Neural Wasserstein Gradient Flows for Discrepancies with Riesz Kernels

Fabian Altekrüger [1] [2] Johannes Hertrich [2] Gabriele Steidl [2]

Abstract

Wasserstein gradient flows of maximum mean discrepancy (MMD) functionals with non-smooth Riesz kernels show a rich structure as singular measures can become absolutely continuous ones and conversely. In this paper we contribute to the understanding of such flows. We propose to approximate the backward scheme of Jordan, Kinderlehrer and Otto for computing such Wasserstein gradient flows as well as a forward scheme for so-called Wasserstein steepest descent flows by neural networks (NNs). Since we cannot restrict ourselves to absolutely continuous measures, we have to deal with transport plans and velocity plans instead of usual transport maps and velocity fields. Indeed, we approximate the disintegration of both plans by generative NNs which are learned with respect to appropriate loss functions. In order to evaluate the quality of both neural schemes, we benchmark them on the interaction energy. Here we provide analytic formulas for Wasserstein schemes starting at a Dirac measure and show their convergence as the time step size tends to zero. Finally, we illustrate our neural MMD flows by numerical examples.

1. Introduction

Wasserstein gradient flows of certain functionals $\mathcal{F}$ gained increasing attention in generative modeling over the last years. If $\mathcal{F}$ is given by the Kullback-Leibler divergence, the corresponding gradient flow can be represented by the Fokker-Planck equation and the Langevin equation (Jordan et al., 1998; Otto, 2001; Otto & Westdickenberg, 2005; Pavliotis, 2014) and is related to the Stein variational gradient descent (Dong et al., 2023; Grathwohl et al., 2020; di Langosco et al., 2022). In combination with deep-learning techniques, these representations can be used for generative modeling, see, e.g., (Ansari et al., 2021; Gao et al., 2019; Glaser et al., 2021; Hagemann et al., 2022; 2023; Song et al., 2021; Song & Ermon, 2019; Welling & Teh, 2011). For approximating Wasserstein gradient flows for more general functionals, a backward discretization scheme in time, known as Jordan-Kinderlehrer-Otto (JKO) scheme (Giorgi, 1993; Jordan et al., 1998) can be used. Its basic idea is to discretize the whole flow in time by applying iteratively the Wasserstein proximal operator with respect to $\mathcal{F}$. In case of absolutely continuous measures, Brenier's theorem (Brenier, 1987) can be applied to rewrite this operator via transport maps having convex potentials and to learn these transport maps (Fan et al., 2022) or their potentials (Alvarez-Melis et al., 2022; Bunne et al., 2022; Mokrov et al., 2021) by neural networks (NNs). In most papers, the objective functional arises from Kullback-Leibler divergence or its relatives, which restricts the considerations to absolutely continuous measures.

In this paper, we are interested in gradient flows with respect to discrepancy functionals which are also defined for singular measures. Moreover, in contrast to Langevin Monte Carlo algorithms, no analytical form of the target measure is required. The *maximum mean discrepancy* (MMD) is defined as $\mathcal{D}_K^2(\mu,\nu) := \mathcal{E}_K(\mu-\nu)$, where $\mathcal{E}_K$ is the *interaction energy* for signed measures

$$\mathcal{E}_K(\eta) := \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\eta(x)\mathrm{d}\eta(y)$$

and $K\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a conditionally positive definite kernel. Then, we consider gradient flows with respect to the MMD functional $\mathcal{F}_\nu\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ given by

$$\mathcal{F}_\nu := \mathcal{E}_K + \mathcal{V}_{K,\nu} = \mathcal{D}_K^2(\cdot,\nu) + \text{const},$$

where $\mathcal{V}_{K,\nu}(\mu)$ is the so-called *potential energy*

$$\mathcal{V}_{K,\nu}(\mu) := -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} K(x,y)\,\mathrm{d}\nu(y)\,\mathrm{d}\mu(x)$$

acting as an attraction term between the masses of μ and ν, while the interaction energy $\mathcal{E}_K$ is a repulsion term enforcing a proper spread of μ. In general, it will be essential for our

[1] Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany [2] Institute of Mathematics, Technische Universität Berlin, Straße des 17. Juni 136, D-10623 Berlin, Germany. Correspondence to: Fabian Altekrüger <fabian.altekrueger@hu-berlin.de>, Johannes Hertrich <j.hertrich@math.tu-berlin.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

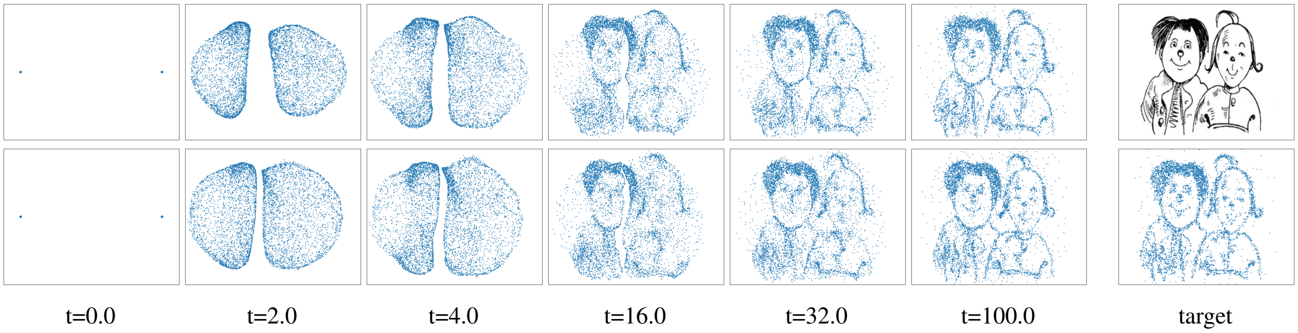


Figure 1: Neural backward (top) and forward (bottom) schemes for the Wasserstein flow of the MMD with distance kernel starting in exactly two points 'sampled' from $\delta_{(-0.5,0)} + \delta_{(0.5,0)}$ toward the 2D density 'Max und Moritz' (Drawing by Wilhelm Busch top right and a sampled version bottom right).

method that the flow's functional can be approximated by samples, which is, e.g., possible if it is defined by an integral

$$\mathcal{F}(\mu) = \int_{\mathbb{R}^d} G(x)\mathrm{d}\mu(x), \quad G\colon \mathbb{R}^d \to \mathbb{R}, \tag{1}$$

like in the potential energy or by a double integral like in the interaction energy. MMD gradient flows are directly related to NN optimization (Arbel et al., 2019). For λ-convex kernels with Lipschitz continuous gradient, MMD Wasserstein gradient flows were thoroughly investigated in (Arbel et al., 2019). In particular, it was shown that these flows can be described as particle flows. However, in certain applications, non-smooth and non-λ-convex kernels like Riesz kernels $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$,

$$K(x, y) = -\|x - y\|^r, \quad r \in (0, 2) \tag{2}$$

and especially negative distance kernels are of interest (Carrillo & Huang, 2017; Chafaï et al., 2023; Ehler et al., 2021; Gräf et al., 2012; Teuber et al., 2011; Wendland, 2005). Here it is known that the Wasserstein gradient flow of the interaction energy starting at an empirical measure cannot remain empirical (Balagué et al., 2013), so that these flows are no longer just particle flows. In particular, Dirac measures (particles) might "explode" and become absolutely continuous measures in two dimensions or "condensated" singular non-Dirac measures in higher dimensions and conversely. For an illustration, see the last example in Appendix H and (Hertrich et al., 2023a) for the one-dimensional setting. Thus, neither the analysis of absolutely continuous Wasserstein gradient flows nor the findings in (Arbel et al., 2019) are applicable in this case. From the computational side, Riesz kernels have exceptional properties which allow a very efficient computation of the corresponding MMD. More precisely, in (Hertrich et al., 2023b) it was shown that MMD with Riesz kernels coincides with its sliced version. Thus, the computation of gradients of MMD can be done in the one-dimensional setting in a fast manner using a simple sorting algorithm.

Contributions. We propose to compute the JKO scheme by learning generative NNs which approximate the disintegration of the transport plans. Using plans instead of maps, we are no longer restricted to absolutely continuous measures, while. Similarly, we consider Wasserstein steepest descent flows (Hertrich et al., 2022) and a forward discretization scheme in time. We approximate the disintegration of the corresponding velocity plans by NNs, where we have to use the loss function corresponding to the steepest descent flow now. Using the disintegration for both schemes, we can handle arbitrary measures in contrast to existing methods, which are limited to the absolutely continuous case. This could be of interest when considering target measures supported on submanifolds, as done in, e.g., (Brehmer & Cranmer, 2020). MMD flows approximated by our neural schemes starting just at two points are illustrated in Fig. 1. Another contribution of our paper is the convergence analysis of the backward, resp. forward schemes starting at a Dirac measure for the interaction energy. Indeed, we provide analytical formulas for the JKO and forward schemes and prove that they converge to the same curve when the time step size goes to zero. This delivers a ground truth for evaluating our neural approximations. We highlight the performance of our neural backward and forward schemes by numerical examples.

Related Work. There exist several approaches to compute neural approximations of the JKO scheme for absolutely continuous measures. Exploiting Brenier's Theorem, in (Alvarez-Melis et al., 2022; Bunne et al., 2022; Mokrov et al., 2021) it was proposed to use input convex NNs (ICNNs) (Amos et al., 2017) within the JKO scheme. More precisely, starting with samples from the initial measure μ, samples from each step of the JKO were iteratively generated by discretizing the functional $\mathcal{F}$ in (Alvarez-Melis et al., 2022; Mokrov et al., 2021), see also Sect. 3. If the potential is strictly convex, they can compute the density in each step using the change-of-variables formula. A similar approach was used in (Bunne et al., 2022), but here the objective is to approximate the functional $\mathcal{F}$ via NNs for a given trajectory of samples. In (Hwang et al., 2021), approximation results for a similar method were provided. Instead of using ICNNs, (Fan et al., 2022) proposed to directly learn the transport map and rewrite the functional $\mathcal{F}$ with a variational formula. Here it is possible to compute $\mathcal{F}$ sample-based, but a minimax problem has to be solved. Finally, motivated by the computational burden of the JKO scheme, the Wasserstein distance in the JKO scheme was replaced by the sliced-Wasserstein distance in (Bonet et al., 2022). All these methods rely on absolutely continuous measures and are not directly applicable for general measures. A slight modification of the JKO scheme for simulating Wasserstein flows is proposed in (Carrillo et al., 2022). For the task of computing strong and weak optimal transport plans, a generalization of transport maps to transport plans was done in (Korotin et al., 2023). Here a transport plan, represented by a so-called stochastic transport map, was learned exploiting the dual formulation of the Wasserstein distance as a minimax problem. Another approach for learning transport plans by training a NN in an adversarial fashion was proposed in (Lu et al., 2020). Recently, it was shown in (Arbel et al., 2019) that Wasserstein flows of MMDs with smooth and λ-convex kernels can be fully described by particle flows. However, here we are interested in non-smooth and non-λ-convex kernels, where this characterization does not hold true. Finally, closely related to gradient flows are Wasserstein natural gradient methods which replace Euclidean gradients by more general ones, see (Arbel et al., 2020; Chen & Li, 2020; Lin et al., 2021).

Outline. We introduce Wasserstein gradient flows and Wasserstein steepest descent flows as well as a backward and forward scheme for their time discretization in Sect. 2. In Sect. 3, we derive a neural backward scheme and in Sect. 4 a neural forward scheme. Analytic formulas for backward and forward schemes of Wasserstein flows of the interaction energy starting at a Dirac measure are given in Sect. 5. These ground truths are used in the first examples in Sect. 6 and were subsequently accomplished by examples for MMD flows. Proofs are postponed to the appendix.

2. Wasserstein Flows

We are interested in gradient flows in the *Wasserstein space* $\mathcal{P}_2(\mathbb{R}^d)$ of Borel probability measures with finite second moments equipped with the *Wasserstein distance*

$$W_2^2(\mu,\nu) := \min_{\boldsymbol{\pi}\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \mathrm{d}\boldsymbol{\pi}(x,y), \qquad (3)$$

where $\Gamma(\mu,\nu) := \{\boldsymbol{\pi} \in \mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d) : (\pi_1)_\#\boldsymbol{\pi} = \mu,\ (\pi_2)_\#\boldsymbol{\pi} = \nu\}$. Here $T_\#\mu := \mu\circ T^{-1}$ denotes the *push-forward* of μ via the measurable map T and $\pi_i(x) := x_i$, $i = 1,2$ for $x = (x_1,x_2) \in \mathbb{R}^{d\times d}$. In the case that μ is absolutely continuous, the Wasserstein distance can be reformulated by Breniers' theorem (Brenier, 1987) using transport maps $T:\mathbb{R}^d\to\mathbb{R}^d$ instead of transport plans as

$$W_2^2(\mu,\nu) = \min_{T_\#\mu=\nu} \int_{\mathbb{R}^d} \|x-T(x)\|^2 \mathrm{d}\mu(x). \qquad (4)$$

Then the optimal transport map $\hat{T}$ is unique and implies the unique optimal transport plan by $\hat{\boldsymbol{\pi}} = (\mathrm{Id},\hat{T})_\#\mu$. Further, $\hat{T} = \nabla\psi$ for some convex, lower semi-continuous (lsc) and μ-a.e. differentiable function $\psi\colon \mathbb{R}^d \to (-\infty,+\infty]$.

A curve $\gamma\colon I \to \mathcal{P}_2(\mathbb{R}^d)$ on the interval $I \subseteq \mathbb{R}$ is called *absolutely continuous* if there exists a Borel velocity field $v_t\colon \mathbb{R}^d\to\mathbb{R}^d$ with $\int_I \|v_t\|_{L_{2,\gamma(t)}}\mathrm{d}t < \infty$ such that the continuity equation

$$\partial_t\gamma(t) + \nabla\cdot(v_t\gamma(t)) = 0$$

is fulfilled on $I\times\mathbb{R}^d$ in a weak sense. An absolutely continuous curve $\gamma\colon(0,\infty)\to\mathcal{P}_2(\mathbb{R}^d)$ with velocity field $v_t \in \mathrm{T}_{\gamma(t)}\mathcal{P}_2(\mathbb{R}^d)$ is a **Wasserstein gradient flow with respect to** $\mathcal{F}\colon\mathcal{P}_2(\mathbb{R}^d)\to(-\infty,\infty]$ if

$$v_t \in -\partial\mathcal{F}(\gamma(t)), \quad \text{for a.e. } t>0, \qquad (5)$$

where $\partial\mathcal{F}(\mu)$ denotes the reduced Fréchet subdiffential at μ and $\mathrm{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ the regular tangent space, see Appendix A.

A pointwise formulation of Wasserstein flows using steepest descent directions was suggested by (Hertrich et al., 2022). In order to describe all "directions" in $\mathcal{P}_2(\mathbb{R}^d)$, it is not sufficient to consider velocity fields. Instead, we need velocity plans $\boldsymbol{v}\in\boldsymbol{V}(\mu)$, where $\boldsymbol{V}(\mu) := \{\boldsymbol{v}\in\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d) : (\pi_1)_\#\boldsymbol{v} = \mu\}$ (Ambrosio et al., 2005; Gigli, 2004). Now the curve $\gamma_{\boldsymbol{v}}$ in direction $\boldsymbol{v}\in\boldsymbol{V}(\mu)$ starting at μ is defined by

$$\gamma_{\boldsymbol{v}}(t) = (\pi_1 + t\pi_2)_\#\boldsymbol{v}.$$

The *(Dini-)directional derivative* of a function $\mathcal{F}\colon\mathcal{P}_2(\mathbb{R}^d)\to(-\infty,\infty]$ at $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ in direction $\boldsymbol{v}\in\boldsymbol{V}(\mu)$ is given by

$$\mathrm{D}_{\boldsymbol{v}}\mathcal{F}(\mu) := \lim_{t\to 0+}\frac{\mathcal{F}(\gamma_{\boldsymbol{v}}(t)) - \mathcal{F}(\mu)}{t}.$$

For a velocity plan $\boldsymbol{v}$, we define the *multiplication by* $c\in\mathbb{R}$ as $c\cdot\boldsymbol{v} := (\pi_1, c\pi_2)_\#\boldsymbol{v}$ and the *metric velocity* by $\|\boldsymbol{v}\|_\mu^2 := \int_{\mathbb{R}^d\times\mathbb{R}^d}\|y\|^2\mathrm{d}\boldsymbol{v}(x,y)$. Let $(x)^- := \max(-x,0)$. Then, inspired by properties of the gradient in Euclidean spaces, we define the *set of steepest descent directions at* $\mu\in\mathcal{P}_2(\mathbb{R}^d)$ by

$$\nabla_-\mathcal{F}(\mu) := \Big\{\Big(\frac{\mathrm{D}_{\hat{\boldsymbol{v}}}\mathcal{F}(\mu)}{\|\hat{\boldsymbol{v}}\|_\mu^2}\Big)^- \cdot \hat{\boldsymbol{v}} : \hat{\boldsymbol{v}} \in \operatorname*{arg\,min}_{\boldsymbol{v}\in\boldsymbol{V}(\mu)} \frac{\mathrm{D}_{\boldsymbol{v}}\mathcal{F}(\mu)}{\|\boldsymbol{v}\|_\mu}\Big\}. \qquad (6)$$

An absolutely continuous curve $\gamma\colon [0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ is a **Wasserstein steepest descent flow with respect to** $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ if

$$\dot\gamma(t) \in \nabla_- \mathcal{F}(\gamma(t)), \quad t \in [0,\infty), \tag{7}$$

where $\dot\gamma(t)$ is the tangent of γ at time t, see Appendix A.

Although both Wasserstein flows are different in general, they coincide by the following proposition from (Hertrich et al., 2022) for functions which are λ-convex along generalized geodesics, see (17) in the appendix.

