfies $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, see [1], we obtain

$$\mathcal{I}_d[b_{m,k}](s) = \frac{c_d}{2} \sum_{n=0}^{m} \binom{m}{n} s^n (-a)^{m-n} \frac{\Gamma(\frac{d-1}{2})\Gamma(\frac{n+1}{2})}{\Gamma(\frac{d+n}{2})}.$$

If $a > 0$ and $s \le a$, we have $\mathcal{I}_d[f_{m,a}](s) = 0$. Otherwise, i.e. for $0 < a < s$, we have

$$\begin{aligned}
\mathcal{I}_d[b_{m,a}](s) &= c_d \int_{a/s}^{1} (st-a)^m (1-t^2)^{\frac{d-3}{2}} \,\mathrm{d}t \\
&= \frac{c_d}{2} \sum_{n=0}^{m} \binom{m}{n} s^n (-a)^{m-n} \int_{a^2/s^2}^{1} t^{\frac{n-1}{2}} (1-t)^{\frac{d-3}{2}} \,\mathrm{d}t \\
&= \frac{c_d}{2} \sum_{n=0}^{m} \binom{m}{n} s^n (-a)^{m-n} \left( B(\tfrac{n+1}{2}, \tfrac{d-1}{2}) - B_{a^2/s^2}(\tfrac{n+1}{2}, \tfrac{d-1}{2}) \right).
\end{aligned}$$

The claim follows by collecting the terms and Lemma 4.5. □

**Proof of Theorem 4.9.**

i) Since $f \in \mathrm{CP}_1(\mathbb{R})$ by Proposition 3.1, we obtain by Proposition 4.3 that $\Phi \in \mathrm{CP}_1(\mathbb{R}^d)$.

ii) By (8) we know that

$$\Phi(x) = F(\|x\|) = \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} f(\langle x, \xi \rangle) \,\mathrm{d}\xi \quad \text{for all } x \in \mathbb{R}^d.$$

Hence we get $\Phi(0) = F(0) = f(0) = -(\mathrm{abs} * M_2)(0) < 0$.

iii) By Proposition 4.6 we directly conclude iii).

iv) Since $f$ is $n+2$ times continuously differentiable by Proposition 3.1 iii), we obtain for any multi-index $\alpha \in \mathbb{N}^d$ with $|\alpha| \le n+2$ that

$$\partial^\alpha \Phi(x) = \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} \xi^\alpha f^{|\alpha|}(\langle x, \xi \rangle) \,\mathrm{d}\xi \quad \text{for all } x \in \mathbb{R}^d.$$

v) Since $f'' = 2u$ is bounded, the first derivative $f'$ is $2\|u\|_\infty$-Lipschitz continuous. Hence, for $x, y \in \mathbb{R}^d$, we can estimate by (iv))

$$\begin{aligned}
|\partial_i \Phi(x) - \partial_i \Phi(y)| &\le \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} |\xi_i| \left| f'(\langle x, \xi \rangle) - f'(\langle y, \xi \rangle) \right| \mathrm{d}\xi \\
&\le \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} 2\|u\|_\infty \|x - y\| \,\mathrm{d}\xi = 2\|u\|_\infty \|x - y\|,
\end{aligned}$$

so that we obtain

$$\|\nabla \Phi(x) - \nabla \Phi(y)\|^2 = \sum_{i=1}^{d} |\partial_i \Phi(x) - \partial \Phi(y)|^2 \le d(2\|u\|_\infty \|x - y\|)^2.$$

vi) Since $\nabla \Phi$ is $L$-Lipschitz continuous with $L = \sqrt{d} 2\|u\|_\infty$, we know, by [40, Thm. 2.1.5] that $\Phi$ is $-L$-convex. Furthermore, $\Phi$ is concave because, for $t \in [0,1]$ and $x, y \in \mathbb{R}^d$, it holds by the concavity of $f$ from Proposition 3.1 iii) that

$$\begin{aligned}
\Phi((1-t)x + ty) &\ge \frac{1}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} (1-t) f(\langle x, \xi \rangle) + t f(\langle y, \xi \rangle) \,\mathrm{d}\xi \\
&= (1-t)\Phi(x) + t\Phi(y).
\end{aligned}$$

□

## Proofs of Section 6

**Proof of Proposition 6.1.** Recall that $f = -\operatorname{abs} * u$ is even and $F = \mathcal{I}_d[f]$ satisfies (8). Let $\mu \in \mathcal{M}_{1/2}(\mathbb{R}^d)$, then the following integral exists

$$
\begin{aligned}
\|\mu\|_{\mathcal{H}_K}^2 &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left(F(\|x-y\|) - F(\|x\|) - F(\|y\|)\right)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \frac{1}{\omega_{d-1}}\int_{\mathbb{S}^{d-1}} f(\langle x-y,\xi\rangle) - f(\langle x,\xi\rangle) - f(\langle -y,\xi\rangle) + f(0)\,\mathrm{d}\xi\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&\quad - F(0)|\mu(\mathbb{R}^d)|^2.
\end{aligned}
$$

