NEURAL SAMPLING FROM BOLTZMANN DENSITIES: FISHER-RAO CURVES IN THE WASSERSTEIN GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

We deal with the task of sampling from an unnormalized Boltzmann density ρ_D by learning a Boltzmann curve given by energies f_t starting in a simple density ρ_Z. First, we examine conditions under which Fisher-Rao flows are absolutely continuous in the Wasserstein geometry. Second, we address specific interpolations f_t and the learning of the related density/velocity pairs (ρ_t, v_t). It was numerically observed that the linear interpolation, which requires only a parametrization of the velocity field v_t, suffers from a "teleportation-of-mass" issue. Using tools from the Wasserstein geometry, we give an analytical example, where we can precisely measure the explosion of the velocity field. Inspired by Máté and Fleuret, who parametrize both f_t and v_t, we propose an interpolation which parametrizes only f_t and fixes an appropriate v_t. This corresponds to the Wasserstein gradient flow of the Kullback-Leibler divergence related to Langevin dynamics. We demonstrate by numerical examples that our model provides a well-behaved flow field which successfully solves the above sampling task.

1 INTRODUCTION

In this paper, we consider the problem of sampling from a Boltzmann density $\rho_D = e^{-f_D}/Z_D$ with unknown normalizing constant Z_D. Our approach is in the spirit of recent developments in generative modelling and aims to construct a curve $\rho_t = e^{-f_t}/Z_t$ interpolating between a simple density ρ_Z and the target ρ_D i.e. $\rho_0 = \rho_Z$ and $\rho_1 = \rho_D$. If such curve admits a velocity field v_t and there exists a solution φ of

$$\partial_t \varphi_t = v_t(\varphi_t), \quad \varphi_0 = \mathrm{Id} \quad \text{where} \quad (\varphi_t)_\sharp(\rho_Z \mathrm{d}x) = \rho_t \mathrm{d}x, \tag{1}$$

then we can use φ_1 and ρ_Z to sample from ρ_D. An important question is, whether for a given family of functions f_t, such velocity fields exists. While, for a fixed time t, the existence of such velocity fields is well established, see e.g. Laugesen et al. (2014), global existence with integrability of $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ in time and space is addressed in this paper. This is directly related to the question whether a large class of Fisher-Rao curves ρ_t is absolutely continuous in the Wasserstein geometry, and in particular, fulfills a continuity equation. This can be reduced to finding solutions of a certain family of PDEs. With the appropriate Hilbert spaces at hand, we show the existence of the solutions of these PDEs and the existence of an integrable velocity field v_t in time and space which admits moreover a minimality property of $\|v_t\|_{L_2(\mathbb{R}^d,\rho_t)}$.

Determining φ_1 requires both finding a curve f_t and the associated vector field v_t. One way is to choose a curve interpolating between f_Z and f_D first and then to learn the velocity field. For the linear interpolation $f_t = (1-t)f_Z + tf_D$, using above considerations, we show that there indeed exists an integrable velocity field implying the absolute continuity of the curve in the Wasserstein space. Yet, it was numerically shown that the corresponding velocity field is often badly behaved as noted in (Máté & Fleuret, 2023, Figure 4 and 6). We highlight by an analytic example the bad regularity of this flow and show in particular that its velocity field may have an exploding norm $\|v_t\|_{L_2(\mathbb{R}^d,\rho_t)}$. Another approach, recently proposed by Máté & Fleuret (2023), considers $f_t = (1-t)f_Z + tf_D + t(1-t)\psi_t$ with unknown ψ and learns (ψ_t, v_t) simultaneously. Henceforth, we refer to this method as the "learned interpolation". However, it is theoretically unclear whether the learned curve is well-behaved or the velocity field is optimal with respect to the above norm. As

an alternative approach, we propose instead to deal *directly* with a well-behaved and mathematically accessible curve on $[0, T]$ and consider the parameterization $f_t = \frac{T-t}{T} f_D + t\psi_t$. Fixing the velocity field as $v_t := \nabla(f_t - f_Z)$ we only have to learn ψ_t. The resulting PDE is the well-known Fokker-Planck equation related to Langevin dynamics which has a solution with Boltzmann densities ρ_t, if, e.g., ρ_Z is a Gaussian. Then, the corresponding SDE is the Ornstein-Uhlenbeck process. Here, the backward ODE of (1) must be applied for sampling. Finally, we learn the networks for the above three cases and demonstrate by numerical examples the effectiveness of our approach.