Proposition 2.1. *Let $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ be locally Lipschitz continuous and λ-convex along generalized geodesics. Then, there exist unique Wasserstein steepest descent and gradient flows starting at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ and these flows coincide.*

Unfortunately, neither interaction energies nor MMD functionals with Riesz kernels (2) are λ-convex along generalized geodesics.

Remark 2.2. We slightly simplified the definitions in (Hertrich et al., 2022) as follows: i) The steepest descent directions $\boldsymbol{v}$ are originally defined to be in the so-called geometric tangent space. Although a formal proof is lacking, we conjecture that the minimizer $\hat{\boldsymbol{v}}$ in (6) is always contained in the geometric tangent space. ii) The original analysis uses Hadamard-directional derivatives, whose definition is stronger than the Dini-directional derivative. However, in case of locally Lipschitz continuous functions $\mathcal{F}$ as, e.g., the discrepancy functional with the Riesz kernel for $r \in [1,2)$, both definitions coincide.

3. Neural Backward Scheme

For computing Wasserstein gradient flows numerically, a backward scheme known as generalized *minimizing movement scheme* (Giorgi, 1993), or *Jordan-Kinderlehrer-Otto* (JKO) *scheme* (Jordan et al., 1998) can be applied which we explain next. For a proper, lsc function $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$, $\tau > 0$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the *Wasserstein proximal mapping* is the set-valued function

$$\operatorname{prox}_{\tau\mathcal{F}}(\mu) := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \tfrac{1}{2\tau} W_2^2(\mu,\nu) + \mathcal{F}(\nu) \right\}. \tag{8}$$

Note that the existence and uniqueness of the minimizer in (8) is assured if $\mathcal{F}$ is λ-convex along generalized geodesics, where $\lambda > -1/\tau$ and $\mu \in \operatorname{dom}\mathcal{F}$, see Lemma 9.2.7 in (Ambrosio et al., 2005).

The **backward scheme** (JKO) starting at $\mu_\tau^0 := \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with time step size $\tau > 0$ is the curve γ_τ given by $\gamma_\tau|_{((n-1)\tau, n\tau]} := \mu_\tau^n$, $n \in \mathbb{N}$, where

$$\mu_\tau^n := \operatorname{prox}_{\tau\mathcal{F}}(\mu_\tau^{n-1}). \tag{9}$$

If $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,+\infty]$ is coercive and λ-convex along generalized geodesics, then the JKO curves γ_τ starting at $\mu \in \overline{\operatorname{dom}\mathcal{F}}$ converge for $\tau \to 0$ locally uniformly to a locally Lipschitz curve $\gamma\colon (0,+\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, which is the unique Wasserstein gradient flow of $\mathcal{F}$ with $\gamma(0+) = \mu$, see Theorem 11.2.1 in (Ambrosio et al., 2005). For a scenario with more general regular functionals, we refer to Theorem 11.3.2 in (Ambrosio et al., 2005).

In general, Wasserstein proximal mappings are hard to compute, so that their approximation with NNs became an interesting topic. Most papers on neural Wasserstein gradient flows rely on the assumption that all μ_τ^n arising in the JKO scheme are absolutely continuous. Then, by (4), the scheme simplifies to $\mu_\tau^n = T_{n\#}\mu_\tau^{n-1}$, where T_n is contained in

$$\operatorname*{arg\,min}_T \Big\{ \frac{1}{2\tau} \int_{\mathbb{R}^d} \|x - T(x)\|^2 \mathrm{d}\mu_\tau^{n-1}(x) + \mathcal{F}(T_\#\mu_\tau^{n-1}) \Big\}. \tag{10}$$

In (Alvarez-Melis et al., 2022; Bunne et al., 2022; Mokrov et al., 2021), it was proposed to learn the transport map via its convex potential $T_n = \nabla\psi_n$ using input convex NNs, while (Fan et al., 2022) directly learned T_n.

Since we are interested in Wasserstein gradient flows for arbitrary measures, we extend (10) and the existing methods and consider the JKO scheme (9) with plans instead of just maps , i.e.,

$$\begin{aligned}\hat{\boldsymbol{\pi}} \in \operatorname*{arg\,min}_{\substack{\boldsymbol{\pi} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)\\ \pi_{1\#}\boldsymbol{\pi} = \mu_\tau^{n-1}}} \Big\{ \frac{1}{2\tau} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \mathrm{d}\boldsymbol{\pi}(x,y) \quad (11)\\ + \mathcal{F}(\pi_{2\#}\boldsymbol{\pi}) \Big\}, \quad \mu_\tau^n := (\pi_2)_\#\hat{\boldsymbol{\pi}}.\end{aligned}$$

We can describe such a plan $\hat\pi$ by a Markov kernel, as we will see in the next lemma.

Lemma 3.1. *For a measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the following equality holds*

$$\begin{aligned}&\{\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : \pi_{1\#}\pi = \mu\}\\ &\quad = \{\mu \times \pi_x : \pi_x \textit{ is a Markov kernel}\}\\ &\quad = \{\mu \times (\mathcal{T}(x,\cdot)_\# P_Z) : \mathcal{T}\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d \textit{ measurable}\}.\end{aligned}$$

Proof. The first equality directly follows from the disintegration theorem B.1 and the second equality follows from Brenier's theorem (Brenier, 1987). □

Details towards the disintegration can be found in Appendix B. By Lemma 3.1 we can rewrite (11) to

$$\begin{aligned}\hat{\mathcal{T}} \in \operatorname*{arg\,min}_{\mathcal{T}} \Big\{ \frac{1}{2\tau} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \mathrm{d}(\mathcal{T}(x,\cdot)_\# P_Z)(y) \mathrm{d}\mu_\tau^{n-1}(x)\\ + \mathcal{F}(\pi_{2\#}(\mu_\tau^{n-1} \times \mathcal{T}(x,\cdot)_\# P_Z)) \Big\}.\end{aligned}$$

Reformulating the pushforward measures we finally obtain

$$\begin{aligned}\hat{\mathcal{T}} \in \operatorname*{arg\,min}_{\mathcal{T}} \Big\{ \frac{1}{2\tau} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - \mathcal{T}(x,z)\|^2 \mathrm{d}P_Z(z) \mathrm{d}\mu_\tau^{n-1}(x)\\ + \mathcal{F}(\mathcal{T}_\#(\mu_\tau^{n-1} \times P_Z)) \Big\}.\end{aligned} \tag{12}$$

Now we propose to parameterize the map $\mathcal{T}$ by a NN $\mathcal{T}_\theta(x,\cdot)\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ for a standard Gaussian latent distribution $P_Z \sim \mathcal{N}(0, \mathrm{Id}_d)$. We learn the NN using the sampled function in (12) as loss function, see Alg. 1.

For this, it is essential that the function $\mathcal{F}$ can be approximated by samples as in (1). In summary, we obtain the sample-based approximation of the JKO scheme outlined in Alg. 1, which we call *neural backward scheme*.

Algorithm 1 Neural backward scheme

```
Input: Samples $x_1^0, ..., x_N^0$ from $\mu_\tau^0$ and $\mathcal{F}$ in (1).
for $n = 1, 2, ...$ do
    - Learn a Markov kernel $\boldsymbol{\pi}_x^n(\cdot) = \mathcal{T}_\theta^n(x,\cdot)_\# P_Z$
      by minimizing
      $$\mathcal{L}(\theta) := \mathbb{E}_{z_i\sim P_Z}\Big[\frac{1}{2\tau N}\sum_{i=1}^N \|x_i^{n-1} - \mathcal{T}_\theta(x_i^{n-1}, z_i)\|^2 + \frac{1}{N}\sum_{i=1}^N G(\mathcal{T}_\theta(x_i^{n-1}, z_i))\Big].$$
    - Sample $x_i^n$ from $\boldsymbol{\pi}_{x_i^{n-1}}$, i.e., draw $z_i \sim P_Z$ and
      set $x_i^n := \mathcal{T}_\theta(x_i^{n-1}, z_i)$.
    - Approximate $\mu_\tau^n := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^n}$.
end for
```

4. Neural Forward Scheme

For computing Wasserstein steepest descent flows numerically, we propose a time discretization by an Euler forward scheme.
The **forward scheme** starting at $\mu_\tau^0 := \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with time step size τ is the curve γ_τ given by $\gamma_\tau|_{((n-1)\tau, n\tau]} = \mu_\tau^n$ with

$$\mu_\tau^n := \gamma_{\boldsymbol{v}^{n-1}}(\tau), \quad \text{where} \quad \boldsymbol{v}^{n-1} \in \nabla_- \mathcal{F}(\mu_\tau^{n-1}). \tag{13}$$

The hard part consists in the computation of the velocity plans $\boldsymbol{v}^{n-1}$ which requires to solve the minimization problem $\hat{\boldsymbol{v}}^{n-1} \in \arg\min_{\boldsymbol{v}\in \boldsymbol{V}(\mu_\tau^{n-1})} \mathrm{D}_{\boldsymbol{v}}\mathcal{F}(\mu_\tau^{n-1})/\|\boldsymbol{v}\|_{\mu_\tau^{n-1}}$ in (6). For approximating these plans, we use again the disintegration $\boldsymbol{v} = \mu \times \boldsymbol{v}_x$ with respect to μ with Markov kernel $\boldsymbol{v}_x$. We propose to parameterize the Markov kernel $\boldsymbol{v}_x$ by a NN $\mathcal{T}_\theta(x,\cdot)\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ via $\boldsymbol{v}_x = \mathcal{T}_\theta(x,\cdot)_\# P_Z$ for a standard Gaussian latent distribution P_Z. Using again the form of $\mathcal{F}$ in (1), we learn the network by minimizing the loss function

$$\mathcal{L}(\theta) = \frac{\mathbb{E}_{z_i\sim P_Z}\left[\frac{1}{N}\sum_{i=1}^N \nabla_{\mathcal{T}_\theta(x_i,z_i)} G(x_i)\right]}{\left(\mathbb{E}_{z_i\sim P_Z}\left[\frac{1}{N}\sum_{i=1}^N \|\mathcal{T}_\theta(x_i,z_i)\|^2\right]\right)^{1/2}} \approx \frac{\mathrm{D}_{\boldsymbol{v}}\mathcal{F}(\mu)}{\|\boldsymbol{v}\|_\mu}, \tag{14}$$

where the x_i, $i = 1, ..., N$ are samples from μ and the above approximation of $\mathrm{D}_{\boldsymbol{v}}\mathcal{F}(\mu)$ follows from

$$\begin{aligned}
&\mathrm{D}_{\boldsymbol{v}}\mathcal{F}(\mu) = \lim_{t\to 0+} \tfrac{1}{t}\big(\mathcal{F}(\gamma_{\mu\times\boldsymbol{v}_x}(t)) - \mathcal{F}(\mu)\big)\\
&= \lim_{t\to 0+} \int \tfrac{1}{t} G(x) \mathrm{d}(\pi_1 + t\pi_2)_\#(\mu \times \boldsymbol{v}_x)(x) - \int \tfrac{1}{t} G(x)\mathrm{d}\mu(x)\\
&= \lim_{t\to 0+} \int \tfrac{1}{t}(G(x+ty) - G(x))\mathrm{d}\boldsymbol{v}_x(y)\mathrm{d}\mu(x)\\
&= \int \nabla_y G(x) \mathrm{d}\boldsymbol{v}_x(y)\mathrm{d}\mu(x).
\end{aligned}$$

Here $\nabla_y G(x) := \lim_{t\to 0+} \frac{G(x+ty)-G(x)}{t}$ denotes the right-sided directional derivative of G at x in direction y, which can be computed by the forward-mode of algorithmic differentiation. By (6), we need the rescaling

$$\mathcal{T}_{\theta,-}(x,z) = \left(\mathrm{D}_{\hat{\boldsymbol{v}}}\mathcal{F}(\mu)/\|\hat{\boldsymbol{v}}\|_\mu^2\right)^- \mathcal{T}_\theta(x,z), \tag{15}$$

where $\left(\mathrm{D}_{\hat{\boldsymbol{v}}}\mathcal{F}(\mu)/\|\hat{\boldsymbol{v}}\|_\mu^2\right)^-$ is discretized as in the second formula in (14). Finally, the steepest descent direction $\boldsymbol{v}^{n-1}$ is given by

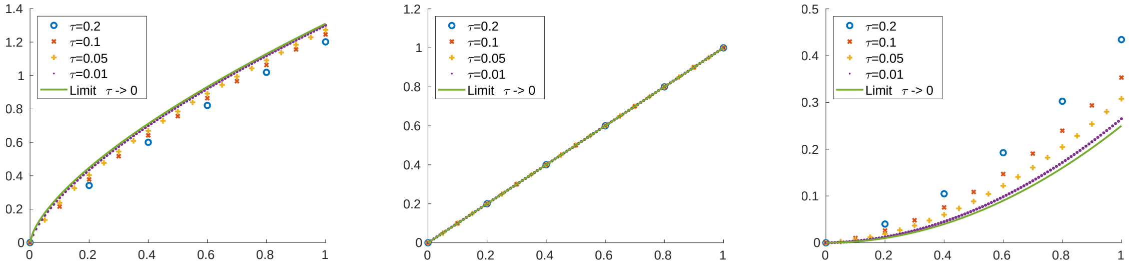
$$\boldsymbol{v}^{n-1} = \mu_\tau^{n-1} \times \mathcal{T}_{\theta,-}(x,\cdot)_\# P_Z.$$

In summary, the explicit Euler scheme of Wasserstein steepest descent flows can be implemented as in Alg. 2, which we call *neural forward scheme*.

Remark 4.1. In the case that all involved measures are absolutely continuous, it was shown in (Ambrosio et al., 2005), Theorem 12.4.4, that the geometric tangent space can be fully described by velocity fields. Then, we can use maps instead of plans for approximating the steepest descent direction and thus simplify the neural forward scheme. This might increase the approximation power of the neural forward scheme in high dimensions.

Algorithm 2 Neural forward scheme

```
Input: Samples $x_1^0, ..., x_N^0$ from $\mu_\tau^0$ and $\mathcal{F}$ in (1).
for $n = 1, 2, ...$ do
    - Learn $\mathcal{T}_\theta^{n-1}$ by minimizing the loss function (14).
    - Compute the network $\mathcal{T}_{\theta,-}^{n-1}(x,z)$ from (15) by
      $$\left(\frac{\mathbb{E}_{z_i\sim P_Z}\left[\frac{1}{N}\sum_{i=1}^N \nabla_{\mathcal{T}_\theta(x_i,z_i)} G(x_i)\right]}{\mathbb{E}_{z_i\sim P_Z}\left[\frac{1}{N}\sum_{i=1}^N \|\mathcal{T}_\theta(x_i,z_i)\|^2\right]}\right)^- \mathcal{T}_\theta(x,z).$$
    - Apply an explicit Euler step by computing for each $i$,
      $$x_i^n = x_i^{n-1} + \tau\mathcal{T}_{\theta,-}^{n-1}(x_i^{n-1}, z_i), \quad z_i \sim P_Z.$$
    - Approximate $\mu_\tau^n := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^n}$.
end for
```



(a) $r = 0.5, \quad f_\tau(n\tau) < f(n\tau)$ (b) $r = 1, \quad f_\tau(n\tau) = f(n\tau)$ (c) $r = 1.5, \quad f_\tau(n\tau) > f(n\tau)$

Figure 2: Visualization of the different convergence behavior $\gamma_\tau \to \gamma$ as $\tau \to 0$ in Theorem 5.2 via $f_\tau(n\tau) \to f(n\tau)$, $n = 0, 1, \dots$ in Remark 5.3 for $\mathcal{F} = \mathcal{E}_K$ and Riesz kernels with $r \in \{0.5, 1, 1.5\}$.

5. Flows for the Interaction Energy

For evaluating the performance of our neural schemes, we examine the Wasserstein flows of the interaction energy with Riesz kernels starting at a Dirac measure. We will provide analytical formulas for the steps in the backward and forward schemes and prove convergence of the schemes to the respective curves. In particular, we will see for the negative distance kernel the following: in two dimensions, the Wasserstein flow $\gamma(t)$, $t > 0$ starting at δ_0 becomes an absolutely continuous measure supported on the ball $t\frac{\pi}{4}\mathbb{B}^2$ with increasing density towards its boundary. In contrast, in three dimensions, the flow becomes uniformly distributed on the 2-sphere $t\frac{2}{3}\mathbb{S}^2$, i.e., it "condensates" on the surface of the ball $t\frac{2}{3}\mathbb{B}^3$.

The following analytical formula for the Wasserstein proximal mapping at a Dirac measure was proven in (Hertrich et al., 2022) based on partial results in (Carrillo & Huang, 2017; Gutleb et al., 2022; Chafaï et al., 2023). Let $\mathcal{U}_A$ denote the uniform distribution on A.

Theorem 5.1. *Let K be a Riesz kernel with $r \in (0, 2)$. Then $\operatorname{prox}_{\tau\mathcal{E}_K}(\delta_0) = (\tau^{1/(2-r)}\mathrm{Id})_\# \eta^*$, where $\eta^* := \operatorname{prox}_{\mathcal{E}_K}(\delta_0)$ is given as follows:*

(i) *For $d + r < 4$, it holds*

$$\eta^* = \rho_s \mathcal{U}_{s\mathbb{B}^d}, \quad \rho_s(x) := A_s\,(s^2 - \|x\|_2^2)^{1-\frac{r+d}{2}},$$

where $x \in s\mathbb{B}^d$ and

$$A_s := \frac{\Gamma\big(\frac{d}{2}\big)s^{-(2-r)}}{\pi^{\frac{d}{2}}B\big(\frac{d}{2}, 2-\frac{r+d}{2}\big)}, s := \left(\frac{\Gamma(2-\frac{r}{2})\,\Gamma(\frac{d+r}{2})\,r}{\frac{d}{2}\,\Gamma(\frac{d}{2})}\right)^{\frac{1}{2-r}}$$

with the Beta function B and the Gamma function Γ.