Denote by $\mathcal{T}_y$ the translation operator $\mathcal{T}_y[g](y) = g(x-y)$, by $\mathcal{M}_y$ the modulation operator $\mathcal{M}_y[g](x) = \mathrm{e}^{-2\pi\mathrm{i}\langle x,y\rangle} g(x)$ and by $g_m(x) = \sqrt{m/\pi}\,\mathrm{e}^{-mx^2}$ the Gaussian approximate identity as in [57, Thm. 5.20]. Since $f$ is continuous and slowly increasing, [57, Thm. 5.20 (4)] yields

$$
f(\langle x-y,\xi\rangle) - f(\langle x,\xi\rangle) - f(\langle -y,\xi\rangle) + f(0) = \lim_{m\to\infty} \langle (\mathrm{id} - \mathcal{T}_{\langle \xi,x\rangle})[(\mathrm{id} - \mathcal{T}_{\langle \xi,-y\rangle})[g_m]], f\rangle.
$$

Let $\varphi_m := (\mathrm{id} - \mathcal{M}_{-\langle \xi,x\rangle})[(\mathrm{id} - \mathcal{M}_{-\langle \xi,-y\rangle})[\hat{g}_m]] \in \mathcal{S}_2(\mathbb{R}^1)$. The function $f$ has the generalized Fourier transform

$$
\hat{f}(r) = \frac{\hat{u}(r)}{2\pi^2 r^2}
$$

of order 1 by Lemma 2.2 and (3). As $g_m$ is even, we have $\hat{\hat{g}}_m = g_m$ and for all $m \in \mathbb{N}$ we can write

$$
\begin{aligned}
\langle (\mathrm{id} - \mathcal{T}_{\langle \xi,x\rangle})[(\mathrm{id} - \mathcal{T}_{\langle \xi,-y\rangle})[g_m]], f\rangle &= \langle \hat{\varphi}_m, f\rangle = \langle \varphi_m, \hat{f}\rangle\\
&= \int_{\mathbb{R}} (1 - \mathrm{e}^{2\pi\mathrm{i}\langle \xi,x\rangle r})(1 - \mathrm{e}^{-2\pi\mathrm{i}\langle \xi,y\rangle r})\hat{g}_m(r)\hat{f}(r)\,\mathrm{d}r.
\end{aligned}
$$

Since $u$ is continuous with compact support, its Fourier transform is bounded. The term $r \mapsto (1 - \mathrm{e}^{2\pi\mathrm{i}\langle \xi,x\rangle r})(1 - \mathrm{e}^{-2\pi\mathrm{i}\langle \xi,y\rangle r})$ is bounded and has a zero of order 2 at zero, so that

$$
\hat{f}(r)(1 - \mathrm{e}^{2\pi\mathrm{i}\langle \xi,x\rangle r})(1 - \mathrm{e}^{-2\pi\mathrm{i}\langle \xi,y\rangle r}) = \frac{\hat{u}(r)}{2\pi^2 r^2}(1 - \mathrm{e}^{2\pi\mathrm{i}\langle \xi,x\rangle r})(1 - \mathrm{e}^{-2\pi\mathrm{i}\langle \xi,y\rangle r})
$$

is integrable. Moreover, we have

$$
|\hat{g}_m(r)| = \mathrm{e}^{-\frac{\pi^2 r^2}{m}} \le 1 \quad \text{for all} \quad r \in \mathbb{R} \text{ and } m \in \mathbb{N}.
$$

Since $\hat{g}_m$ converges pointwise to the constant 1, Lebesgue's convergence theorem yields

$$
f(\langle x-y,\xi\rangle) - f(\langle x,\xi\rangle) - f(\langle -y,\xi\rangle) + f(0) = \int_{\mathbb{R}} (1 - \mathrm{e}^{2\pi\mathrm{i}\langle \xi,x\rangle r})(1 - \mathrm{e}^{-2\pi\mathrm{i}\langle \xi,y\rangle r})\hat{f}(r)\,\mathrm{d}r.
$$

Further, we obtain using Fubini's theorem

$$\begin{aligned}
&\omega_{d-1}(\|\mu\|_{\mathcal{H}_K}^2 + F(0)\mu(\mathbb{R}^d)^2) \\
&= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} \int_{\mathbb{R}^d} (1 - \mathrm{e}^{2\pi \mathrm{i}\langle r\xi, x\rangle}) \,\mathrm{d}\mu(x) \int_{\mathbb{R}^d} (1 - \mathrm{e}^{-2\pi \mathrm{i}\langle r\xi, y\rangle}) \,\mathrm{d}\mu(y) \hat{f}(r) \,\mathrm{d}r \,\mathrm{d}\xi \\
&= \int_{\mathbb{S}^{d-1}} \int_{\mathbb{R}} |\mu(\mathbb{R}^d) - \hat{\mu}(r\xi)|^2 \hat{f}(r) \,\mathrm{d}r \,\mathrm{d}\xi \\
&= 2 \int_{\mathbb{S}^{d-1}} \int_0^\infty |\mu(\mathbb{R}^d) - \hat{\mu}(r\xi)|^2 \frac{\hat{f}(\|r\xi\|)}{\|r\xi\|^{d-1}} r^{d-1} \,\mathrm{d}r \,\mathrm{d}\xi \\
&= 2 \int_{\mathbb{R}^d} |\mu(\mathbb{R}^d) - \hat{\mu}(x)|^2 \frac{\hat{f}(\|x\|)}{\|x\|^{d-1}} \,\mathrm{d}x.
\end{aligned}$$