Contributions.

1. We prove for the first time that under mild conditions every reasonable curve of Boltzmann densities is an absolutely continuous Wasserstein curve, see Theorems 1 and 2 and Appendix A. In particular, this includes the linear interpolation.
2. We verify by a nontrivial analytic example that the velocity field corresponding to the linear interpolation can admit an exploding norm $\|v_t\|_{L_2(\mathbb{R}^d,\rho_t)}$, leading to a "non-smooth" particle transport, see Figure 1 and Appendix B.
3. We propose to learn a curve using the parameterization $f_t = \frac{T-t}{T} f_D + t\psi_t$, which we call "gradient flow interpolation" by the following reason. With an appropriate velocity field which can be computed directly from f_t, this resembles the Wasserstein *gradient flow* of the Kullback-Leibler divergence with fixed second argument ρ_Z. This curve can be also described as the Fokker-Planck equation of a Langevin SDE.
4. We learn the networks for the linear, learned and gradient flow interpolations and demonstrate the performance of these three methods on certain sampling problems.

Related Work. Classical approaches for sampling from an unnormalized density are based on the Markov Chain Monte Carlo (MCMC) method, see Gilks et al. (1995), including variants as the Hamiltonian Monte Carlo (HMM) in Hoffman & Gelman (2014) or the Metropolis Adjusted Langevin Algorithm (MALA) in Girolami & Calderhead (2011). Arising problems with the correct distribution of mass among different modes lead to the use of importance sampling Neal (2001) and sequential Monte Carlo samplers Moral et al. (2006).
Newer approaches include the simulation of gradient flows of probability measures. In Chen et al. (2024) Gaussian mixtures are used to approximate the Fisher-Rao gradient flow of the Kullback-Leibler KL divergence. This corresponds to a time rescaling of the linear interpolation. In Nüsken (2024); Maurais & Marzouk (2024) they approximate the latter curve by kernelizing the associated Poisson equation and simulating a corresponding interacting particle system. Closely related Stein variational gradient descent Liu & Wang (2016) kernelizes the Wasserstein gradient flow of the reverse KL divergence. Recently in Chehab & Korba (2024) they consider a different path which they term the "dilation path". The associated vector field is given as the gradient of the dilation of the target density. The convolution of the density of the dilation path with (scaled) Gaussian kernels corresponds to the density of an Ornstein-Uhlenbeck process.
Recently, (stochastic) normalizing flows have been widely used to sample from unnormalized densities Noé et al. (2019); Wu et al. (2020); Hagemann et al. (2022). In Midgley et al. (2023) this problem is tackled without using a curve of probability measures by augmenting normalizing flows, which are trained using an α-divergence loss, with annealed importance sampling. In Xu et al. (2024) the Wasserstein gradient flow of $KL(\cdot, \rho_D)$ is approximated. They learn a network for approximating the corresponding JKO scheme. In a similar approach in Hertrich & Gruhlke (2024) a functional F with minimum ρ_D is considered. They then learn a vector field by approximating a JKO scheme for the Wasserstein gradient flow of F and combine it with an importance based rejection method. They focus on F being the Kullback-Leibler divergence with target distribution ρ_D. Using KL for sampling is a natural choice and it is shown in Chen et al. (2023) that it is the only f-divergence such that the Wasserstein gradient flow does not depend on the normalization constant of ρ_D.
Note that parametrizing the vector field using the score $\nabla \log p_t$ is common practice for sampling via SDEs. A method that generates samples and scores from trajectories of the Fokker Planck equation of certain SDEs is presented in Boffi & Vanden-Eijnden (2023). In Nüsken et al. (2024) the vector field is learned by using the KL divergence on the path measures of a forward and the corresponding backward SDE assuming knowledge of $\nabla \log p_t$. In Zhang & Chen (2022) for example where they formulate the sampling problem in terms of a stochastic control problem and

explicitly include the score into the parametrization of the vector field. The stochastic control viewpoint is also used in (Berner et al., 2024, Section 3) where the KL divergence between path measures associated to a controlled and an uncontrolled SDE is used to approximate the score. A different approach using the KL on path measures is formulated in Vargas et al. (2023). Another way of approximating a flow ending in ρ_D is to consider the variance exploding diffusion SDE and approximating the score $\nabla \log p_t$ which is done in Akhound-Sadegh et al. (2024).