(ii) *For $d + r \geq 4$, it holds*

$$\eta^* = \mathcal{U}_{c\mathbb{S}^{d-1}}, \; c := \left(\tfrac{r}{2}\,{}_2F_1\big(-\tfrac{r}{2}, \tfrac{2-r-d}{2}; \tfrac{d}{2}; 1\big)\right)^{1/(2-r)}$$

with the hypergeometric function ${}_2F_1$.

Now the steps of the JKO scheme and its limit curve are given by the following theorem, which we prove in Appendix C.

Theorem 5.2. *Let K be a Riesz kernel with $r \in (0, 2)$, $\mathcal{F} := \mathcal{E}_K$ and $\eta^* := \operatorname{prox}_{\mathcal{E}_K}(\delta_0)$.*

(i) *Then, the measures μ_τ^n from the JKO scheme (9) starting at $\mu_\tau^0 = \delta_0$ are given by*

$$\mu_\tau^n = \big(t_{\tau,n}^{\frac{1}{2-r}}\,\mathrm{Id}\big)_\# \eta^*,$$

where $t_{\tau,0} = 0$ and $t_{\tau,n}$, $n \in \mathbb{N}$, is the unique positive zero of the function $t \mapsto t_{\tau,n-1}^{\frac{1}{2-r}} t^{\frac{1-r}{2-r}} - t + \tau$. In particular, we have for $r = 1$ that $t_{\tau,n} = n\tau$.

(ii) *The associated curves γ_τ in (9) converge for $\tau \to 0$ locally uniformly to the curve $\gamma\colon [0, \infty) \to \mathcal{P}_2(\mathbb{R}^d)$ given by*

$$\gamma(t) := ((t(2-r))^{\frac{1}{2-r}} Id)_\# \eta^*.$$

In particular, we have for $r = 1$ that $\gamma(t) = (t\,\mathrm{Id})_\# \eta^*$.

For $r \geq 1$, in (Hertrich et al., 2022) it was shown that the curve γ from part (ii) from the previous theorem is a Wasserstein steepest descent flow.

Remark 5.3 (Illustration of Theorem 5.2). By the above theorem, we can represent the curves $\gamma_\tau|_{((n-1)\tau, n\tau]} := \mu_\tau^n$ and their limit γ as $\tau \to 0$ by

$$\gamma(t) = (f(t)\,\mathrm{Id})_\# \eta^* \quad \text{and} \quad \gamma_\tau(t) = (f_\tau(t)\,\mathrm{Id})_\# \eta^*,$$

where the functions $f, f_\tau\colon [0, \infty) \to \mathbb{R}$ are given by

$$f(t) = ((2-r)t)^{\frac{1}{2-r}}, \quad \text{and} \quad f_\tau|_{((n-1)\tau, n\tau]} = t_{\tau,n}^{\frac{1}{2-r}}.$$

Hence, the convergence behavior of γ_τ to γ can be visualized by the convergence behavior of f_τ to f as in Fig. 2. The values $t_{\tau,n}$ from Theorem 5.2 are computed by Newton's method. For $r \in (0, 1)$, it holds $f_\tau(n\tau) < f(n\tau)$, $n = 1, 2, \dots$; for $r \in (1, 2)$ we have the opposite relation, and for $r = 1$ the approximation points lie on the limit curve.

The next theorem, which we prove in Appendix C, shows that also the Euler forward scheme converges for $r = 1$. Note that for $r \in (0, 1)$ there does not exist a steepest descent direction in δ_0 such that the Euler forward scheme is not well-defined. For $r \in (1, 2)$, we have $\nabla_- \mathcal{F}(\delta_0) = \delta_{0,0}$ which implies that $\mu_\tau^n = \delta_0$ for all n, i.e., γ_τ in (13) coincides with the constant curve $\gamma(t) = \delta_0$, which is a Wasserstein steepest descent flow with respect to $\mathcal{F}$, see (Hertrich et al., 2022).

Theorem 5.4. *Let K be a Riesz kernel with $r = 1$, $\mathcal{F} := \mathcal{E}_K$ and $\eta^* := \operatorname{prox}_{\mathcal{E}_K}(\delta_0)$. Then the measures μ_τ^n from the Euler forward scheme (13) starting at $\mu_\tau^0 = \delta_0$ coincides with those of the JKO scheme $\mu_\tau^n = (\tau n \operatorname{Id})_\# \eta^*$.*

6. Numerical Examples

In the following, we evaluate our results based on numerical examples. In Subsection 6.1, we benchmark the different numerical schemes based on the interaction energy flow starting at δ_0. Here, we can evaluate the quality of the outcome based on the analytic results in Sect. 5. In Subsection 6.2, we apply the different schemes for MMD Wasserstein flows. Since no ground truth is available, we can only compare the visual impression. The implementation details and advantages of the both neural schemes are given in Appendix E[1].

Comparison with Particle Flows. We compare our neural forward and backward schemes with particle flows. The main idea is to approximate the Wasserstein flow with respect to a function $\mathcal{F}$ by the gradient flow with respect to the functional $F_M(x_1, ..., x_M) = \mathcal{F}(\frac{1}{M}\sum_{i=1}^M \delta_{x_i})$, where $x_1, ..., x_M \in \mathbb{R}^d$ are distinct particles. We include a detailed description in Appendix D. For smooth and λ-convex kernels, such flows were considered in (Arbel et al., 2019). In this particular case, the authors showed that MMD flows starting in point measures can be fully described by this representation. However, for the Riesz kernels, this is no longer true. Instead, we show in Appendix D that the particle flow is a Wasserstein gradient flow *but with respect to a restricted functional*. Nevertheless, the mean field limit $M \to \infty$ may provide a meaningful approximation of Wasserstein gradient flows with respect to $\mathcal{F}$.

However, for computing the particle flow, the assumption $x_i \neq x_j$ for $i \neq j$ is crucial. Consequently, it is not possible to simulate a particle flow starting in a Dirac measure. As a remedy, we start the particle flow in M samples randomly located in a very small area around the initial point. The optimal initial structure of the initial points depends on the choice of the functional $\mathcal{F}$ and is non-trivial to compute. For a detailed analysis of the influence of the initial point distribution, we refer to Appendix F. We will observe that particle flows provide a reasonable baseline for the approximation of Wasserstein flows even though the initial distribution of the samples significantly influences the the results.

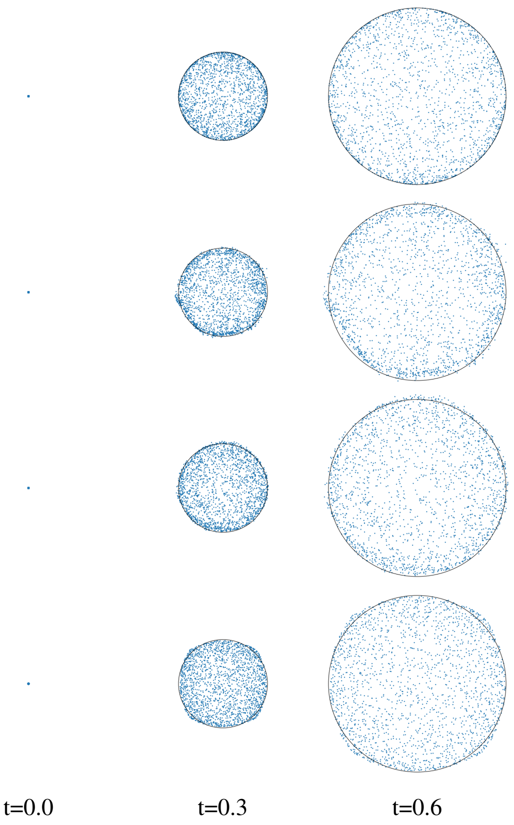


Figure 3: Comparison of different approaches for approximating the Wasserstein gradient flow of $\mathcal{E}_K$ with step size $\tau = 0.05$. **From top to bottom**: limit curve, neural backward scheme, neural forward scheme and particle flow. The black circle is the border of the limit $\operatorname{supp} \gamma(t)$. Here the forward flow shows the best fit. While our neural flows start in a single point, the particle flow starts with uniform samples in a square of radius 10^{-9}, a structure which remains visible over the time.

6.1. Interaction Energy Flows with Benchmark

We compare the different approaches for simulating the Wasserstein flow of the interaction energy $\mathcal{F} = \mathcal{E}_K$ starting in δ_0.

A visual comparison in two dimensions is given in Fig. 3. While our neural schemes start in a single point, the particle flow starts with uniformly distributed particles in a square of size 10^{-9}. This square structure remains visible over the time. For particle flows with other starting point

[1] The code is available at https://github.com/FabianAltekrueger/NeuralWassersteinGradientFlows

configurations, see Appendix F.

A quantitative comparison between the analytic flow and its approximation with the discrepancy $\mathcal{D}_K^2$ (negative distance kernel) as distance is given in Fig. 4. The time step size is again $\tau = 0.05$ and simulated 10000 samples. In the left part of the figure, we compare the different approaches for $d = 2$ and different Riesz kernel exponents. For the Riesz exponent $r = 1$ the neural forward scheme gives the best approximation of the Wasserstein flow. While for $r = 0.5$ the neural backward scheme and the particle flow approximate the limit curve nearly similarly, the neural backward scheme performs better for $r = 1.5$. The poor approximation ability of the particle flow can be explained by the inexact starting measure and the relatively high step size $\tau = 0.05$. Reducing the step size leads to an improvement of the approximation. As outlined in the text before Theorem 5.4, the neural forward scheme is not defined for $r \neq 1$. The right part of Fig. 4 shows results for $r = 1$ and different dimensions d. While in the three-dimensional case, the particle flow is not able to push the particles from the initial cube onto the sphere (condensation), for higher dimensions it approximates the limit curve almost perfectly. For the two network-based methods a higher dimension leads to a higher approximation error.

6.2. MMD Flows

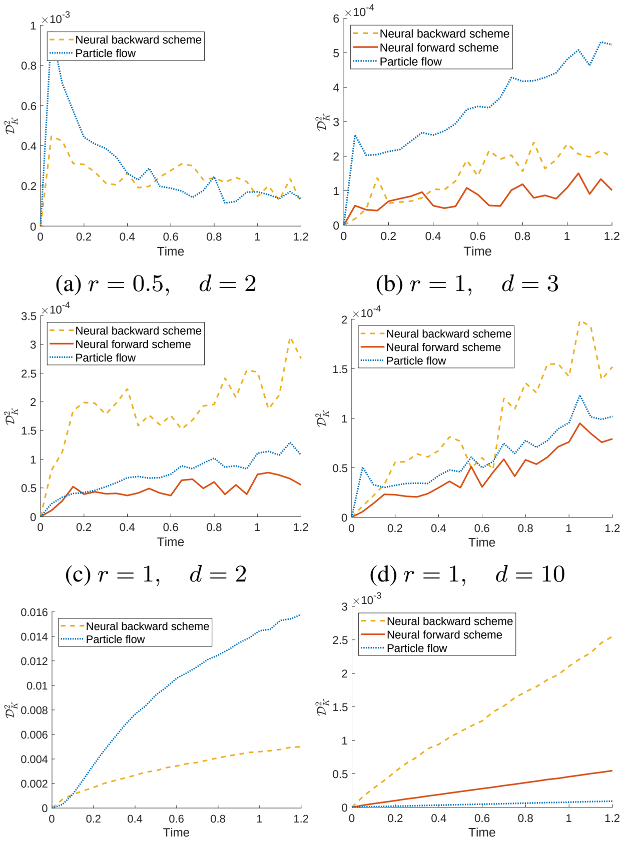
Next, we consider MMD Wasserstein flows $\mathcal{F}_\nu$. We can use the proposed methods to sample from a target measure ν which is given by some samples as it was already shown in Fig. 1 'Max und Moritz' in the introduction with 6000 particles and $\tau = 0.5$. More examples are given in Appendix H.

In order to show the scalability of the methods, we can use the proposed methods to sample from the MNIST dataset (LeCun et al., 1998). Each 28×28 MNIST digit can be interpreted as a point in the 784 dimensional space such that our target measure ν is a weighted sum of Dirac measures. Here we only use the first 100 digits of MNIST for ν. Fig. 5 illustrates samples and their trajectories using the proposed methods. The effect of the inexact starting of the particle flow can be seen in the trajectory of the particle flow, where the first images of the trajectory contain a lot of noise. For more details, we refer to Appendix E. In Appendix I we illustrate the same example when starting in an absolutely continuous measure instead of a singular measure.

7. Conclusions

We introduced neural forward and backward schemes to approximate Wasserstein flows of MMDs with non-smooth Riesz kernels. Both neural schemes are realized by approximating the disintegration of transport plans and velocity plans. This enables us to handle non-absolutely continuous measures, which were excluded in prior works. In order to benchmark the schemes, we derive analytic formulas for the schemes with respect to the interaction energy starting at Dirac measures. Finally, the performance of our neural approximations was demonstrated by numerical examples. Here, additionally particle flows were considered, which show a good performance as well, but may depend on the start distribution of the points.

Our work can be extended in different directions. Even

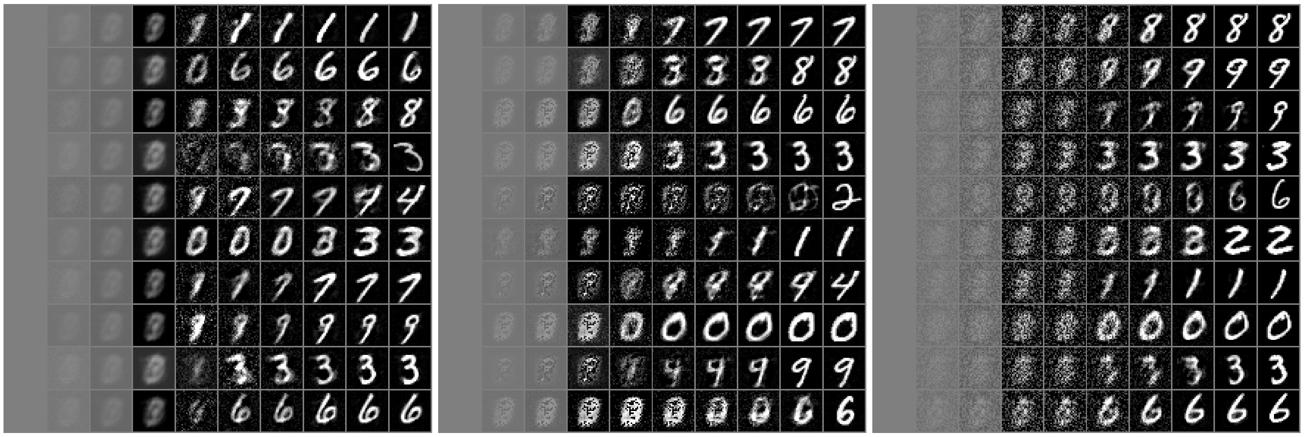


(a) $r = 0.5, \quad d = 2$ (b) $r = 1, \quad d = 3$

(c) $r = 1, \quad d = 2$ (d) $r = 1, \quad d = 10$

(e) $r = 1.5, \quad d = 2$ (f) $r = 1, \quad d = 1000$

Figure 4: Discrepancy between the analytic Wasserstein flow of $\mathcal{E}_K$ and its approximations for $\tau = 0.05$. **Left**: dimension $d = 2$ and different exponents of the Riesz kernel. Note that the neural forward flow only exists for $r = 1$, where it gives the best approximation. For $r = 0.5$ the neural backward scheme and the particle flow approximate the limit curve nearly similar, while the neural backward scheme performs better for $r = 1.5$ which is due to the relatively large time step size. **Right**: Different dimensions $d \in \{3, 10, 1000\}$ and $r = 1$. While the particle flow suffers from the inexact initial samples in lower dimensions, it performs very well in higher dimensions. The neural forward scheme gives a more accurate approximation than the neural backward scheme.



Neural backward scheme

Neural forward scheme

Particle flow

Figure 5: Samples and their trajectories from MNIST starting in δ_x for $x = 0.5 \cdot \mathbf{1}_{784}$. The inexact starting of the particle flow leads to noisy images at the beginning.

though the forward scheme converges nicely in all our numerical examples, it would be desirable to derive both analytic formulas for steepest descent directions as well as an analytic convergence result for (13) for other functions than the interaction energy. It will be also interesting to restrict measures to certain supports, as, e.g., curves and to examine corresponding flows. Moreover, we aim to extend our framework to posterior sampling in a Bayesian setting. Here a sampling based approach appears to be useful for several applications. So far we used fully connected NNs for approximating the corresponding measures. Nevertheless, the usage of convolutional NNs could be a key ingredient for applying the proposed methods on image data. Finally, we can use the findings from (Hertrich et al., 2023b) and compute the MMD functional by its sliced version. We hope that this can lead to a significant acceleration of our proposed schemes.

Acknowledgement

F.A. acknowledges support from the German Research Foundation (DFG) under Germany‘s Excellence Strategy – The Berlin Mathematics Research Center MATH+ within the project EF3-7 and J.H. by the DFG within the project STE 571/16-1. We would like to thank Robert Beinert and Manuel Gräf for fruitful discussions on Wasserstein gradient flows.

References

Alvarez-Melis, D., Schiff, Y., and Mroueh, Y. Optimizing functionals on the space of probabilities with input convex neural networks. *Transactions on Machine Learning Research*, 2022.

Ambrosio, L., Gigli, N., and Savare, G. *Gradient Flows*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 2005. ISBN 978-3-7643-2428-5.

Amos, B., Xu, L., and Kolter, J. Z. Input convex neural networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 146–155. PMLR, 2017.

Ansari, A. F., Ang, M. L., and Soh, H. Refining deep generative models via discriminator gradient flow. In *International Conference on Learning Representations*, 2021.