Inserting $\hat{f}$, we can write

$$\|\mu\|_{\mathcal{H}_K}^2 = \frac{1}{\pi^2 \omega_{d-1}} \int_{\mathbb{R}^d} |\hat{\mu}(0) - \hat{\mu}(x)|^2 \frac{\hat{u}(\|x\|)}{\|x\|^{d+1}} \,\mathrm{d}x - F(0)\hat{\mu}(0)^2. \tag{32}$$

Now assume that $\mu \in \mathcal{M}_{1/2}(\mathbb{R}^d)$ with $\|\mu\|_{\mathcal{H}_K} = 0$. Because $F(0) < 0$ by Theorem 4.9 and $\hat{u} \geq 0$ by (4), both summands in (32) are nonnegative and therefore must vanish. The second term yields that $\mu(\mathbb{R}^d) = \hat{\mu} = 0$. Since $\operatorname{supp} \hat{u}(\|\cdot\|)\|\cdot\|^{-(d+1)} = \mathbb{R}^d$, cf. [43, Lem 2.39], it follows that $\hat{\mu}$ is constant with $\hat{\mu} \equiv \hat{\mu}(0) = 0$. This implies $\mu = 0$ as the Fourier transform $\mathcal{F}\colon \mathcal{M}(\mathbb{R}^d) \to C_b(\mathbb{R}^d)$ is injective. Consequently, the KME is injective, which means that $K$ is characteristic. $\square$

**Proof of Theorem 6.2.** Since $\Phi(x) \in \mathcal{O}(\|x\|^\alpha)$, we can estimate $|\Phi(x)| \leq C(1 + \|x\|^\alpha)$ for all $x \in \mathbb{R}^d$. For $\alpha \geq 1$ we have by convexity of $\|\cdot\|^\alpha$ that

$$\|x + y\|^\alpha \leq 2^{\alpha-1}(\|x\|^\alpha + \|y\|^\alpha).$$

For $\alpha \in (0, 1)$, we define the function $f\colon [0, \infty) \to \mathbb{R}$ by $f(x) := x^\alpha$, which is concave and monotone increasing. Then we obtain for $x, y \geq 0$ with $x + y > 0$ that

$$\begin{aligned}
f(x) &\geq \frac{y}{x+y} f(0) + \frac{x}{x+y} f(x+y), \\
f(y) &\geq \frac{x}{x+y} f(0) + \frac{y}{x+y} f(x+y).
\end{aligned}$$

Adding both equation yields

$$f(x) + f(y) \geq f(x+y).$$

Since $f$ is monotone increasing, we obtain by the triangle inequality

$$\|x + y\|^\alpha \leq (\|x\| + \|y\|)^\alpha \leq \|x\|^\alpha + \|y\|^\alpha.$$

Summarizing, we have for $\alpha \geq 0$ that

$$\|x + y\|^\alpha \leq 2^\alpha(\|x\|^\alpha + \|y\|^\alpha).$$

Therefore, we can guarantee the existence of the integral

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |\Phi(x-y)| \,\mathrm{d}\sigma(x)\,\mathrm{d}\sigma(y) \le \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} C(1+2^\alpha(\|x\|^\alpha+\|y\|^\alpha))\,\mathrm{d}\sigma(x)\,\mathrm{d}\sigma(y)$$
$$\le C\sigma(\mathbb{R}^d)\left(\sigma(\mathbb{R}^d)+2^{\alpha+1}\int_{\mathbb{R}^d}\|x\|^\alpha\,\mathrm{d}\sigma(x)\right) < \infty.$$

Hence, the discrepancy $d_{\tilde{K}}(\mu,\nu)$ is well-defined for $\mu,\nu \in \mathcal{M}_\alpha(\mathbb{R}^d)$.

By (10), we see that $K(x,x) \in \mathcal{O}(\|x\|^{2\beta})$, $\beta := \max\{r-1,(\alpha+r-1)/2\}$ such that $d_K$ is by (15) well-defined for measures in $\mathcal{M}_\beta$.