The authors of Máté & Fleuret (2023) present a loss based on the continuity equation and the special form $\frac{e^{-f_t}}{Z_t}$ of Boltzmann densities. This loss allows for using fixed interpolations f_t and only learning v_t as well as learning f_t and v_t simultaneously. Using a fixed interpolation and a loss based on the continuity equation is also an idea showing in the physics informed neural networks (PINN) literature, see e.g. Albergo & Vanden-Eijnden (2024). When finishing this paper, we became aware of the recent preprint of Sun et al. (2024) and the workshop paper Sun et al. comparing a variety of PDE based sampling methods. They included a similar approach as the one proposed here, using the same parametrization to simulate an Ornstein-Uhlenbeck SDE. However, coming from the Wasserstein perspective, we came to the conclusion that the associated gradient flow path is especially well-behaved. We simulate the corresponding *ODE* and highlight its benefits.
For the proof that many Boltzmann density curves are absolutely continuous Wasserstein curves, the main object of interest is

$$-\nabla \cdot (\rho_t \nabla s_t) = (\alpha_t - \overline{\alpha_t})\, \rho_t, \quad t \in [0,1] \tag{2}$$

which relates Fisher-Rao curves and Wasserstein absolutely continuous curves. Here ∇s_t corresponds to the vector field of a Wasserstein absolutely continuous curves and $\alpha_t - \bar{\alpha}_t$ for $\bar{\alpha}_t = \mathbb{E}_{\rho_t}[\alpha_t]$ corresponds to the gradient in the Fisher-Rao geometry. This equation was studied for a fixed time t in various contexts. In nonlinear filtering it is used to study weak solutions of nonlinear filters see e.g. (Laugesen et al., 2014, Section 2.1). Their theory for fixed time steps t can be generalized to the case of a time dependent equation as seen in Section A. The close relationship between the Poincare inequality and fixed time solutions of (2) is discussed in Dhara & Kałamajska (2015).

2 Wasserstein Flows of Boltzmann Densities

2.1 General Background

Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability measures on $\mathbb{R}^d$. For $\mu \in \mathcal{P}(\mathbb{R}^d)$, let $L^p(\mathbb{R}^d, \mu)$, $p \in [1, \infty)$, denotes the Banach space of (equivalence classes of) real-valued functions on $\mathbb{R}^d$ with finite norm $\|f\|_{L_p(\mathbb{R}^d,\mu)} := \left(\int_{\mathbb{R}^d} |f|^p \,\mathrm{d}\mu\right)^{1/p}$, and $L_p(\mathbb{R}^d, \mathbb{R}^d, \mu)$ be the Banach space of Borel vector fields $v : \mathbb{R}^d \to \mathbb{R}^d$ with finite norm $\|v\|_{L_p(\mathbb{R}^d,\mathbb{R}^d,\mu)} := (\int_{\mathbb{R}^d} \|\|v\|\|^p \,\mathrm{d}\mu)^{1/p}$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$. For the Lebesgue measure, we skip the μ in the notation. The space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures having finite second moments equipped with the Wasserstein(-2) metric

$$W_2(\mu_0, \mu_1) := \min_{\pi \in \Gamma(\mu_0,\mu_1)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \,\mathrm{d}\pi(x,y) \right)^{\frac{1}{2}},$$

where $\Gamma(\mu_0, \mu_1)$ denotes the set of couplings or plans $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ having marginals $\mu_i \in \mathcal{P}_2(\mathbb{R}^d)$, $i = 0, 1$, is a complete metric space. In complete metric spaces, we can consider absolutely continuous curves depending on the metric of the space. In the above Wasserstein space, a weakly continuous curve $\mu : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$, $t \mapsto \mu_t$ is *absolutely continuous*, if there exists a Borel vector field $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ with $\int_0^1 \|v_t\|_{L_2(\mathbb{R}^d,\mathbb{R}^d,\mu_t)} \,\mathrm{d}t < \infty$ such that the *continuity equation*

$$\partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0 \tag{3}$$

is fulfilled in the sense of distributions [1] see, e.g. Ambrosio et al. (2005). While for fixed μ_t there are many vector fields such that the continuity equation is fulfilled, there exists a unique one such

[1] $\int_0^1 \int_{\mathbb{R}^d} \partial_t \phi + \langle \nabla_x \phi, v_t \rangle \,\mathrm{d}\mu_t \mathrm{d}t = 0$ for all $\phi \in C_c^\infty\left((0,1) \times \mathbb{R}^d\right)$ where $C_c^\infty((0,1) \times \mathbb{R}^d)$ denotes the space of smooth and compactly supported functions.