Arbel, M., Korba, A., Salim, A., and Gretton, A. Maximum mean discrepancy gradient flow. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 1–11, New York, USA, 2019.

Arbel, M., Gretton, A., Li, W., and Montufar, G. Kernelized wasserstein natural gradient. In *International Conference on Learning Representations*, 2020.

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

Bonet, C., Courty, N., Septier, F., and Drumetz, L. Efficient gradient flows in sliced-Wasserstein space. *Transactions on Machine Learning Research*, 2022.

Brehmer, J. and Cranmer, K. Flows for simultaneous manifold learning and density estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 442–453. Curran Associates, Inc., 2020.

Brenier, Y. Décomposition polaire et réarrangement monotone des champs de vecteurs. *Comptes Rendus de l'Académie des Sciences Paris Series I Mathematics*, 305 (19):805–808, 1987.

Bunne, C., Papaxanthos, L., Krause, A., and Cuturi, M. Proximal optimal transport modeling of population dynamics. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6511–6528. PMLR, 2022.

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

Carrillo, J. A., Craig, K., Wang, L., and Wei, C. Primal dual methods for Wasserstein gradient flows. *Foundations of Computational Mathematics*, 22(2):389–443, 2022.

Chafaï, D., Saff, E., and Womersley, R. Threshold condensation to singular support for a Riesz equilibrium problem. *Analysis and Mathematical Physics*, 13(19), 2023.

Chen, Y. and Li, W. Optimal transport natural gradient for statistical manifolds with continuous sample space. *Information Geometry*, 3(1):1–32, 2020.

Cohen, S., Arbel, M., and Deisenroth, M. P. Estimating barycenters of measures in high dimensions. *arXiv preprint arXiv:2007.07105*, 2021.

di Langosco, L. L., Fortuin, V., and Strathmann, H. Neural variational gradient descent. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2022.

Dong, H., Wang, X., Yong, L., and Zhang, T. Particle-based variational inference with preconditioned functional gradient flow. In *The Eleventh International Conference on Learning Representations*, 2023.

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

Fan, J., Zhang, Q., Taghvaei, A., and Chen, Y. Variational Wasserstein gradient flow. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6185–6215. PMLR, 2022.

Gao, Y., Jiao, Y., Wang, Y., Wang, Y., Yang, C., and Zhang, S. Deep generative learning via variational gradient flow. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2093–2101. PMLR, 2019.

Gigli, N. *On the geometry of the space of probability measures endowed with the quadratic optimal transport distance*. PhD thesis, Scuola Normale Superiore di Pisa, 2004.

Giorgi, E. D. New problems on minimizing movements. In Ciarlet, P. and Lions, J.-L. (eds.), *Boundary Value Problems for Partial Differential Equations and Applications*, pp. 81–98. Masson, 1993.

Glaser, P., Arbel, M., and Gretton, A. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8018–8031. Curran Associates, Inc., 2021.

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

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D. K., and Zemel, R. S. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, 2020.

Gutleb, T. S., Carrillo, J. A., and Olver, S. Computation of power law equilibrium measures on balls of arbitrary dimension. *Constructive Approximation*, 2022.

Hagemann, P., Hertrich, J., and Steidl, G. Stochastic normalizing flows for inverse problems: A Markov chain viewpoint. *SIAM Journal on Uncertainty Quantification*, 10:1162–1190, 2022.

Hagemann, P., Hertrich, J., and Steidl, G. *Generalized normalizing flows via Markov chains*. Series: Elements in Non-local Data Interactions: Foundations and Applications. Cambridge University Press, 2023.

Hertrich, J., Gräf, M., Beinert, R., and Steidl, G. Wasserstein steepest descent flows of discrepancies with Riesz kernels. *arXiv preprint arXiv:2211.01804*, 2022.

Hertrich, J., Beinert, R., Gräf, M., and Steidl, G. Wasserstein gradient flows of the discrepancy with distance kernel on the line. In *Scale Space and Variational Methods in Computer Vision*, pp. 431–443. Springer, 2023a.

Hertrich, J., Wald, C., Altekrüger, F., and Hagemann, P. Generative sliced MMD flows with Riesz kernels. *arXiv preprint arXiv:2305.11463*, 2023b.

Hwang, H. J., Kim, C., Park, M. S., and Son, H. The deep minimizing movement scheme. *arXiv preprint arXiv:2109.14851*, 2021.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lin, A. T., Li, W., Osher, S., and Montúfar, G. Wasserstein proximal of GANs. In Nielsen, F. and Barbaresco, F. (eds.), *Geometric Science of Information*, pp. 524–533, Cham, 2021. Springer International Publishing.

Lu, G., Zhou, Z., Shen, J., Chen, C., Zhang, W., and Yu, Y. Large-scale optimal transport via adversarial training with cycle-consistency. *arXiv preprint arXiv:2003.06635*, 2020.

Mattila, P. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Cambridge University Press, 1999.

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J. M., and Burnaev, E. Large-scale wasserstein gradient flows. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15243–15256, 2021.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26:101–174, 2001.

Otto, F. and Westdickenberg, M. Eulerian calculus for the contraction in the Wasserstein distance. *SIAM Journal on Mathematical Analysis*, 37(4):1227–1255, 2005.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, 2019.

Pavliotis, G. A. *Stochastic processes and applications: Diffusion Processes, the Fokker-Planck and Langevin Equations*. Number 60 in Texts in Applied Mathematics. Springer, New York, 2014.

Rockafellar, R. T. and Royset, J. O. Random variables, monotone relations, and convex analysis. *Mathematical Programming*, 148:297–331, 2014. doi: 10.1007/s10107-014-0801-1.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Teuber, T., Steidl, G., Gwosdek, P., Schmaltz, C., and Weickert, J. Dithering by differences of convex functions. *SIAM Journal on Imaging Sciences*, 4(1):79–108, 2011.

Villani, C. *Topics in Optimal Transportation*. Number 58 in Graduate Studies in Mathematics. American Mathematical Society, Providence, 2003.

Welling, M. and Teh, Y.-W. Bayesian learning via stochastic gradient Langevin dynamics. In Getoor, L. and Scheffer, T. (eds.), *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 681–688, Madison, 2011.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2005.

Wu, H., Köhler, J., and Noe, F. Stochastic normalizing flows. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5933–5944, 2020.

A. Wasserstein Spaces as Geodesic Spaces

A curve $\gamma\colon I \to \mathcal{P}_2(\mathbb{R}^d)$ on an interval $I \subset \mathbb{R}$ is called a *geodesics*, if there exists a constant $C \geq 0$ such that $W_2(\gamma(t_1), \gamma(t_2)) = C|t_2 - t_1|$ for all $t_1, t_2 \in I$. The Wasserstein space is a geodesic space, meaning that any two measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ can be connected by a geodesics.

For $\lambda \in \mathbb{R}$, a function $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ is called λ-*convex along geodesics* if, for every $\mu, \nu \in \operatorname{dom} \mathcal{F} := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathcal{F}(\mu) < \infty\}$, there exists at least one geodesics $\gamma\colon [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ between μ and ν such that

$$\mathcal{F}(\gamma(t)) \leq (1-t)\,\mathcal{F}(\mu) + t\,\mathcal{F}(\nu) - \tfrac{\lambda}{2}\, t(1-t)\, W_2^2(\mu,\nu), \qquad t \in [0,1].$$

Every function being λ-convex along generalized geodesics is also λ-convex along geodesics since generalized geodesics with base $\sigma = \mu$ are actual geodesics. To ensure uniqueness and convergence of the JKO scheme, a slightly stronger condition, namely being λ-*convex along generalized geodesics* will be in general needed. Based on the set of three-plans with base $\sigma \in \mathcal{P}_2(\mathbb{R}^d)$ given by

$$\Gamma_\sigma(\mu,\nu) := \big\{\boldsymbol{\alpha} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d) : (\pi_1)_\#\boldsymbol{\alpha} = \sigma, (\pi_2)_\#\boldsymbol{\alpha} = \mu, (\pi_3)_\#\boldsymbol{\alpha} = \nu\big\},$$

the so-called *generalized geodesics* $\gamma\colon [0,\epsilon] \to \mathcal{P}_2(\mathbb{R}^d)$ joining μ and ν (with base σ) is defined as

$$\gamma(t) := \big((1 - \tfrac{t}{\epsilon})\pi_2 + \tfrac{t}{\epsilon}\pi_3\big)_\#\boldsymbol{\alpha}, \qquad t \in [0,\epsilon], \tag{16}$$

where $\boldsymbol{\alpha} \in \Gamma_\sigma(\mu,\nu)$ with $(\pi_{1,2})_\#\boldsymbol{\alpha} \in \Gamma^{\mathrm{opt}}(\sigma,\mu)$ and $(\pi_{1,3})_\#\boldsymbol{\alpha} \in \Gamma^{\mathrm{opt}}(\sigma,\nu)$, see Definition 9.2.2 in (Ambrosio et al., 2005). Here $\Gamma^{\mathrm{opt}}(\mu,\nu)$ denotes the set of optimal transport plans $\boldsymbol{\pi}$ realizing the minimum in (3). The plan $\boldsymbol{\alpha}$ may be interpreted as transport from μ to ν via σ. Then a function $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ is called λ-*convex along generalized geodesics* (Ambrosio et al., 2005), Definition 9.2.4, if for every $\sigma, \mu, \nu \in \operatorname{dom}\mathcal{F}$, there exists at least one generalized geodesics $\gamma\colon [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ related to some $\boldsymbol{\alpha}$ in (16) such that

$$\mathcal{F}(\gamma(t)) \leq (1-t)\,\mathcal{F}(\mu) + t\,\mathcal{F}(\nu) - \tfrac{\lambda}{2}\, t(1-t)\, W_{\boldsymbol{\alpha}}^2(\mu,\nu), \qquad t \in [0,1], \tag{17}$$

where

$$W_{\boldsymbol{\alpha}}^2(\mu,\nu) := \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|y-z\|_2^2 \,\mathrm{d}\boldsymbol{\alpha}(x,y,z).$$

Wasserstein spaces are manifold-like spaces. In particular, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (Ambrosio et al., 2005), § 8, there exists the *regular tangent space* at μ, which is defined by

$$\mathrm{T}_\mu\mathcal{P}_2(\mathbb{R}^d) := \overline{\{\nabla\phi : \phi \in C_{\mathrm{c}}^\infty(\mathbb{R}^d)\}}^{L_2(\mu,\mathbb{R}^d)}.$$

Note that $\mathrm{T}_\mu\mathcal{P}_2(\mathbb{R}^d)$ is an infinite dimensional subspace of $L_2(\mu,\mathbb{R}^d)$ if $\mu \in \mathcal{P}_2^r(\mathbb{R}^d)$ and it is just $\mathbb{R}^d$ if $\mu = \delta_x$, $x \in \mathbb{R}^d$

For a proper and lsc function $\mathcal{F}\colon \mathcal{P}_2(\mathbb{R}^d) \to (-\infty,\infty]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, the *reduced Fréchet subdiffential at* μ is defined as

$$\partial\mathcal{F}(\mu) := \Big\{\xi \in L_{2,\mu} : \mathcal{F}(\nu) - \mathcal{F}(\mu) \geq \inf_{\pi\in\Gamma^{\mathrm{opt}}(\mu,\nu)} \int\limits_{\mathbb{R}^{2d}} \langle \xi(x), y - x\rangle \,\mathrm{d}\pi(x,y) + o(W_2(\mu,\nu))\ \forall \nu \in \mathcal{P}_2(\mathbb{R}^d)\Big\}. \tag{18}$$

For general $\mathcal{F}$, the velocity field $v_t \in \mathrm{T}_{\gamma(t)}\mathcal{P}_2(\mathbb{R}^d)$ in (5) is only determined for almost every $t > 0$, but we want to give a definition of the steepest descent flows pointwise. We equip the space of velocity plans $\boldsymbol{V}(\mu) := \{\boldsymbol{v} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) : (\pi_1)_\#\boldsymbol{v} = \mu\}$ with the metric W_μ defined by

$$W_\mu^2(\boldsymbol{v},\boldsymbol{w}) := \inf_{\boldsymbol{\alpha}\in\Gamma_\mu(\boldsymbol{v},\boldsymbol{w})} W_{\boldsymbol{\alpha}}^2((\pi_2)_\#\boldsymbol{v}, (\pi_2)_\#\boldsymbol{w}),$$

Then the *geometric tangent space* at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ is given by

$$\mathbf{T}_\mu\mathcal{P}_2(\mathbb{R}^d) := \overline{\boldsymbol{G}(\mu)}^{W_\mu},$$

where

$$\boldsymbol{G}(\mu) := \big\{ \boldsymbol{v} \in \boldsymbol{V}(\mu) : \exists \epsilon > 0 \text{ such that } \boldsymbol{\pi} = (\pi_1, \pi_1 + \tfrac{1}{\epsilon}\pi_2)_\# \boldsymbol{v} \in \Gamma^{\mathrm{opt}}(\mu, (\pi_2)_\# \boldsymbol{\pi}) \big\}$$

consists of all geodesic directions at $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ (correspondingly all geodesics starting in $\mu \in \mathcal{P}_2(\mathbb{R}^d)$). We define the *exponential map* $\exp_\mu \colon \mathbf{T}_\mu \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ by

$$\exp_\mu(\boldsymbol{v}) := \gamma_{\boldsymbol{v}}(1) = (\pi_1 + \pi_2)_\# \boldsymbol{v}.$$

The *"inverse" exponential map* $\exp_\mu^{-1} \colon \mathcal{P}_2(\mathbb{R}^d) \to \mathrm{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$ is given by the (multivalued) function

$$\exp_\mu^{-1}(\nu) := \big\{ (\pi_1, \pi_2 - \pi_1)_\# \boldsymbol{\pi} : \boldsymbol{\pi} \in \Gamma^{\mathrm{opt}}(\mu, \nu) \big\}$$

and consists of all velocity plans $\boldsymbol{v} \in \boldsymbol{V}(\mu)$ such that $\gamma_{\boldsymbol{v}}|_{[0,1]}$ is a geodesics connecting μ and ν. For a curve $\gamma \colon I \to \mathcal{P}_2(\mathbb{R}^d)$, a velocity plan $\boldsymbol{v}_t \in \mathbf{T}_{\gamma(t)} \mathcal{P}_2(\mathbb{R}^d)$ is called a *(geometric) tangent vector* of γ at $t \in I$ if, for every $h > 0$ and $\boldsymbol{v}_{t,h} \in \exp_{\gamma(t)}^{-1}(\gamma(t+h))$, it holds

$$\lim_{h \to 0+} W_{\gamma(t)}(\boldsymbol{v}_t, \tfrac{1}{h} \cdot \boldsymbol{v}_{t,h}) = 0.$$

If a tangent vector $\boldsymbol{v}_t$ exists, then, since $W_{\gamma(t)}$ is a metric on $\boldsymbol{V}(\gamma(t))$, the above limit is uniquely determined, and we write

$$\dot\gamma(t) := \boldsymbol{v}_t.$$

In Theorem 4.19 in (Gigli, 2004) it is shown that $\dot\gamma_{\boldsymbol{v}}(0) = \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbf{T}_\mu \mathcal{P}_2(\mathbb{R}^d)$. Therefore, the definition of a tangent vector of a curve is consistent with the interpretation of $\gamma_{\boldsymbol{v}}$ as a curve in direction of $\boldsymbol{v}$. For $\boldsymbol{v} \in \boldsymbol{G}(\mu)$, we can also compute the (geometric) tangent vector of a geodesics $\gamma_{\boldsymbol{v}}$ on $[0, \epsilon]$ by $\dot\gamma_{\boldsymbol{v}}(t) = (\pi_1 + t\,\pi_2, \pi_2)_\# \boldsymbol{v}$, $t \in [0, \epsilon)$.

B. Disintegration of measures

Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel algebra of $\mathbb{R}^d$. A map $k \colon \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0, \infty]$ is called Markov kernel, if $k(x, \cdot) \in \mathcal{P}(\mathbb{R}^d)$ for all $x \in \mathbb{R}^d$ and $k(\cdot, A)$ is measurable for all $A \in \mathcal{B}(\mathbb{R}^d)$. Next we state the disintegration theorem, see, e.g., Theorem 5.3.1 in (Ambrosio et al., 2005).

Theorem B.1. *Let $\boldsymbol{\pi} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ and assume that $\pi_{1\#}\boldsymbol{\pi} = \mu \in \mathcal{P}_2(\mathbb{R}^d)$. Then there exists a μ-a.e. uniquely determined family of probability measures $(\pi_x)_{x \in \mathbb{R}^d} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ such that for all functions $f \colon \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty]$ it holds*

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} f(x,y) \mathrm{d}\boldsymbol{\pi}(x,y) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(x,y) \mathrm{d}\pi_x(y) \mathrm{d}\mu(x).$$

Note that the family of probability measures $(\pi_x)_{x \in \mathbb{R}^d} \subseteq \mathcal{P}_2(\mathbb{R}^d)$ in Theorem B.1 can be described by a Markov kernel $x \to \pi_x$.

C. Proof of Theorems 5.2 and 5.4

The proofs rely on the fact that all measures μ_τ^n computed by the JKO and forward schemes (9) for the function $\mathcal{F} = \mathcal{E}_K$ with Riesz kernel K which start at a point measure are orthogonally invariant (radially symmetric). We prove that fact first. This implies in Proposition C.3 that we can restrict our attention to flows on $\mathcal{P}_2(\mathbb{R})$, where the Wasserstein distance is just defined via quantile functions.