Now assume additionally that the first $r-1$ moments of $\mu$ and $\nu$ coincide. This implies that for all $p_j \in \Pi_{r-1}(\mathbb{R}^d)$ that

$$\int_{\mathbb{R}^d} p_j(x)\,\mathrm{d}(\mu-\nu)(x) = 0.$$

Then we obtain

$$\begin{aligned} d_K(\mu,\nu)^2 = d_{\tilde{K}}(\mu,\nu)^2 &- \sum_{j=1}^N \int_{\mathbb{R}^d} p_j(x)\,\mathrm{d}(\mu-\nu)(x)\int_{\mathbb{R}^d}\Phi(\xi_j-y)\,\mathrm{d}(\mu-\nu)(y)\\ &- \sum_{k=1}^N \int_{\mathbb{R}^d} p_k(y)\,\mathrm{d}(\mu-\nu)(y)\int_{\mathbb{R}^d}\Phi(x-\xi_k)\,\mathrm{d}(\mu-\nu)(x)\\ &+ \sum_{k,j=1}^N \Phi(\xi_j-\xi_k)\int_{\mathbb{R}^d} p_j(x)\,\mathrm{d}(\mu-\nu)(x)\int_{\mathbb{R}^d} p_k(y)\,\mathrm{d}(\mu-\nu)(y)\\ &= d_{\tilde{K}}(\mu,\nu)^2. \end{aligned}$$

□

**Proof of Proposition 6.4.** 1. First, we show that $K$ is $\lfloor\frac{n+2}{2}\rfloor$ times continuously differentiable. Let $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le \lfloor\frac{n+2}{2}\rfloor$. The case $|\alpha| = 0$ is clear. For $|\alpha| \ge 1$, we obtain

$$\begin{aligned} \partial_x^\alpha\partial_y^\alpha K(x,y) &= \partial_x^\alpha\partial_y^\alpha F(\|x-y\|) = \partial_x^\alpha\partial_y^\alpha \frac{1}{\omega_{d-1}}\int_{\mathbb{S}^{d-1}} f(\langle\xi,x-y\rangle)\,\mathrm{d}\xi\\ &= \frac{(-1)^{|\alpha|}}{\omega_{d-1}}\int_{\mathbb{S}^{d-1}} \xi^{2\alpha} f^{2|\alpha|}(\langle\xi,x-y\rangle)\,\mathrm{d}\xi. \end{aligned} \tag{33}$$

By [52, Cor. 4.36], this implies that every $h \in \mathcal{H}_K$ is at least $\lfloor\frac{n+2}{2}\rfloor$-times continuously differentiable.

2. For the second part, assume that $n \ge 2$. By [52, Lem 4.34] and (33), we obtain

$$\langle\partial_{x_i}K(x,\cdot),\partial_{x_i}K(y,\cdot)\rangle_{\mathcal{H}_K} = \partial_{x_i}\partial_{y_i}K(x,y) = \frac{-1}{\omega_{d-1}}\int_{\mathbb{S}^{d-1}}\xi_i^2 f''(\langle\xi,x-y\rangle)\,\mathrm{d}\xi.$$

By Proposition 3.1, the function $f$ is even and $f''' = 2u'$. Hence $u'$ is odd and $\|u''\|_\infty$-Lipschitz continuous and $f'''(0) = 2u'(0) = 0$. Thus, we obtain for $s > 0$ that

$$|f''(s) - f''(0)| \le \int_0^s |f'''(t)|\, \mathrm{d}t = \int_0^s |f'''(t) - f'''(0)|\, \mathrm{d}t \le \int_0^s 2\|u''\|_\infty t\, \mathrm{d}t = \|u''\|_\infty s^2.$$

Hence we can estimate

$$\begin{aligned}
\|\partial_{x_i} K(x,\cdot) - \partial_{x_i} K(y,\cdot)\|^2_{\mathcal{H}_K} &= \|\partial_{x_i} K(x,\cdot)\|^2_{\mathcal{H}_K} + \|\partial_i K(y,\cdot)\|^2_{\mathcal{H}_K} - 2\langle \partial_{x_i} K(x,\cdot), \partial_{x_i} K(y,\cdot)\rangle_{\mathcal{H}_K} \\
&= -\frac{2}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} \xi_i^2 \left(f''(0) - f''(\langle \xi, x-y\rangle)\right) \mathrm{d}\xi \\
&\le \frac{2\|u''\|_\infty}{\omega_{d-1}} \int_{\mathbb{S}^{d-1}} |\langle \xi, x-y\rangle|^2 \,\mathrm{d}\xi \\
&\le 2\|u''\|_\infty \|x-y\|^2.
\end{aligned}$$

Therefore, $\partial_{x_i} K(x,\cdot)$ is Lipschitz continuous with constant $L := \sqrt{2\|u''\|_\infty}$. Finally, we see again by (the proof of) [52, Cor. 4.36], for any $h \in \mathcal{H}_K$, that

$$\begin{aligned}
|\partial_{x_i} h(x) - \partial_{x_i} h(y)| &= |\langle h, \partial_{x_i} K(x,\cdot)\rangle_{\mathcal{H}_K} - \langle h, \partial_{x_i} K(y,\cdot)\rangle_{\mathcal{H}_K}| \\
&= |\langle h, \partial_{x_i} K(x,\cdot) - \partial_{x_i} K(y,\cdot)\rangle_{\mathcal{H}_K}| \\
&\le \|h\|_{\mathcal{H}_K} L \|x-y\|,
\end{aligned}$$

which gives the assertion by

$$\|\nabla h(x) - \nabla h(y)\| \le \sqrt{d} L \|h\|_{\mathcal{H}_K} \|x-y\|. \qquad \square$$