Proposition C.1. *Let $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ be orthogonally invariant and $\mathcal{F} := \mathcal{E}_K$ for the Riesz kernel K with $r \in (0, 2)$. Then, any measure $\mu_* \in \mathrm{prox}_{\tau\mathcal{F}}(\nu)$ is orthogonally invariant.*

Proof. Fix $\tau > 0$ and assume that $\mu_* \in \mathrm{prox}_{\tau\mathcal{F}}(\nu)$ is not orthogonally invariant. Then we can find an orthogonal matrix $O \in O(d)$ such that $\mu_* \neq O_\# \mu_*$. Define

$$\tilde\mu := \tfrac{1}{2}\mu_* + \tfrac{1}{2} O_\# \mu_*.$$

Then, for an optimal plan $\boldsymbol{\pi} \in \Gamma^{\mathrm{opt}}(\nu, \mu_*)$, we take the radial symmetry of ν into account and consider

$$\tilde{\boldsymbol{\pi}} := \tfrac{1}{2}\boldsymbol{\pi} + \tfrac{1}{2}(O\pi_1, O\pi_2)_\# \boldsymbol{\pi} \in \Gamma(\nu, \tilde\mu).$$

Now it follows

$$\begin{aligned} W_2^2(\nu,\tilde\mu) &\le \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\tilde{\boldsymbol{\pi}}(x,y)\\ &= \tfrac12 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\boldsymbol{\pi}(x,y) + \tfrac12 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|Ox-Oy\|_2^2 \,\mathrm{d}\boldsymbol{\pi}(x,y)\\ &= W_2^2(\nu,\mu_*). \end{aligned}$$

By definition of $\mathcal{E}_K$ we have further orthogonal invariance the Euclidean distance

$$\begin{aligned} \mathcal{E}_K(\mu_*) &= \mathcal{E}_K\left(\tilde\mu + \tfrac12\mu_* - \frac12 O_\# \mu_*\right)\\ &= \mathcal{E}_K(\tilde\mu) + \mathcal{E}_K\left(\frac12 \mu_* - \frac12 O_\#\mu_*\right) - \frac12 \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (\|x-y\|_2^r - \|x-Oy\|_2^r)\,\mathrm{d}\tilde\mu(x)\,\mathrm{d}\mu_*(y)\\ &= \mathcal{E}_K(\tilde\mu) + \tfrac14 \mathcal{E}_K(\mu_* - O_\#\mu_*). \end{aligned}$$

Since $\mu_* \neq O_\#\mu_*$ and $\mathcal{D}_K^2(\mu,\nu) = 0$ if and only if $\mu=\nu$, we infer that $\mathcal{E}_K(\mu_* - O_\#\mu_*) = \mathcal{D}_K^2(\mu_*, O_\#\mu_*) > 0$, which implies

$$\tfrac{1}{2\tau} W_2^2(\nu,\tilde\mu) + \mathcal{F}(\tilde\mu) < \tfrac{1}{2\tau} W_2^2(\nu,\mu_*) + \mathcal{F}(\mu_*).$$

This contradicts the assertion that $\mu_* \in \operatorname{prox}_{\tau\mathcal{F}}(\nu)$ and concludes the proof. □

In the following, we embed the set of orthogonally (radially) symmetric measures with finite second moment

$$\operatorname{RS}(\mathbb{R}^d) \coloneqq \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : O_\#\mu = \mu \text{ for all } O \in O(d)\} \subseteq \mathcal{P}_2(\mathbb{R}^d)$$

isometrically into $L^2((0,1))$. Here, we proceed in two steps. First, we embed the $\operatorname{RS}(\mathbb{R}^d)$ isometrically in the set of one-dimensional probability measures $\mathcal{P}_2(\mathbb{R})$.

One-dimensional probability measures can be identified by their quantile function (Rockafellar & Royset, 2014), § 1.1. More precisely, the *cumulative distribution function* $F_\mu \colon \mathbb{R} \to [0,1]$ of $\mu \in \mathcal{P}_2(\mathbb{R})$ is defined by

$$F_\mu(x) \coloneqq \mu((-\infty,x]), \qquad x \in \mathbb{R}.$$

It is non-decreasing and right-continuous with $\lim_{x\to-\infty} F_\mu(x) = 0$ and $\lim_{x\to\infty} F_\mu(x) = 1$. The *quantile function* $Q_\mu \colon (0,1) \to \mathbb{R}$ is the generalized inverse of F_μ given by

$$Q_\mu(p) \coloneqq \min\{x \in \mathbb{R} \;:\; F_\mu(x) \ge p\}, \qquad p \in (0,1). \tag{19}$$

It is non-decreasing and left-continuous. By the following theorem, the mapping $\mu \mapsto Q_\mu$ is an isometric embedding of $\mathcal{P}_2(\mathbb{R})$ into $L_2((0,1))$.

Theorem C.2 (Theorem 2.18 in (Villani, 2003))**.** *Let $\mu,\nu \in \mathcal{P}_2(\mathbb{R})$. Then the quantile function satisfies*

$$Q_\mu \in \mathcal{C}((0,1)) \subset L_2((0,1)) \qquad \textit{and} \qquad \mu = (Q_\mu)_\# \lambda_{(0,1)}, \tag{20}$$

with the cone $\mathcal{C}((0,1)) \coloneqq \{Q \in L_2((0,1)) : Q \text{ nondecreasing}\}$ and

$$W_2^2(\mu,\nu) = \int_0^1 |Q_\mu(s) - Q_\nu(s)|^2 \mathrm{d}s.$$

Using this theorem, we can now embed $\operatorname{RS}(\mathbb{R}^d)$ isometrically into $L_2((0,1))$ by the following proposition.

Proposition C.3. *The mapping $\iota\colon \operatorname{RS}(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R})$ defined by $\iota(\mu) = (\|\cdot\|_2)_\#\mu$ is an isometry from $(\operatorname{RS}(\mathbb{R}^d), W_2)$ to $(\mathcal{P}_2(\mathbb{R}), W_2)$ with range $\mathcal{P}_2(\mathbb{R}_{\ge 0}) \subseteq \mathcal{P}_2(\mathbb{R})$. Moreover, the inverse $\iota^{-1} : \mathcal{P}_2(\mathbb{R}_{\ge0}) \to \operatorname{RS}(\mathbb{R}^d)$ is given by*

$$\iota^{-1}(\tilde\mu)(A) = \mu(A) = \int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(x)\,\mathrm{d}\mathcal{U}_{\partial B_r(0)}(x)\mathrm{d}\tilde\mu(r), \quad A \in \mathcal{B}(\mathbb{R}^d), \tag{21}$$

where $B_r(0)$ is the ball in $\mathbb{R}^d$ around 0 with radius r. The mapping

$$\Psi : \mathrm{RS}(\mathbb{R}^d) \to \mathcal{C}_0((0,1)), \quad -\mu \mapsto Q_{\iota(\mu)}$$

to the convex cone $\mathcal{C}_0((0,1)) \coloneqq \{Q \in L^2((0,1)) : Q$ is non-decreasing and $Q \geq 0\} \subset L_2((0,1))$ with the quantile functions $Q_{\iota(\mu)}$ defined in (19), is a bijective isometry. In particular, it holds for all $\mu, \nu \in \mathrm{RS}(\mathbb{R}^d)$ that

$$W_2(\mu,\nu) = \int_0^1 (f(s) - g(s))^2 \,\mathrm{d}s, \quad f \coloneqq \Psi(\mu),\ g \coloneqq \Psi(\nu).$$

Proof. 1. First, we show the inversion formula (21). Let $\mu \in \mathrm{RS}(\mathbb{R}^d)$ and $\tilde{\mu} = (\|\cdot\|_2)_\# \mu$. Then, we obtain by the transformation $x \to (\frac{x}{\|x\|_2}, \|x\|_2)$ that there exist $\tilde{\mu}$-almost everywhere unique measures μ_r on $\partial B_r(0)$ such that for all $A \in \mathcal{B}(\mathbb{R}^d)$,

$$\mu(A) = \int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(x) \mathrm{d}\mu_r(x) \mathrm{d}\tilde{\mu}(r).$$

Since μ is orthogonally invariant, we obtain for any $O \in O(d)$ that

$$\begin{aligned}&\int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(x)\, \mathrm{d}\mu_r(x) \mathrm{d}\tilde{\mu}(r) = \mu(A) = O_\#\mu(A)\\ &= \int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(Ox)\, \mathrm{d}\mu_r(x) \mathrm{d}\tilde{\mu}(r) = \int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(x)\, \mathrm{d}O_\#\mu_r(x) \mathrm{d}\tilde{\mu}(r).\end{aligned}$$

Due to the uniqueness of the μ_r, we have $\tilde{\mu}$-almost everywhere that $O_\#\mu_r = \mu_r$ for all $O \in \mathrm{O}(d)$. By § 3, Theorem 3.4 in (Mattila, 1999), this implies that $\mu_r = U_{\partial B_r(0)}$. Hence, we have that

$$\mu(A) = \int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(x)\, \mathrm{d}\mathcal{U}_{\partial B_r(0)}(x) \mathrm{d}\tilde{\mu}(r), \tag{22}$$

which proves (21) and the statement about the range of ι.

2. Next, we show the isometry property. Let $\mu, \nu \in \mathrm{RS}(\mathbb{R}^d)$, $\tilde{\mu} = \iota(\mu)$, $\tilde{\nu} = \iota(\nu)$ and $\pi \in \Gamma^{\mathrm{opt}}(\mu,\nu)$. Then, it holds for $\tilde{\pi} \coloneqq (\|\pi_1 \cdot\|_2, \|\pi_2 \cdot\|_2)_\#\pi \in \Gamma(\tilde{\mu},\tilde{\nu})$ that

$$\begin{aligned}W_2^2(\mu,\nu) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\pi(x,y) \geq \int_{\mathbb{R}^d\times\mathbb{R}^d} (\|x\|_2 - \|y\|_2)^2 \,\mathrm{d}\pi(x,y)\\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} (x-y)^2 \,\mathrm{d}\tilde{\pi}(x,y) \geq W_2^2(\tilde{\mu},\tilde{\nu}).\end{aligned}$$

To show the reverse direction let $\tilde{\pi} \in \Gamma^{opt}(\tilde{\mu},\tilde{\nu})$ and define $\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ by $\pi_{1\#}\pi = \mu$ and the disintegration

$$\pi_x(A) = \tilde{\pi}_{\|x\|_2}(\{c \geq 0 : c\tfrac{x}{\|x\|_2} \in A\}).$$

In the following, we show that $\pi_{2\#}\pi = \nu$ so that $\pi \in \Gamma(\mu,\nu)$. Let $A \in \mathcal{B}(\mathbb{R}^d)$ be given by $A \coloneqq \{cx : c \in [a,b], x \in B\}$ for some $B \in \partial B_1(0)$. We show that $\pi_{2\#}\pi(A) = \nu(A)$. As the set of all such A is a $\cap$-stable generator of $\mathcal{B}(\mathbb{R}^d)$, this implies that $\pi_{2\#}\pi = \nu$. By definition, it holds

$$\int_{\mathbb{R}^d} \pi_x(A)\,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \tilde{\pi}_{\|x\|_2}(\{c \geq 0 : c\tfrac{x}{\|x\|_2} \in A\})\mathrm{d}\mu(x),$$

and using the identity (22) further

$$\begin{aligned}\pi_{2\#}\pi(A) &= \int_{[0,\infty)} \int_{\partial B_r(0)} \tilde{\pi}_r(\{c \geq 0 : cx/r \in A\}) \mathrm{d}U_{\partial B_r(0)}(x) \mathrm{d}\tilde{\mu}(r)\\ &= \int_{[0,\infty)} \int_{\partial B_1(0)} \tilde{\pi}_r(\{c \geq 0 : cx \in A\}) \mathrm{d}U_{\partial B_1(0)}(x) \mathrm{d}\tilde{\mu}(r).\end{aligned}$$

Now, by definition of A, it holds that $\{c \geq 0 : cx \in A\} = [a,b]$ for $x \in B$ and $\{c \geq 0 : cx \in A\} = \emptyset$ for $x \notin B$. Thus, the above formula is equal to

$$\begin{aligned}\pi_{2\#}\pi(A) &= \int_{[0,\infty)} \int_{\partial B_1(0)} \tilde{\pi}_r([a,b]) 1_B(x) \,\mathrm{d}\mathcal{U}_{\partial B_1(0)}(x) \,\mathrm{d}\tilde{\mu}(r)\\ &= \int_{[0,\infty)} \tilde{\pi}_r([a,b]) \mathcal{U}_{\partial B_1(0)}(B) \,\mathrm{d}\tilde{\mu}(r) = \mathcal{U}_{\partial B_1(0)}(B) \int_{\mathbb{R}^d} \tilde{\pi}_r([a,b]) \,\mathrm{d}\tilde{\mu}(r)\\ &= \mathcal{U}_{\partial B_1(0)}(B) \pi_{2\#}\tilde{\pi}([a,b]) = \mathcal{U}_{\partial B_1(0)}(B) \tilde{\nu}([a,b])\end{aligned}$$

and applying (22) for ν to

$$\begin{aligned}\pi_{2\#}\pi(A) = \mathcal{U}_{\partial B_1(0)}(B)\tilde{\nu}([a,b]) &= \int_{[0,\infty)} \int_{\partial B_r(0)} 1_{[a,b]}(r) 1_B(x/r) \,\mathrm{d}\mathcal{U}_{\partial B_r(0)}(x) \,\mathrm{d}\tilde{\nu}(r)\\ &= \int_{[0,\infty)} \int_{\partial B_r(0)} 1_A(x) \mathrm{d}U_{\partial B_r(0)}(x) \mathrm{d}\tilde{\nu}(r) = \nu(A).\end{aligned}$$

Finally, note that for π-almost every (x,y) there exists by construction some $c \geq 0$ such that $x = cy$, which implies $\|x-y\|_2 = |\|x\|_2 - \|y\|_2|$. Further, it holds by construction that $(\|\cdot\|_2)_\#\pi_x = \tilde{\pi}_{\|x\|_2}$. Therefore, we can conclude that

$$\begin{aligned}W_2^2(\mu,\nu) &\leq \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 \,\mathrm{d}\pi(x,y) = \int_{\mathbb{R}^d \times \mathbb{R}^d} (\|x\|_2 - \|y\|_2)^2 \,\mathrm{d}\pi(x,y)\\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (\|x\|_2 - \|y\|_2)^2 \mathrm{d}\pi_x(y) \,\mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (x-y)^2 \,\mathrm{d}\tilde{\pi}_x(y) \,\mathrm{d}\tilde{\mu}(x)\\ &= \int_{\mathbb{R}^d \times \mathbb{R}^d} (x-y)^2 \,\mathrm{d}\tilde{\pi}(x,y) = W_2^2(\tilde{\mu},\tilde{\nu})\end{aligned}$$

and we are done.

3. The statement that Ψ is a bijective isometry follows directly by the previous part and Proposition C.2. □

Applying the isometry Ψ from the proposition, we can compute the steps from the JKO scheme (9) explicitly for the function $\mathcal{F} := \mathcal{E}_K$ starting at δ_0. We need the following lemma.

Lemma C.4. *For $r \in (0,2)$ and $\tau > 0$, we consider the functions*

$$h_\tau \colon \mathbb{R}_{>0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}, \quad h_\tau(t,s) := s^{\frac{1}{2-r}} t^{\frac{1-r}{2-r}} - t + \tau. \tag{23}$$

Then, for any $s \geq 0$, the function $t \mapsto h_\tau(t,s)$ has a unique positive zero $\hat{t}$ and it holds

$$h_\tau(t,s) > 0 \text{ for } t < \hat{t} \qquad \text{and} \qquad h_\tau(t,s) < 0 \text{ for } t > \hat{t}.$$

In particular, $h_\tau(t,s) \geq 0$ implies $t \leq \hat{t}$ and $h_\tau(t,s) \leq 0$ implies $t \geq \hat{t}$.

Proof. For $r \in [1,2)$, it holds $\frac{1-r}{2-r} \leq 0$ so that the first summand of $h_\tau(\cdot,s)$ is decreasing. Since the second one is also strictly decreasing, we obtain that h_τ is strictly decreasing. Moreover, we have by definition that $h(\tau,s) = s^{1/(2-r)}\tau^{(1-r)/(2-r)} \geq 0$ and that $h(t,s) \to -\infty$ as $t \to \infty$ such that it admits a unique zero $\hat{t} \geq \tau$.
For $r \in (0,1)$, we have $h_\tau(0,s) = \tau > 0$ and $h_\tau(t,s) \to -\infty$ as $t \to \infty$. This ensures the existence of a positive zero. Moreover, we have that $h_\tau(\cdot,s)$ is concave on $(0,\infty)$. Assume that there are $0 < \hat{t}_1 < \hat{t}_2$ with $h_\tau(\hat{t}_1,s) = h_\tau(\hat{t}_2,s) = 0$. Then, it holds by concavity that

$$h_\tau(\hat{t}_1,s) \geq (1 - \hat{t}_1/\hat{t}_2) h_\tau(0,s) + \hat{t}_1/\hat{t}_2 h_\tau(\hat{t}_2,s) = (1 - \hat{t}_1/\hat{t}_2)\tau > 0,$$

which is a contradiction. □

The following lemma implies Theorem 5.2(i).

Lemma C.5 (and **Theorem 5.2(i)**). *Let $\mathcal{F} = \mathcal{E}_K$, where K is the Riesz kernel with $r \in (0,2)$ and η^* be the unique element of $\operatorname{prox}_{\mathcal{F}}(\delta_0)$. Then, for $\mu = (t_0^{1/(2-r)} \operatorname{Id})_\# \eta^*$, $t_0 \geq 0$, there exists a unique measure $\hat{\mu} \in \operatorname{prox}_{\tau\mathcal{F}}(\mu)$ given by*

$$\hat{\mu} = \big(\hat{t}^{\frac{1}{2-r}} \operatorname{Id}\big)_\# \eta^*,$$

where $\hat{t}$ is the unique positive zero of the strictly decreasing function $t \mapsto h_\tau(t, t_0)$ in (23). In particular, this implies Theorem 5.2(i).