# D Geodesic $\lambda$-Convexity of MMD Functional with Smoothed Distance Kernel

A *generalized geodesic* is an interpolating curve $\gamma_t \colon [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ that connects two measures $\mu^2$ and $\mu^3 \in \mathcal{P}_2(\mathbb{R}^d)$ via a three-plan $\boldsymbol{\mu}$. More specifically, for a base $\mu^1 \in \mathcal{P}(\mathbb{R}^d)$, this three-plan $\boldsymbol{\mu} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ has marginals $\pi^i_\# \boldsymbol{\mu} = \mu^i, i = 1,2,3$ and must be optimal in the sense that $\pi^{1,i}_\# \boldsymbol{\mu} \in \Pi_{\mathrm{opt}}(\mu^1, \mu^i)$ for $i = 1, 2$. A generalized geodesic $\gamma_t$ joining $\mu^2$ with $\mu^3$ via $\mu^1$ is defined as $\gamma_t := ((1-t)\pi^2 + t\pi^3)_\# \boldsymbol{\mu}$. For any choice of $\mu^1, \mu^2, \mu^3 \in \mathcal{P}_2(\mathbb{R}^d)$, we can always find optimal plans $\boldsymbol{\mu}^{1,i} \in \Pi_{\mathrm{opt}}(\mu^1, \mu^i)$ for $i = 1, 2$, and by the Gluing Lemma [56] there exists a three plan $\boldsymbol{\mu} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ with marginals $\pi^i_\# \boldsymbol{\mu} = \mu^i, i = 1,2,3$ and $\pi^{1,i}_\# \boldsymbol{\mu} \in \Pi_{\mathrm{opt}}(\mu^1, \mu^i)$ for $i = 1, 2$. This means that there always exists at least one generalized geodesic joining $\mu^2$ with $\mu^3$ via $\mu^1$. However, this generalized geodesic is not necessarily unique.

Given $\lambda \in \mathbb{R}$, a function $G \colon \mathcal{P}_2(\mathbb{R}^d) \to [0, \infty]$ is *$\lambda$-convex along generalized geodesics*, if for any choice $\mu^1, \mu^2, \mu^3 \in \mathrm{dom}(G)$ there always exists a generalized geodesic $\gamma_t$ joining $\mu^2$ with $\mu^3$ via $\mu^1$, such that

$$G(\gamma_t) \le (1-t)G(\mu^2) + tG(\mu^3) - \frac{\lambda}{2} t(1-t) \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \|x_2 - x_3\|^2 \,\mathrm{d}\boldsymbol{\mu}(x_1, x_2, x_3).$$

For a more detailed description we refer to [2, Section 9.2].

In [2] sufficient conditions for the $\lambda$-convexity of the following two typical energy functionals were given. The *potential energy* $V\colon \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathcal{V}(\mu) := \int_{\mathbb{R}^d} V(x)\,\mathrm{d}\mu(x).$$

**Lemma D.1.** *Let $V$ be lower semi-continuous and have quadratic grow, i.e.*

$$V(x) \geq -A - B\|x\|^2 \quad \textit{for all} \quad x \in \mathbb{R}^d$$

*with $A, B \in \mathbb{R}$. If $V$ is $\lambda$-convex, then $\mathcal{V}$ is $\lambda$-convex along generalized geodesics.*

For $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the *interaction energy* is given by

$$\mathcal{K}(\mu) := \int_{\mathbb{R}^d \times \mathbb{R}^d} K(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y).$$

Since the interaction energy can be seen as a potential energy on the product space, Proposition D.1 also applies to the interaction energy.

**Lemma D.2.** *Let $K$ be lower semi-continuous and have quadratic grow, i.e.*

$$K(x,y) \geq -A - B(\|x\|^2 + \|y\|^2) \quad \textit{for all} \quad x, y \in \mathbb{R}^d$$

*with $A, B \in \mathbb{R}$. If $K$ is $\lambda$-convex, then $\mathcal{K}$ is $\lambda$-convex along generalized geodesics.*

Let $F$ be defined as in (9) and $K(x,y) = F(\|x-y\|)$. Then, the MMD functional $G$ from (21) can be rewritten as

$$G(\mu) = \frac{1}{2}\mathcal{K}(\mu) + \mathcal{V}(\mu) + \frac{1}{2}c_\nu, \quad V(x) := -\int_{\mathbb{R}^d} F(\|x-y\|)\,\mathrm{d}\nu(y)) \tag{34}$$

where $c_\nu \geq 0$ is a constant. Both $V$ and $K$ suffice the conditions in Lemmas D.1 and D.2.

**Proposition D.3.** *Let $F$ be defined as in* (9) *and $V$ and $K$ in* (34)*, then $\mathcal{V}$ and $\mathcal{K}$ are lower semi-continuous and have quadratic grow. Moreover, $V$ is convex and $K$ is $\lambda$-convex with $\lambda = -4\sqrt{d}\|u\|_\infty$. In summary, the MMD functional $G$ from* (34) *is lower semi-continuous and $\lambda$-convex with $\lambda$ above.*

*Proof.* By Theorem 4.9 , we can write $F(s) = -C_d|s| + \varphi(s)$ with $\varphi \in \mathcal{C}_0(\mathbb{R})$. For the lower semi-continuity of $V$, we have

$$\begin{aligned}
|V(x_1) - V(x_2)| &\leq \int_{\mathbb{R}^d} |F(\|x_1 - y\|) - F(\|x_2 - y\|)|\,\mathrm{d}\nu(y) \\
&= \int_{\mathbb{R}^d} |C_d\|x_1 - y\| + \varphi(\|x_1 - y\|) - C_d\|x_2 - y\| - \varphi(\|x_2 - y\|)|\,\mathrm{d}\nu(y) \\
&\leq \int_{\mathbb{R}^d} C_d|\|x_1 - y\| - \|x_2 - y\|| + |\varphi(\|x_1 - y\|) - \varphi(\|x_2 - y\|)|\,\mathrm{d}\nu(y) \\
&\leq C_d\|x_1 - x_2\| + \int_{\mathbb{R}^d} \|\varphi(\|x_1 - y\|) - \varphi(\|x_2 - y\|)|\,\mathrm{d}\nu(y).
\end{aligned}$$

Since $\varphi \in \mathcal{C}_0(\mathbb{R})$, it follows by Lebesgue's dominated convergence theorem that $V$ is continuous. Moreover, choosing $A, B = 0$, we see that $V$ has quadratic grow. By Theorem 4.9 iii), the function $-F(\|x\|)$ is convex. For $x_1, x_2 \in \mathbb{R}^d$ and $t \in [0,1]$, it holds

$$\begin{aligned} V((1-t)x_1 + tx_2) &= \int_{\mathbb{R}^d} -F(\|(1-t)(x_1 - y) + t(x_2 - y)\|) \,\mathrm{d}\nu(y) \\ &\leq \int_{\mathbb{R}^d} -(1-t)F(\|x_1 - y\|) - tF(\|x_2 - y)\|) \,\mathrm{d}\nu(y) \\ &= (1-t)V(x_1) + tV(x_2). \end{aligned}$$

Hence $V$ is convex, too.

For the interaction energy, it is clear that $K$ is continuous, because $F$ is continuous, and we can choose $A = \|\varphi\|_\infty + 2C_d$ and $B = C_d$ to obtain

$$\begin{aligned} F(\|x-y\|) &= -C_d\|x-y\| + \varphi(\|x-y\|) \geq -\|\varphi\|_\infty - C_d(\|x\| + \|y\|) \\ &\geq -\|\varphi\|_\infty - C_d(1 + \|x\|^2 + 1 + \|y\|^2) \geq -A - B(\|x\|^2 + \|y\|^2). \end{aligned}$$

By Corollary 4.9 v), we get that $K$ is $\lambda$-convex with $\lambda = -4\sqrt{d}\|u\|_\infty$. □

By [2, 11.2.1b], a lower bounded $\lambda$-convex functional is always coercive, so that [2, Thm.11.2.1] can be formulated as follows:

**Theorem D.4.** *Let $G\colon \mathcal{P}_2(\mathbb{R}^d) \to [0,\infty]$ be proper lower semi-continuous and $\lambda$ convex. Then, for $\gamma^{(0)} \in \mathrm{dom}(G)$, there is a unique Wasserstein gradient flow $\gamma_t$ starting in $\gamma^{(0)}$. Moreover, the piecewise constant curve $\gamma_t^\tau := \gamma^{(k)}$, $t \in ((k-1)\tau, k\tau]$ given by the implicit Euler scheme (JKO scheme)*

$$\gamma^{(k+1)} \in \operatorname*{arg\,min}_{\gamma \in \mathcal{P}_2(\mathbb{R}^d)} \frac{1}{2\tau} W_2^2(\gamma^{(k)}, \gamma) + \phi(\gamma),$$

*converges locally uniformly to $\gamma_t$. In particular, this holds true for our MMD functional with smooed kernel $G$ in* (34)*.*

It was shown in [27, Prop. 9] that for certain functionals, e.g. $G$ in (34), the so-called Wasserstein steepest descent flows (explicit scheme) and the Wasserstein gradient flows (implicit scheme) coincide.

Since the MMD functional with our SND kernel (same for the Gaussian kernel) is only $\lambda$-convex with $\lambda < 0$ it is not ensured that its Wasserstein gradient flow, resp. its approximation by an Euler forward scheme converges towards the target $\nu$. Here is an example.

**Example D.5.** *In general, it is not clear whether the gradient flow $\gamma_t$ the MMD functional with smooth kernels converges towards the target measure $\nu$ as $t \to \infty$. To this end, consider the symmetric setup with the target and initial measures*

$$\begin{aligned} \nu &:= \tfrac{1}{2}(\delta_{y_1} + \delta_{y_2}), \quad y_1 = e_1,\ y_2 = -e_1, \\ \mu &:= \tfrac{1}{2}(\delta_{x_1} + \delta_{x_2}), \quad x_1 = \tfrac{1}{\sqrt{3}}e_2,\ x_2 = -\tfrac{1}{\sqrt{3}}e_2. \end{aligned}$$