Proof. Set $\alpha := (t_0)^{\frac{1}{2-r}}$ so that $\mu = (\alpha \operatorname{Id})_\# \eta^*$. Since it holds by definition that $\eta^* \in \operatorname{prox}_{\mathcal{F}}(\delta_0)$, we have by Proposition C.1 that $\mu \in \operatorname{RS}(\mathbb{R}^d)$ and then $\operatorname{prox}_{\tau\mathcal{F}}(\mu) \subseteq \operatorname{RS}(\mathbb{R}^d)$. Using $f = \Psi(\eta^*)$ and the fact that $\Psi(c \operatorname{Id}_\# \nu) = c\Psi(\nu)$ for any $\nu \in \operatorname{RS}(\mathbb{R}^d)$ and $c \geq 0$, we obtain

$$\begin{aligned}\operatorname{prox}_{\tau\mathcal{F}}(\mu) &= \operatorname*{arg\,min}_{\nu \in \operatorname{RS}(\mathbb{R}^d)} \tfrac{1}{2\tau} W_2^2(\nu, \mu) + \mathcal{F}(\nu) \\ &= \Psi^{-1}\Big(\operatorname*{arg\,min}_{g \in \mathcal{C}_0((0,1))} \tfrac{1}{2\tau} \int_0^1 (g(s) - \alpha f(s))^2 \mathrm{d}s + \mathcal{F}(\Psi^{-1}(g))\Big). \end{aligned} \tag{24}$$

Then, we know by Theorem 5.1 that

$$\big\{\Psi^{-1}\big(\hat{t}^{\frac{1}{2-r}} f\big)\big\} = \big\{\big(\hat{t}^{\frac{1}{2-r}} \operatorname{Id}\big)_\# \eta^*\big\} = \operatorname{prox}_{\hat{t}\mathcal{F}}(\delta_0)$$

such that

$$\{\hat{t}^{\frac{1}{2-r}} f\} = \operatorname*{arg\,min}_{g \in \mathcal{C}_0((0,1))} \tfrac{1}{2\hat{t}} \int_0^1 (g(s))^2 \, \mathrm{d}s + \mathcal{F}(\Psi^{-1}(g)). \tag{25}$$

Now, we consider the optimization problem

$$\begin{aligned} &\operatorname*{arg\,min}_{g \in \mathcal{C}_0((0,1))} \tfrac{1}{2\tau} \int_0^1 (g(s) - \alpha f(s))^2 \, \mathrm{d}s - \tfrac{1}{2\hat{t}} \int_0^1 (g(s))^2 \, \mathrm{d}s \\ = &\operatorname*{arg\,min}_{g \in \mathcal{C}_0((0,1))} \int_0^1 \big(\tfrac{1}{2\tau} - \tfrac{1}{2\hat{t}}\big) g(s)^2 + \tfrac{\alpha}{2\tau} f(s)^2 - \tfrac{\alpha}{\tau} f(s) g(s) \, \mathrm{d}s. \end{aligned} \tag{26}$$

By Lemma C.4 we know that $\hat{t} > \tau$. Hence this problem is convex and any critical point is a global minimizer. By setting the derivative in $L^2((0,1))$ to zero, we obtain that the minimizer fulfills

$$0 = \tfrac{1}{\tau}(g(s) - \alpha f(s)) - \tfrac{1}{\hat{t}} g(s) \quad \Leftrightarrow \quad g(s) = \frac{\hat{t}\alpha}{\hat{t} - \tau} f(s).$$

Since

$$h_\tau(\hat{t}, t_0) = 0 \quad \Leftrightarrow \quad \frac{\hat{t}^{\frac{1-r}{2-r}} t_0^{\frac{1}{2-r}}}{\hat{t} - \tau} = 1 \quad \Leftrightarrow \quad \frac{\hat{t}\alpha}{\hat{t} - \tau} = \hat{t}^{\frac{1}{2-r}}$$

it follows that $\hat{t}^{\frac{1}{2-r}} f$ is the minimizer of (26). As it is also the unique minimizer in (25), we conclude by adding the two objective functions that

$$\big\{\hat{t}^{\frac{1}{2-r}} f\big\} \in \operatorname*{arg\,min}_{g \in \mathcal{C}_0((0,1))} \tfrac{1}{2\tau} \int_0^1 (g(s) - \alpha f(s))^2 \mathrm{d}s + \mathcal{F}(\Psi^{-1}(g)).$$

By (24), this implies that

$$\operatorname{prox}_{\tau\mathcal{F}}(\mu) = \big\{\Psi^{-1}\big(\hat{t}^{\frac{1}{2-r}} f\big)\big\} = \big\{\big(\hat{t}^{\frac{1}{2-r}} \operatorname{Id}\big)_\# \eta^*\big\}$$

and we are done. □

Finally, we have to invest some effort to show the convergence of the curves induced by the JKO scheme. We need two preliminary lemmata to prove finally Theorem 5.2(ii).

Lemma C.6. *Let $r \in (0,2)$, $t_{\tau,0} = 0$ and let $t_{\tau,n}$ be the unique positive zero of $h_\tau(\cdot, t_{\tau,n-1})$ in (23). Then, the following holds true:*

(i) *If $r \in [1,2)$, then $t_{\tau,n} \geq t_{\tau,n-1} + (2-r)\tau$, and thus $t_{\tau,n} \geq (2-r)n\tau$.*
If $r \in (0,1]$, then $t_{\tau,n} \leq t_{\tau,n-1} + (2-r)\tau$, and thus $t_{\tau,n} \leq (2-r)n\tau$.

(ii) *Let $n \geq 2$. For $r \in [1,2)$, we have*

$$t_{\tau,n} - t_{\tau,n-1} \leq (2-r)\tau + c_{\tau,n}\tau, \quad \textit{with} \quad c_{\tau,n} = \tfrac{r-1}{(4-2r)(n-1)} \geq 0. \tag{27}$$

For $r \in (0,1]$, the same inequality holds true with $\geq$ instead of $\leq$.

Proof. (i) For $r \in [1,2)$, the function $x \mapsto x^{\frac{1-r}{2-r}}$ is convex. Then the identity $f(y) \geq f(x) + (y-x)f'(x)$ for convex, differentiable functions yields

$$(t + (2-r)\tau)^{\frac{1-r}{2-r}} \geq t^{\frac{1-r}{2-r}} + (2-r)\tau \, \tfrac{1-r}{2-r} \, t^{\frac{1-r}{2-r}-1} = t^{\frac{1-r}{2-r}} + (1-r)\tau t^{-\frac{1}{2-r}}.$$

Hence, we obtain

$$\begin{aligned} h_\tau(t + (2-r)\tau, t) &= t^{\frac{1}{2-r}} (t + (2-r)\tau)^{\frac{1-r}{2-r}} - t - (2-r)\tau + \tau \\ &\geq t^{\frac{1}{2-r}} \big(t^{\frac{1-r}{2-r}} + (1-r)\tau t^{-\frac{1}{2-r}}\big) - t + (r-1)\tau = 0. \end{aligned}$$

In particular, we have that $h_\tau(t_{\tau,n-1} + (2-r)\tau, t_{\tau,n-1}) \geq 0$ which implies the assertion by Lemma C.4. The proof for $r \in (0,1]$ works analogously by using the concavity of $x \mapsto x^{\frac{1-r}{2-r}}$.

(ii) Let $r \in [1,2)$. Using Taylor's theorem, we obtain with $\xi \in [t, t + (2-r)\tau]$ that

$$\begin{aligned} (t + (2-r)\tau)^{\frac{1-r}{2-r}} &= t^{\frac{1-r}{2-r}} + (2-r)\tau \, \tfrac{1-r}{2-r} \, t^{-\frac{1}{2-r}} + \tfrac{r-1}{2(2-r)^2} \xi^{\frac{-1}{2-r}-1} (2-r)^2 \tau^2 \\ &= t^{\frac{1-r}{2-r}} + (1-r)\tau t^{-\frac{1}{2-r}} + \tfrac{r-1}{2} \xi^{-\frac{1}{2-r}-1} \tau^2 \\ &\leq t^{\frac{1-r}{2-r}} + (1-r)\tau t^{-\frac{1}{2-r}} + \tfrac{r-1}{2} t^{-\frac{1}{2-r}-1} \tau^2. \end{aligned}$$

Thus, by monotonicity of $x \mapsto x^{\frac{1-r}{2-r}}$ it holds for $t > 0$ and $c = \frac{r-1}{2t}$ that

$$\begin{aligned} & h_\tau\left(t + (2-r)\tau + c\tau^2, t\right) = t^{\frac{1}{2-r}} \left(t + (2-r)\tau + c\tau^2\right)^{\frac{1-r}{2-r}} - t - (2-r)\tau - c\tau^2 + \tau \\ & \leq t^{\frac{1}{2-r}} \left(t + (2-r)\tau\right)^{\frac{1-r}{2-r}} - t + (r-1)\tau - c\tau^2 \\ & \leq t^{\frac{1}{2-r}} \Big(t^{\frac{1-r}{2-r}} + (1-r)\tau t^{-\frac{1}{2-r}} + \tfrac{r-1}{2} t^{-\frac{1}{2-r}-1} \tau^2\Big) - t + (r-1)\tau - c\tau^2 = 0. \end{aligned}$$

Inserting $t_{\tau,n-1} \geq (2-r)(n-1)\tau > 0$ for t and setting $c := \frac{r-1}{2t_{\tau,n}}$, we obtain

$$h(t_{\tau,n-1} + (2-r)\tau + c\tau^2, t_{\tau,n-1}) \leq 0,$$

which yields by Lemma C.4 that

$$t_{\tau,n} \leq t_{\tau,n-1} + (2-r)\tau + c\tau^2 \quad \text{and} \quad c = \tfrac{r-1}{2t_{\tau,n}} \leq \tfrac{r-1}{(4-2r)(n-1)\tau} = \tfrac{c_{\tau,n}}{\tau}.$$

and consequently the assertion

$$t_{\tau,n} \leq t_{\tau,n-1} + (2-r)\tau + c_{\tau,n}\tau.$$

The proof for $r \in (0,1]$ follows the same lines. □

Lemma C.7. *Let $r \in (0,2)$, $t_{\tau,0} = 0$, and let $t_{\tau,n}$ be the unique positive zero of $h_\tau(\cdot, t_{\tau,n-1})$ defined in (23). Then, it holds for $r \in [1,2)$ that*

$$0 \leq t_{\tau,n} - (2-r)\tau n \leq \tau(r-1)\Big(1 + \tfrac{1}{4-2r} + \tfrac{1}{4-2r}\log(n)\Big),$$

and for $r \in (0,1]$ that

$$0 \leq (2-r)\tau n - t_{\tau,n} \leq \tau(r-1)\Big(1 + \tfrac{1}{4-2r} + \tfrac{1}{4-2r}\log(n)\Big).$$

Proof. In both cases, the first inequality was proven in Lemma C.6 (i). For the second one, we consider the case $r \in [1,2)$. For $r \in (0,1)$, the assertion follows in the same way. Since $h_\tau(\tau,0)=0$, we have that $t_{\tau,1}=\tau$ such that $t_{\tau,1}-(2-r)\tau=(r-1)\tau$. This proves the estimate for $n=1$. Moreover, summing up the equations (27) for $2,...,n$, we obtain

$$\begin{aligned} t_{\tau,n} &\le (2-r)\tau n+(r-1)\tau+\tau\sum_{k=2}^{n}\tfrac{r-1}{(4-2r)(k-1)} \\ &= (2-r)\tau n+\tau(r-1)\Big(1+\tfrac{1}{4-2r}\sum_{k=1}^{n-1}\tfrac{1}{k}\Big) \\ &\le (2-r)\tau n+\tau(r-1)\Big(1+\tfrac{1}{4-2r}+\tfrac{1}{4-2r}\log(n)\Big) \end{aligned}$$

and we are done. □

Proof of Theorem 5.2(ii) For fixed $T>0$, we show that γ_τ converges uniformly on $[0,T]$ to γ. Then, for $n=0,1,\dots$, we have by part (i) of the theorem that

$$\gamma_\tau(t)=\mu_\tau^n=\big((t_{\tau,n})^{\frac{1}{2-r}}\operatorname{Id}\big)_{\#}\eta^*,\quad t\in((n-1)\tau,n\tau]$$

and we want to show convergence to

$$\gamma(t)=\big((t(2-r))^{\frac{1}{2-r}}\operatorname{Id}\big)_{\#}\eta^*.$$

Since the curve $t\mapsto(t\operatorname{Id})_{\#}\eta^*$ is a geodesics, there exists a constant $C>0$ such that

$$W_2(\gamma_\tau(t),\gamma(t))\le C|(t_{\tau,n})^{\frac{1}{2-r}}-(t(2-r))^{\frac{1}{2-r}}|. \tag{28}$$

Now assume that $n\tau\le T$, i.e., $n\le T/\tau$.

For $r\in[1,2)$, the function $t\mapsto t^{\frac{1}{2-r}}$ is Lipschitz continuous on $[0,T]$, such that there exists some $L>0$ such that for $t\in((n-1)\tau,n\tau]$,

$$\begin{aligned} W_2(\gamma_\tau(t),\gamma(t)) &\le LC|t_{\tau,n}-(2-r)t|\le LC\left(|t_{\tau,n}-(2-r)n\tau|+(2-r)|t-n\tau|\right) \\ &\le LC\left(|t_{\tau,n}-(2-r)n\tau|+(2-r)\tau\right), \end{aligned}$$

and by Lemma C.7 further

$$W_2(\gamma_\tau(t),\gamma(t))\le LC\tau(r-1)\Big(1+\tfrac{1}{4-2r}+\tfrac{1}{4-2r}\log(\tfrac{T}{\tau})\Big)+LC(2-r)\tau\to 0\quad\text{as}\quad\tau\to 0$$

which yields the assertion for $r\in[1,2)$.

For $r\in(0,1]$, the function defined by $f(t):=t^{\frac{1}{2-r}}$ is increasing and f' is decreasing for $t>0$. Thus, using $t_{\tau,n}\le(2-r)n\tau$, we get for $t\in[(n-1)\tau,n\tau)$ that

$$\begin{aligned} t_{\tau,n}^{\frac{1}{2-r}}-(t(2-r))^{\frac{1}{2-r}} &\le((2-r)n\tau)^{\frac{1}{2-r}}-((2-r)(n-1)\tau)^{\frac{1}{2-r}} \\ &=\int_{(2-r)(n-1)\tau}^{(2-r)n\tau}f'(t)\,\mathrm{d}t\le\int_0^{(2-r)\tau}f'(t)\,\mathrm{d}t \\ &=((2-r)\tau)^{1/(2-r)}. \end{aligned}\tag{29}$$

Similarly, we obtain

$$\begin{aligned} (t(2-r))^{\frac{1}{2-r}}-t_{\tau,n}^{\frac{1}{2-r}} &\le((2-r)n\tau)^{\frac{1}{2-r}}-t_{\tau,n}^{\frac{1}{2-r}} \\ &=\int_{t_{\tau,n}}^{(2-r)n\tau}f'(t)\,\mathrm{d}t\le\int_0^{(2-r)n\tau-t_{\tau,n}}f'(t)\,\mathrm{d}t \\ &=\big((2-r)n\tau-t_{\tau,n}\big)^{\frac{1}{2-r}} \\ &\le\Big(\tau(r-1)\Big(1+\tfrac{1}{4-2r}+\tfrac{1}{4-2r}\log(\tfrac{t}{\tau})\Big)\Big)^{\frac{1}{2-r}}. \end{aligned}\tag{30}$$

Combining (28), (29) and (30), we obtain the assertion. □

Proof of Theorem 5.4 For $n = 0$ the claim holds true by definition. For $n \geq 1$, assume that

$$\mu_\tau^{n-1} = ((n-1)\tau)_\# \eta^*$$

and consider the geodesics

$$\gamma_{\delta_0 \otimes \eta^*}(t) = (t\,\mathrm{Id})_\# \eta_1^* = (t\,\mathrm{Id})_\# \eta^*.$$

Note that by Corollary 20 and Theorem 22 in (Hertrich et al., 2022) there exists a unique steepest descent direction $\mathrm{D}_-\mathcal{F}((t\,\mathrm{Id})_\#\eta^*)$ for all $t \geq 0$. Moreover, we have by Theorem 23 in (Hertrich et al., 2022) that $\gamma_{\delta_0\otimes\eta^*}(t)$ is a Wasserstein steepest descent flow. Thus, using Lemma 6 in (Hertrich et al., 2022), we obtain that the unique element $\boldsymbol{v} \in \mathrm{D}_-\mathcal{F}(\mu_\tau^{n-1}) = \mathrm{D}_-\mathcal{F}(\gamma_{\delta_0\otimes\eta^*}((n-1)\tau))$ is given by

$$\boldsymbol{v} = \dot\gamma_{\delta_0\otimes\eta^*}((n-1)\tau) = (\pi_1 + (n-1)\tau\pi_2, \pi_2)_\#(\delta_0 \otimes \eta^*) = ((n-1)\tau\,\mathrm{Id}, \mathrm{Id})_\#\eta^*.$$

In particular, we have that

$$\mu_\tau^n = \gamma_{\boldsymbol{v}}(\tau) = (\pi_1 + \tau\pi_2)_\#\boldsymbol{v} = (\pi_1 + \tau\pi_2)_\#((n-1)\tau\,\mathrm{Id}, \mathrm{Id})_\#\eta^* = (n\tau\,\mathrm{Id})_\#\eta^*.$$

Now, the claim follows by induction. □

D. Particle Flows for Numerical Comparison

In order to approximate the Wasserstein gradient flow by particles, we restrict the set of feasible measures to the set of point measures located at exactly M points, i.e., to the set

$$\mathcal{S}_M := \Big\{\frac{1}{M}\sum_{i=1}^M \delta_{x_i} : x_i \in \mathbb{R}^d, x_i \neq x_j \text{ for all } i \neq j\Big\}.$$

Then, we compute the Wasserstein gradient flow of the functional

$$\mathcal{F}_M(\mu) := \begin{cases} \mathcal{F}(\mu), & \text{if } \mu \in \mathcal{S}_M, \\ +\infty, & \text{otherwise.}\end{cases}$$

In order to compute the gradient flow with respect ot $\mathcal{F}_M$, we consider the (rescaled) particle flow for the function $F_M\colon \mathbb{R}^{dM} \to \mathbb{R}$ given by

$$F_M(x_1, ..., x_M) := \mathcal{F}_M\Big(\frac{1}{M}\sum_{i=1}^M \delta_{x_i}\Big).$$

More precisely, we are interested in solutions of the ODE

$$\dot u = -M\nabla F_M(u). \tag{31}$$

Then, the following proposition relates the solutions of (31) with Wasserstein gradient flows with respect to $\mathcal{F}_M$.

Proposition D.1. *Let $u = (u_1, ..., u_M)\colon (0,\infty) \to \mathbb{R}^{dM}$ be a solution of (31) with $u_i(t) \neq u_j(t)$ for all $i \neq j$ and all $t \in (0,\infty)$. Then, the curve*

$$\gamma\colon (0,\infty) \to \mathcal{P}_2(\mathbb{R}^d), \quad \gamma(t) := \frac{1}{M}\sum_{i=1}^M \delta_{u_i(t)}$$

is a Wasserstein gradient flow with respect to $\mathcal{F}_M$.

Proof. Let $x = (x_1, ..., x_M) \in \mathbb{R}^{dM}$ with $x_i \neq x_j$ for all $i \neq j$. Then, there exists $\epsilon > 0$ such that for all $y \in \mathbb{R}^{dM}$ with $\|x - y\| < \epsilon$ it holds that the optimal transport plan between $\frac{1}{M}\sum_{i=1}^M \delta_{x_i}$ and $\frac{1}{M}\sum_{i=1}^M \delta_{y_i}$ is given by $\boldsymbol{\pi} := \frac{1}{M}\sum_{i=1}^M \delta_{(x_i,y_i)}$. In particular, it holds

$$W_2^2\Big(\frac{1}{M}\sum_{i=1}^M \delta_{x_i}, \frac{1}{M}\sum_{i=1}^M \delta_{y_i}\Big) = \frac{1}{M}\sum_{i=1}^M \|x_i - y_i\|_2^2. \tag{32}$$

Moreover, since F_M is locally Lipschitz continuous, we obtain that u is absolute continuous. Together with (32), this yields that γ is (locally) absolute continuous. Thus, we obtain by Proposition 8.4.6 in (Ambrosio et al., 2005) that the velocity field v_t of γ fulfills

$$\begin{aligned}
0 &= \lim_{h\to 0} \frac{W_2^2(\gamma(t+h)), (Id + hv_t)_{\#}\gamma(t))}{|h|^2} \\
&= \lim_{h\to 0} \frac{W_2^2(\frac{1}{M}\sum_{i=1}^M \delta_{u_i(t+h)}, \frac{1}{M}\sum_{i=1}^M \delta_{u_i(t)+hv_t(u_i(t))})}{|h|^2} \\
&= \lim_{h\to 0} \frac{1}{M}\sum_{i=1}^M \Big\| \frac{u_i(t+h)-u_i(t)}{h} - v_t(u_i(t)) \Big\|^2 = \frac{1}{M}\sum_{i=1}^M \|\dot u_i(t) - v_t(u_i(t))\|^2
\end{aligned}$$

for almost every $t \in (0,\infty)$, where the first equality in the last line follows from (32). In particular, this implies $\dot u_i(t) = v_t(u_i(t))$ a.e. such that $M\nabla F_M(u(t)) = \big(v_t(u_1(t)), ..., v_t(u_M(t))\big)$. Now consider for fixed t and ϵ from (32) some $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $W_2(\mu, \gamma(t)) < M\epsilon$. If $\mu \in \mathcal{S}_M$, we find $x_\mu = (x_{\mu,1}, ..., x_{\mu,M}) \in \mathbb{R}^{dM}$ such that $\mu = \frac{1}{M}\sum_{i=1}^M \delta_{x_{\mu,i}}$ and such that the unique element of $\Gamma^{\mathrm{opt}}(\mu, \gamma(t))$ is given by $\boldsymbol{\pi} = \frac{1}{M}\sum_{i=1}^M \delta_{(x_{\mu,i}, u_i(t))}$. Then, we obtain that

$$\begin{aligned}
0 &\le F_M(x_\mu) - F_M(u(t)) + \langle \nabla F_M(u(t)), x_\mu - u(t)\rangle + o(\|x_\mu - u(t)\|) \\
&= \mathcal{F}_M(\mu) - \mathcal{F}_M(\gamma(t)) + \frac{1}{M}\sum_{i=1}^M \langle v_t(u_i(t)), x_{\mu,i} - u_i(t)\rangle + o(W_2(\mu,\gamma(t))) \\
&= \mathcal{F}_M(\mu) - \mathcal{F}_M(\gamma(t)) + \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle v_t(x_1), x_2 - x_1\rangle \mathrm{d}\boldsymbol{\pi}(x_1,x_2) + o(W_2(\mu,\gamma(t))).
\end{aligned}$$

Since $\boldsymbol{\pi}$ is the unique optimal transport plan between μ and $\gamma(t)$, we obtain that for $\mu \in \mathcal{S}_M$ equation (18) is fulfilled. If $\mu \notin \mathcal{S}_M$, we obtain that $\mathcal{F}_M(\mu) = +\infty$ such that (18) holds trivially true. Summarizing, this yields that $v_t \in -\partial\mathcal{F}_M(\gamma(t))$ showing the assertion by (5). □

E. Implementation details

Our code is written in PyTorch (Paszke et al., 2019). For the network-based methods neural backward scheme and neural forward scheme we use the same fully connected network architecture with ReLU activation functions and train the networks with Adam optimizer (Kingma & Ba, 2015) with a learning rate of $1e-3$.

In Sect. 6.1 we use a batch size of 6000 in two and three dimensions, of 5000 in ten dimensions, of 500 in 1000 dimensions and a time step size of $\tau = 0.05$ for all methods. In two, three and ten dimensions we use networks with three hidden layer and 128 nodes for both methods and in 1000 dimensions for the neural backward scheme three hidden layers with 256 nodes, while we use 2048 nodes for the neural forward scheme. We train the neural forward scheme for 25000 iterations in all dimensions and the neural backward scheme for 20000 iterations using a learning rate of $5e-4$.

In Sect. 6.2 we use a full batch size for 5000 iterations in the first two time steps and then 1000 iterations for the neural forward scheme and for the neural backward scheme 20000 and 10000 iterations, respectively. The networks have four hidden layers and 128 nodes and we use a time step size of $\tau = 0.5$. In order to illustrate the given image in form of samples, we use the code of (Wu et al., 2020). For the 784-dimensional MNIST task we use two hidden layers and 2048 nodes and a step size of $\tau = 0.5$ for the first 10 steps. Then we increase the step size to 1, 2 and 3 after a time of 5, 25 and 50, respectively, and finally use a step size of $\tau = 6$ after a time of 6000. While the starting measure of the network-based methods can be chosen as δ_x for $x = 0.5 \cdot \mathbf{1}_{784}$ and $\mathbf{1}_d \in \mathbb{R}^d$ is the vector with all entries equal to 1, the initial particles of the particle flow are sampled uniformly around x with a radius of $R = 10^{-9}$.

Neural forward scheme vs neural backward scheme The advantages of forward and backward scheme mainly follow the case of forward and backward schemes in Euclidean spaces. In our experiments the forward scheme performs better. Moreover, the notion of the steepest descent direction as some kind of derivative might allow some better analysis. In particular, it could be used for the development of "Wasserstein momentum methods" which appears to be an interesting direction of further research. On the other hand, the forward scheme always requires the existence of steepest descent directions. This is much stronger than the assumption that the Wasserstein proxy (9) exists, which is the main assumption for the backward scheme. For instance, in the case $r < 1$ the backward scheme exists, but not the forward scheme. Therefore the backward scheme can be applied for more general functions.

F. Initial Particle Configurations for Particle Flows

Figs. 6 and 7 illustrate the effect of using different initial particles of the particle flow for the interaction energy $\mathcal{F} = \mathcal{E}_K$. While in Fig. 6 we use the Riesz kernel with 2-norm, in Fig. 7 the 1-norm is used for the Riesz kernel. Since for the particle flow we cannot start in an atomic measure, we need to choose the initial particles appropriately, depending on the choice of the kernel for the MMD functional. For the kernel $K(x, y) = -\|x - y\|_2$, the 'best' initial structure is a circle, while a square is the 'best' structure for the kernel $K(x, y) = -\|x - y\|_1$ since it decouples to a sum of 1-dimensional functions. Obviously, the geometry of the initial particles influences the behaviour of the particle flow heavily and leads to severe approximation errors if the time step size is not sufficient small. In particular, for the used time step size of $\tau = 0.05$ the geometry of the initial particles is retained. Note that, since the optimal initial structure depends on the choice of the functional $\mathcal{F}$ and its computation is non-trivial, we decided to choose the non-optimal square initial structure for all experiments.

G. MMD Flows on the line

While in general an analytic solution of the MMD flow is not known, in the one-dimensional case we can compute the exact gradient flow if the target measure is $\nu = \delta_p$ for some $p \in \mathbb{R}$. More explicitly, by (Hertrich et al., 2023a) the exact gradient flow of $\mathcal{F}_{\delta_0}$ starting in the initial measure $\mu_\tau^0 = \delta_{-1}$ is given by

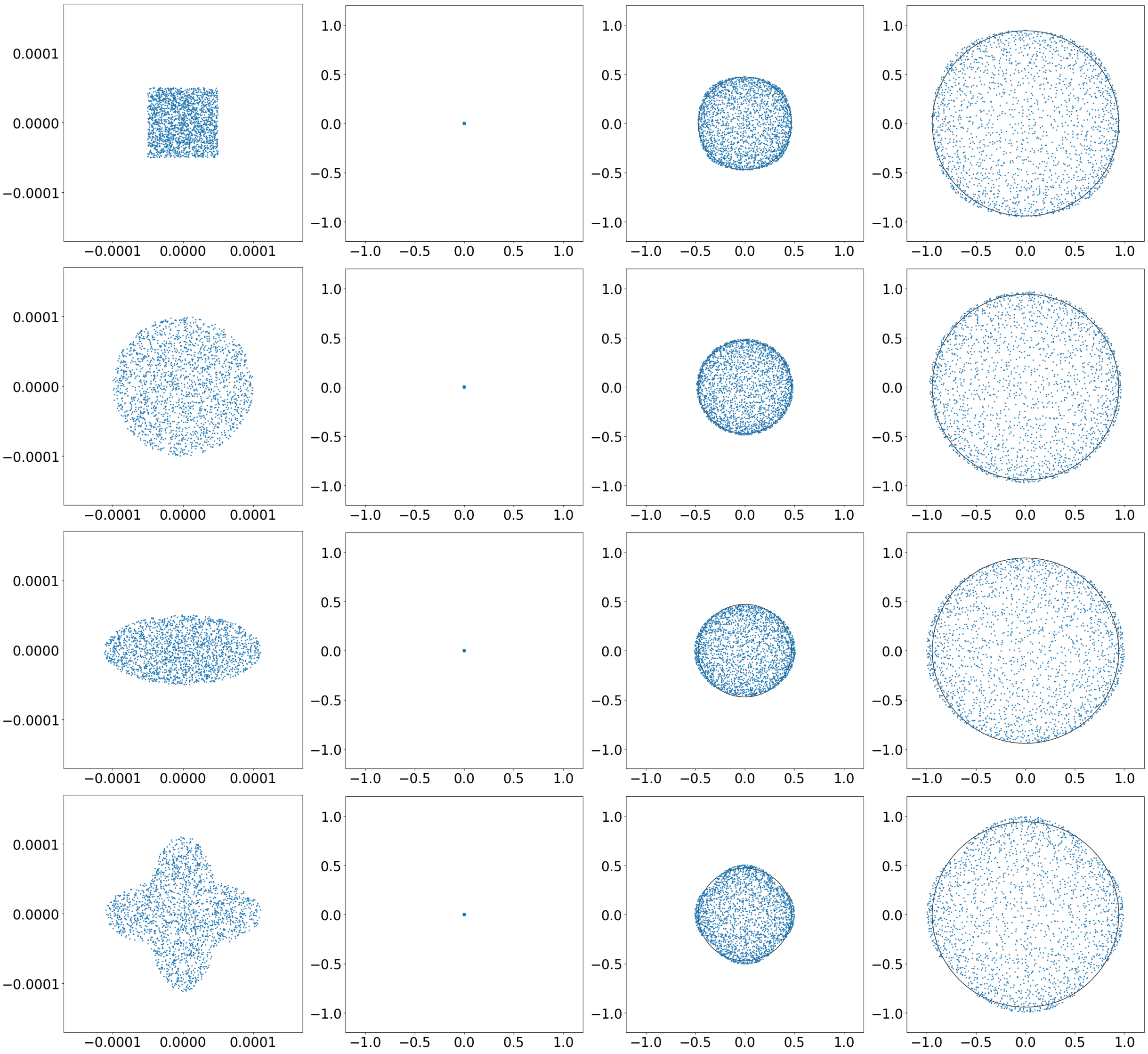
$$\gamma(t) = \begin{cases} \delta_{-1}, & t = 0, \\ \frac{1}{2t}\lambda_{[-1,-1+2t]}, & 0 \le t \le \frac{1}{2}, \\ \frac{1}{2t}\lambda_{[-1,0]} + (1 - \frac{1}{2t})\delta_0, & \frac{1}{2} < t. \end{cases}$$

A quantitative comparison between the analytic flow and its approximations with the discrepancy $\mathcal{D}_K^2$ is given in Fig. 8. We use a time step size of $\tau = 0.01$ and simulated 2000 samples. While our neural schemes start in δ_{-1}, the particle flow starts with uniformly distributed particles in an interval of size 10^{-9} around -1. Until the time $t = 0.5$ all methods behave similarly, and then the neural forward scheme and the particle flow give a much worse approximation than the neural backward scheme. This can be explained by the fact that after time $t = 0.5$ the particles should flow into the singular target δ_0. Nevertheless, the repulsion term in the discrepancy leads to particle explosions in the neural forward scheme and the particle flow such that the approximation error increases. This behaviour can be prevented by decreasing the time step size τ.

H. Further Numerical Examples

Example 1 In Fig. 9, we consider the target measure given from the image 'Smiley'. A sample from the exact target density is illustrated in Fig. 9 (right). For all methods we use a time step size of $\tau = 0.1$. The network-based methods use a network with four hidden layers and 128 nodes and train for 4000 iterations in the first ten steps and then for 2000 iterations. While we can start the network-based methods in $\delta_{(-1,0)} + \delta_{(1,0)}$, the 2000 initial particles of the particle flow needs to be placed in small squares of radius $R = 10^{-9}$ around $\delta_{(-1,0)}$ and $\delta_{(1,0)}$. The effect of this remedy can be seen in Fig. 9 (bottom), where the particles tend to form squares.

Example 2 In Fig. 10, we aim to compute the MMD flows for the target $\nu = \delta_{(-1,-1)} + \delta_{(1,1)}$ starting in $\delta_{(-1,1)} + \delta_{(1,-1)}$ for the network-based methods and small squares with radius $R = 10^{-9}$ around $(-1, 1)$ and $(1, -1)$ for the particle flow. We use a step size of $\tau = 0.1$. The network-based methods use a network with four hidden layers and 128 nodes and train for 4000 iterations in the first ten steps and then for 2000 iterations.



t=0.0 t=0.0 t=0.6 t=1.2

Figure 6: Effect of different initial particle configurations for the particle flow of $\mathcal{E}_K$, $K(x,y) := -\|x-y\|_2$. The black circle is the border of the limit $\operatorname{supp} \gamma(t)$. Using Gaussian distributed samples instead of uniformly distributed samples lead to a similar result. Left: zoomed-in part of the initial particles (note axes!).