*Then, with $\tilde{F}(s) := \frac{F'(s)}{s}$, the velocity field becomes for $i = 1, 2$*

$$\begin{aligned} v_t(x_i) &= \tfrac{1}{2}(x_i - x_j)\tilde{F}(\|x_i - x_j\|) - \tfrac{1}{2}\left((x_i - y_1)\tilde{F}(\|x_i - y_1\|) + (x_i - y_2)\tilde{F}(\|x_i - y_2\|)\right) \\ &= \tfrac{1}{\sqrt{3}} e_2 \tilde{F}(\tfrac{2}{\sqrt{3}}) - \tfrac{1}{2}\left(\tfrac{2}{\sqrt{3}} e_2 - e_1 + e_1\right)\tilde{F}(\tfrac{2}{\sqrt{3}}) = 0, \end{aligned}$$

*so that we get stuck in $\gamma_t = \mu$ for all $t \geq 0$. In general, ensuring convergence towards the target is challenging due to local extrema. However, in the numerical experiments, we observed that for both ND and SND, the flows typically performed well in approximating the target.*

**Proof of Proposition 8.1.** The simplification of the iterations to

$$\begin{aligned} x^{(k+1)} &= x^{(k)} + \tau \frac{x^{(k)} - y}{\|x^{(k)} - y\|} F'(\|x^{(k)} - y\|) \\ &= y + (x^{(k)} - y)\left(1 + \tau \frac{F'(\|x^{(k)} - y\|)}{\|x^{(k)} - y\|}\right) \end{aligned} \tag{35}$$

is straightforward.

i) For $F = -\frac{1}{2}$ abs, we have $F'(s) = -\frac{1}{2}$ for $s > 0$. Then we obtain by (35) and since $\|x^{(0)} - y\| < \frac{\tau}{2}$ that

$$\|y - x^{(1)}\| = \frac{\tau}{2} - \|x^{(0)} - y\| < \frac{\tau}{2}.$$

The second step, $x^{(2)}$ jumps exactly back to $x^{(0)}$ because

$$\begin{aligned} x^{(2)} &= y + (x^{(1)} - y)\left(1 - \frac{\tau}{2\|x^{(1)} - y\|}\right) \\ &= y + (x^{(0)} - y)\left(1 - \frac{\tau}{2\|x^{0} - y\|}\right)\left(1 - \frac{\tau}{2\|x^{(1)} - y\|}\right) \\ &= y + (x^{(0)} - y)\frac{(2\|x^{(0)} - y\| - \tau)(-2\|x^{(0)} - y\|)}{2\|x^{(0)} - y\|(\tau - 2\|x^{(0)} - y\|)} = x^{(0)}. \end{aligned}$$

Consequently, $(x^{(k)})_k$ oscillates between $x^{(0)}$ and $x^{(1)}$.

ii) Generally, for $\lambda$-convex functionals with $\lambda > 0$, Baillon-Haddad's theorem [5, Cor. 18.17] ensures convergence of (27) for $\tau < \lambda^{-1}$. For completeness, we provide a simpler proof for our setting. Let $F = \mathcal{I}_d[-|u|]$, with $u \in \mathcal{U}^0(\mathbb{R})$. We know that $F$ is convex and twice differentiable. In particular, we have for $\tilde{F}(s) := \frac{F'(s)}{s}$ that $\tilde{F}(0) = F''(0) < 0$. Since $\tilde{F} \in \mathcal{C}(\mathbb{R})$ by Proposition 4.6, we can find $\delta > 0$ such that $F(x) < \frac{1}{2}F''(0)$ for $|x| < \delta$. If we assume that $\|x^{(k)} - y\| < \tau$ and $\tau < \min\{\delta, \|\tilde{F}'\|_\infty^{-1}\}$, we obtain

$$\|x^{(k+1)} - y\| = \|x^{(k)} - y\|\left(1 + \tau\tilde{F}(\|x^{(k)} - y\|)\right).$$

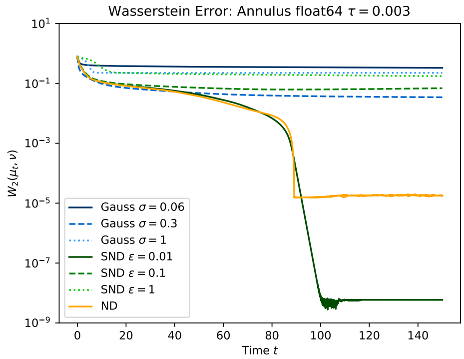



Figure 9: Wasserstein 2 error between Annulus target $\nu$ and flow $\gamma_n^\tau$ after $n$ iterations. Horizontal axis in time $t = \tau n$ with Gaussian (blue), SND (green) and ND (orange). Both computed with double precision with step size $\tau = 0.003$.

We always have

$$1 + \tau \tilde{F}(\|x^{(k)} - y\|) > 1 - \tau \|\tilde{F}\|_\infty > 0.$$

Since $\|x^{(k)} - y\| < \delta$, we know that $F(\|x^{(k)} - y\|) < \frac{1}{2}F''(0) < 0$, which implies

$$1 + \tau \tilde{F}(\|x^{(k)} - y\|) < 1 + \tfrac{\tau}{2}F''(0) < 1.$$

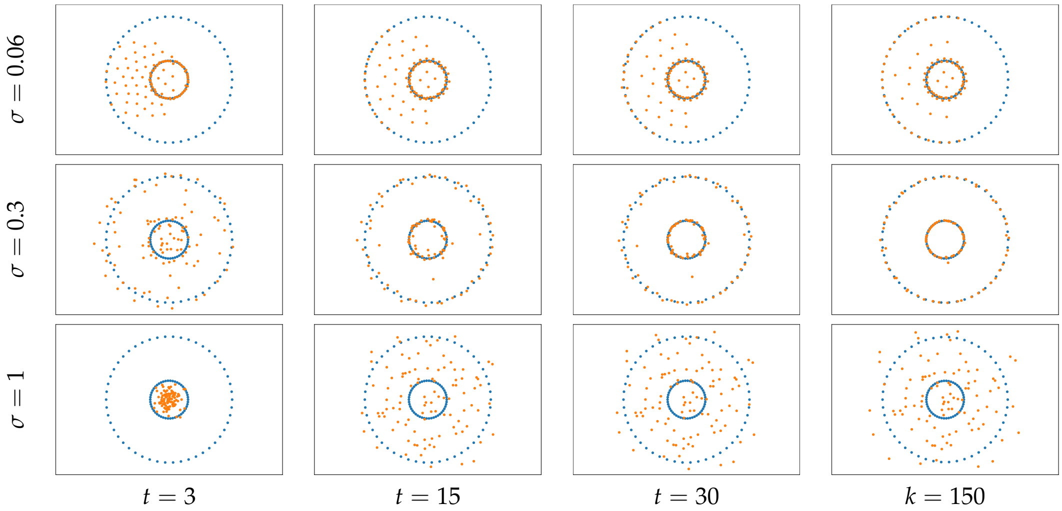
This yields $\|x^{(k+1)} - y\| < \|x^{(k)} - y\| < \delta$, and thus, by induction,

$$\|x^{(k+1)} - y\| \leq \|x^{(0)} - y\|(1 + \tfrac{\tau}{2}F''(0))^k.$$

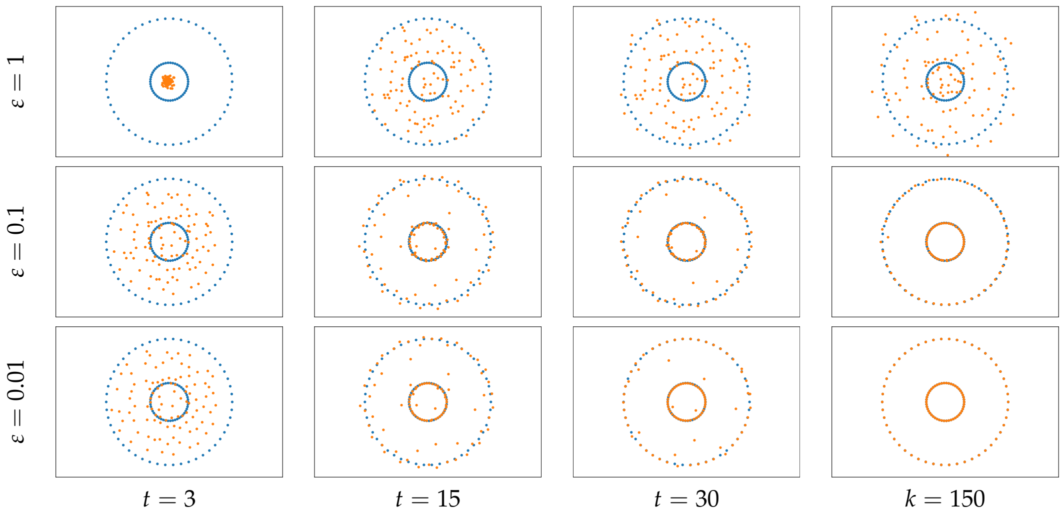
Therefore, we have exponential convergence when $\tau$ and $\|x^{(k)} - y\|$ are sufficiently small. $\square$

# E Numerical Example: Annulus Target

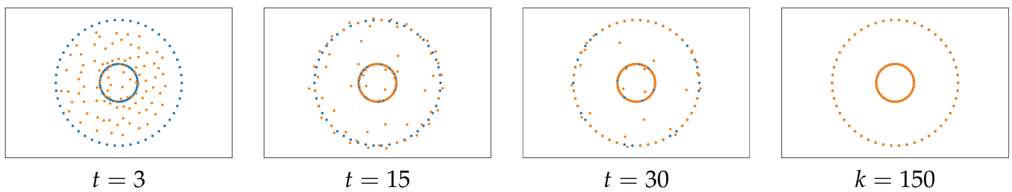
The Annulus target consists of two concentric circles with radius 1 and 0.3. Each is discretized with 50 points, so that $\nu$ consists of $M = 100$ points. Here we use a step size of $\tau = 0.003$ and double precision. The MMD flows are depicted in Figure 10 and the respective errors in Figure 9.




(a) Gaussian




(b) SND




(c) ND

Figure 10: MMD flow (25) with step size $\tau = 0.02$.