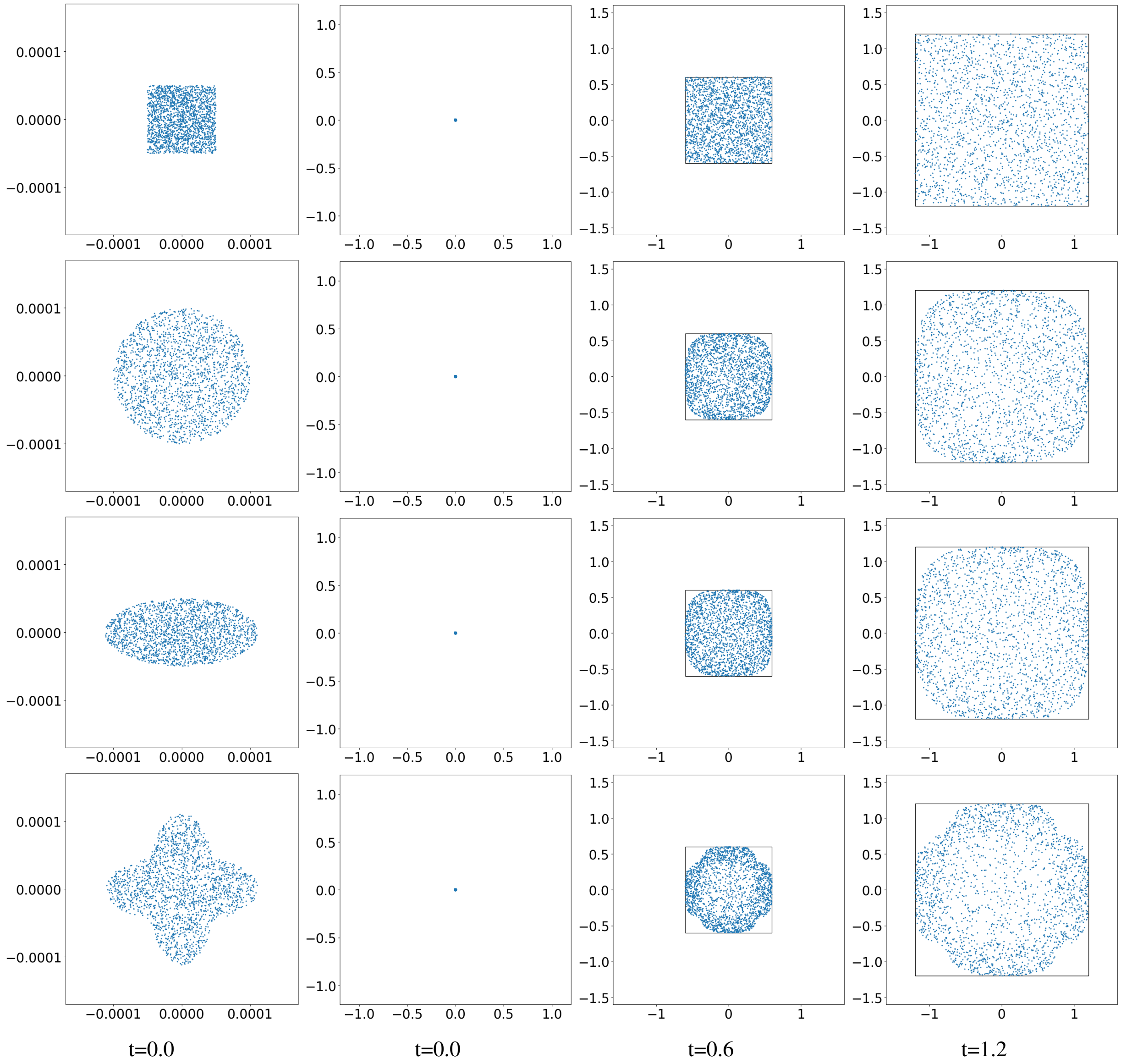


Figure 7: Effect of different initial particle configurations for the particle flow of $\mathcal{E}_K$, $K(x, y) := -\|x - y\|_1$. The black circle is the border of the limit $\operatorname{supp} \gamma(t)$. Left: zoomed-in part of the initial particles (note axes!).

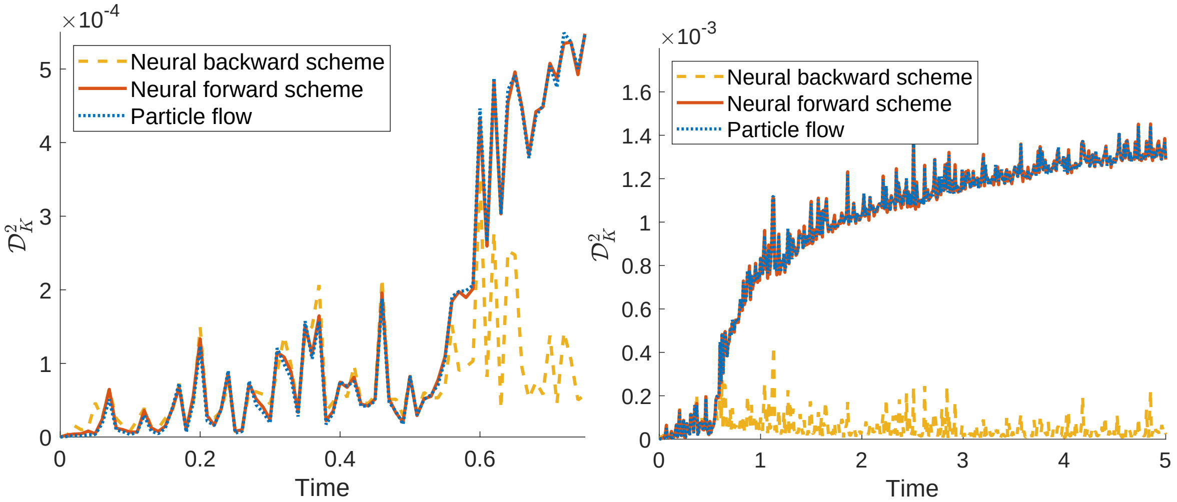


Figure 8: Discrepancy between the analytic Wasserstein gradient flow of $\mathcal{F}_{\delta_0}$ and its approximations for $\tau = 0.01$. **Left:** The discrepancy until time $t = 0.75$. Here we can observe that all methods behave similarly until time $t = 0.5$. After that, the particles of the analytic solution flow into δ_0. **Right:** The discrepancy until time $t = 5$. Obviously, the neural backward scheme is able to approximate the Wasserstein flow well, while the repulsion term leads to an explosion of the particles and therefore to a high approximation error for the neural forward scheme and the particle flow.

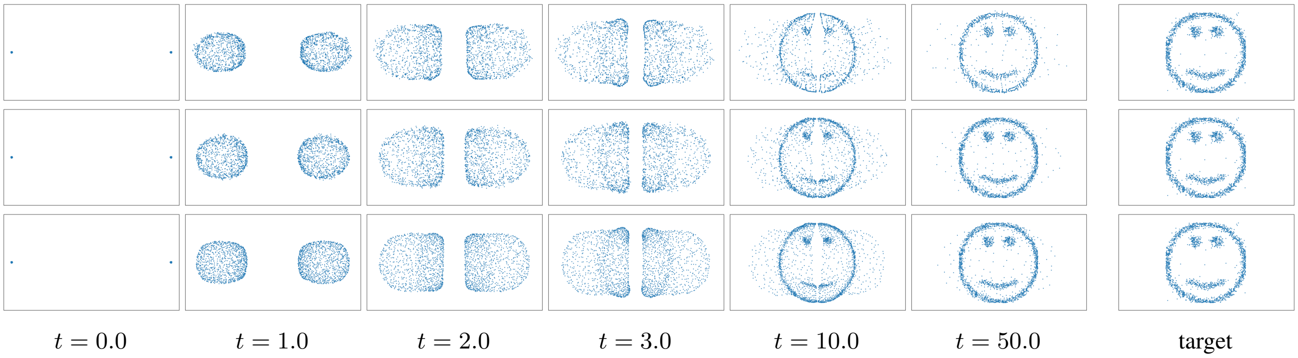


Figure 9: Comparison of neural backward scheme (top), neural forward scheme (middle) and particle flow (bottom) for sampling of the two-dimensional density 'Smiley' starting in $\delta_{(-1,0)} + \delta_{(1,0)}$.

Example 3 Instead of computing a MMD flow, we can also consider a different functional $\mathcal{F}$. Here we define the energy functional as

$$\begin{aligned}\mathcal{F}(\mu) = &\int_{\mathbb{R}^2} 1_{(-\infty,0)}(x)|y| - x \,\mathrm{d}\mu(x,y)\\ &- \frac{1}{2}\int_{\mathbb{R}^2}\int_{\mathbb{R}^2} 1_{[0,\infty)^2}(x_1,x_2)\|y_1 - y_2\|\mathrm{d}\mu(x_1,y_1)\mathrm{d}\mu(x_2,y_2).\end{aligned} \tag{33}$$

The first term in the first integral pushes the particles towards the x-axis until $x = 0$ and the second term in the first integral moves the particles to the right. The second integral is the interaction energy in the y-dimension, pushing the particles away from the x-axis. The corresponding neural backward scheme and neural forward scheme are depiced in Fig. 11. Initial particles are sampled from from the uniform distribution on $[-2,-1] \times [-0.5, 0.5]$, i.e., the initial measure is absolutely continuous.

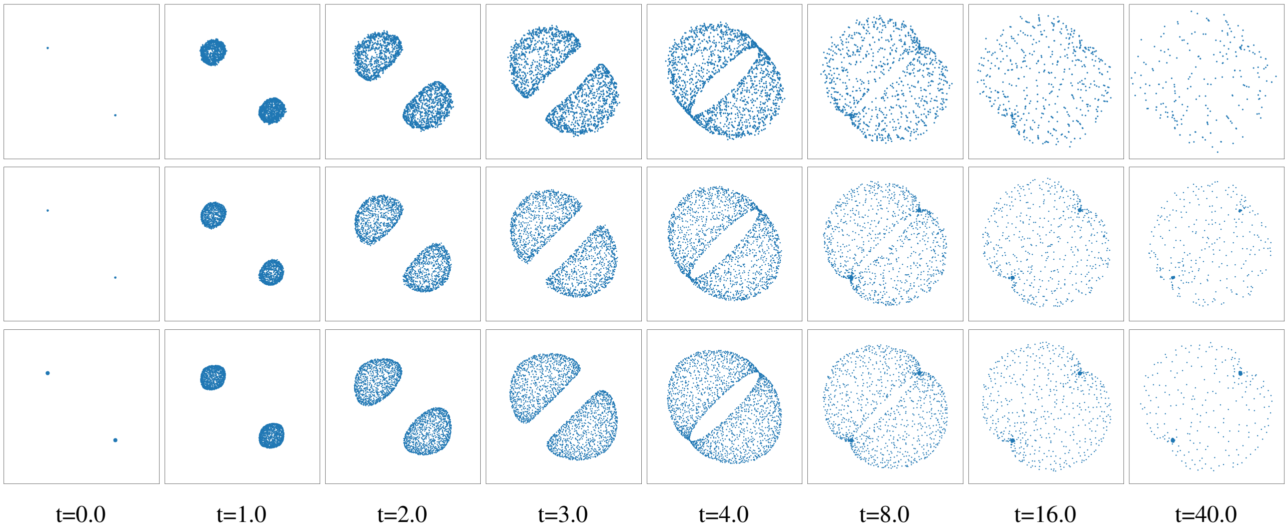


Figure 10: Comparison of neural backward scheme (top), neural forward scheme (middle) and particle flow (bottom) for computing the MMD flow with target $\nu = \delta_{(1,1)} + \delta_{(-1,-1)}$ starting in the opposite diagonal points $\delta_{(-1,1)} + \delta_{(1,-1)}$

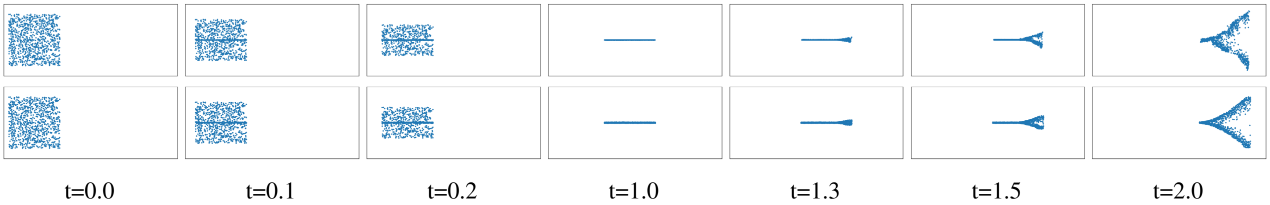


Figure 11: Neural backward scheme (top) and neural forward scheme (bottom) for the energy functional (33) starting at the uniform distribution on the square. The absolutely continuous initial measure becomes for $t > 1$ a *non* absolutely continuous one, which is supported on a line at $t = 1$ and branches afterwards.

Example 4 We can use the proposed schemes for computing the MMD barycenter. More precisely, let $\mu_1, ..., \mu_n \in \mathcal{P}_2(\mathbb{R}^d)$, then we aim to find the measure μ^* given by

$$\mu^* = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^n \alpha_i \mathcal{D}_K^2(\mu, \mu_i), \quad \sum_{i=1}^n \alpha_i = 1.$$

Consequently, we consider the functional $\mathcal{F}$ given by

$$\mathcal{F}_{\mu_1,...,\mu_n}(\mu) = \sum_{i=1}^n \alpha_i \mathcal{D}_K^2(\mu, \mu_i). \tag{34}$$

By Proposition 2 in (Cohen et al., 2021) the barycenter is given by $\mu^* = \sum_{i=1}^n \alpha_i \mu_i$. We illustrate an example, where we compute the MMD barycenter between the measures μ_1 and μ_2, which are uniformly distributed on the unit circle and uniformly distributed on the boundary of the square on $[-1, 1]^2$, respectively. The starting measure is δ_0 for the neural backward and neural forward scheme and a small square of radius $R = 10^{-9}$ around δ_0 for the particle flow. Note that the measures μ_1 and μ_2 are supported on different submanifolds. In Fig. 12 we illustrate the corresponding neural backward scheme (top), the neural forward scheme (middle) and the particle flow (bottom). Obviously, all methods are able to approximate the correct MMD barycenter.

I. MNIST starting in a uniform distribution

Here we recompute the MNIST example from Sect. 6.2 starting in an absolutely continuous measure instad of a singular measure. More precisely, the initial particles of all schemes are uniformly distributed in $\mathcal{U}_{[0,1]^d}$. Then we use the same

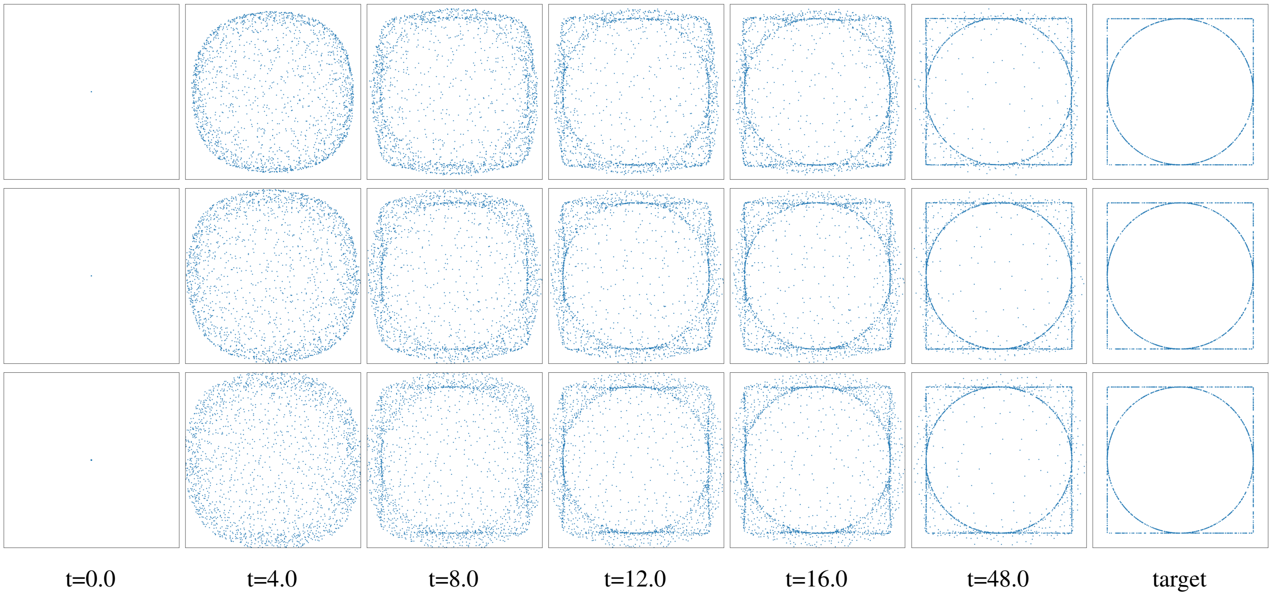


Figure 12: Neural backward scheme (top), neural forward scheme (middle) and particle flow (bottom) for the energy functional (34) starting in δ_0.

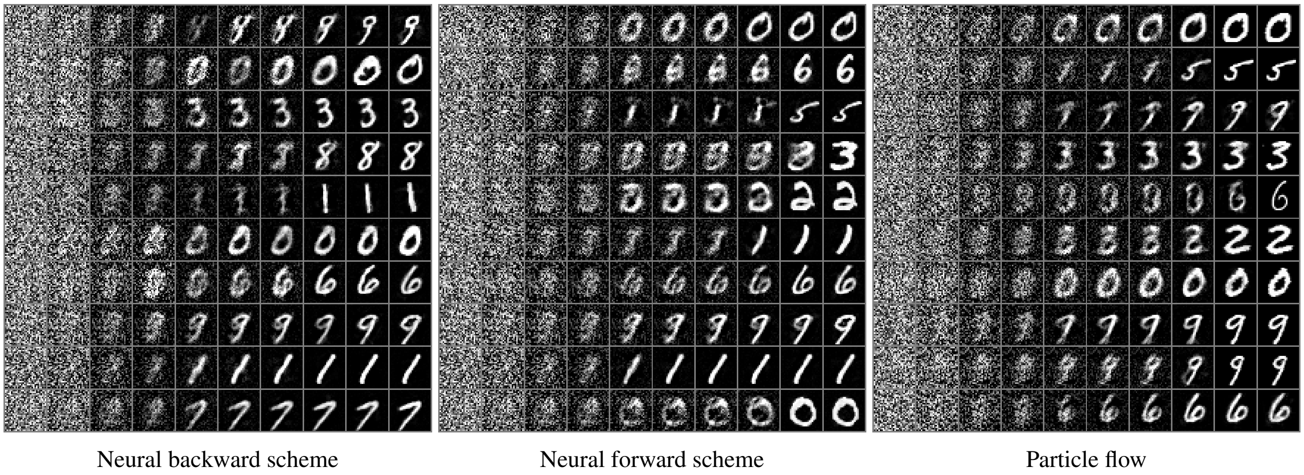


Figure 13: Samples and their trajectories from MNIST starting in $\mathcal{U}_{[0,1]^d}$.

experimental configuration as in Sect. 6.2. In Fig. 13 we illustrate the trajectories from MNIST of the different methods. In contrast to Sect. 6.2, where the particle flow suffered from the inexact starting because of the singular starting measure, in this case the methods behave similarly.