that $\|v_t\|_{L_2(\mathbb{R}^d,\mu_t)}$ becomes minimal for almost every $t \in [0,1]$. In other words, if μ_t fulfills the continuity equation for some vector field, then there exists a unique vector field that solves

$$\operatorname*{arg\,min}_{v_t \in L_2(\mathbb{R}^d,\mathbb{R}^d,\mu_t)} \int_{\mathbb{R}^d} \|v_t\|^2 \,\mathrm{d}\mu_t \quad \text{s.t.} \quad \partial_t \mu_t + \nabla \cdot (\mu_t v_t) = 0. \tag{4}$$

Moreover, this optimal vector field is determined by the condition that $v_t \in T_{\mu_t}\mathcal{P}_2(\mathbb{R}^d)$ for almost every $t \in [0,1]$, with the *regular tangent space*

$$T_\mu \mathcal{P}_2(\mathbb{R}^d) := \overline{\{\nabla\phi : \phi \in C_c^\infty(\mathbb{R}^d)\}}^{L_2(\mathbb{R}^d,\mu)}, \tag{5}$$

where the right hand side is the closure in the Banach space $L_2(\mathbb{R}^d, \mu)$, see (Ambrosio et al., 2005, § 8). Finally, if μ_t is an absolutely continuous curve with Borel vector field v_t such that for every compact Borel set $B \subset \mathbb{R}^d$ it holds $\int_0^1 \sup_B \|v_t\| + \mathrm{Lip}(v_t, B)\,\mathrm{d}t < \infty$, then there exists a solution $\varphi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ of

$$\partial_t \varphi(t,x) = v_t(\varphi(t,x)), \quad \varphi(0,x) = x \tag{6}$$

which fulfills $\mu_t = \varphi_{t,\sharp}\mu_0 := \mu_0 \circ \varphi_t^{-1}$, see e.g. (Ambrosio et al., 2005, Proposition 8.1.8).

2.2 Wasserstein meets Fisher-Rao Flows

In the following, we are interested in absolutely continuous probability measures with respect to the Lebesgue measure, i.e., the space $\mathcal{P}^{ac}(\mathbb{R}^d) := \{\mu = \rho\,\mathrm{d}x : \rho \in L_1(\mathbb{R}^d),\ \rho \geq 0,\ \int_{\mathbb{R}^d} \rho\,\mathrm{d}x = 1\}$ and we will address such measures just by their densities. Moreover, we will only consider Boltzmann/Gibbs densities depending on functions $f : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ with $e^{-f_t} \in L_1(\mathbb{R}^d)$ and $f(\cdot, x) \in C^1([0,1])$ for all $x \in \mathbb{R}^d$ and $\partial_t f(t,\cdot) \in L_2(\mathbb{R}^d, \rho_t)$ such that

$$\rho_t := \frac{e^{-f_t}}{Z_t}, \quad Z_t := \int_{R^d} e^{-f_t}\,\mathrm{d}x. \tag{7}$$

In this case, straightforward computation gives

$$\partial\mu_t = \partial_t \rho_t = -\left(\partial_t f_t - \mathbb{E}_{\rho_t}[\partial_t f_t]\right)\rho_t, \tag{8}$$

where it maybe useful to note that

$$\mathbb{E}_{\rho_t}[\partial_t f_t] = -\partial_t \log Z_t.$$

This differential equation is a Fisher-Rao flow equation. More precisely, a weakly continuous curve $\rho : [0,1] \to \mathcal{P}^{ac}(\mathbb{R}^d)$, $t \mapsto \rho_t$ is a *Fisher-Rao curve*, if there exists a measurable map $\alpha : [0,1] \times \mathbb{R}^d \to \mathbb{R}$ such that $\alpha_t := \alpha(t,\cdot) \in L_2(\mathbb{R}^d, \rho_t)$ and we have in a distibutional sense

$$\partial_t \rho_t = (\alpha_t - \bar{\alpha}_t)\rho_t, \quad \bar{\alpha}_t := \mathbb{E}_{\rho_t}[\alpha_t]. \tag{9}$$

Obviously, (8) delivers a Fisher-Rao curve with $\alpha_t := -\partial_t f_t$. For more information on Fisher-Rao flows see, e.g., Gallouët & Monsaingeon (2017); Wang & Li (2022). We are interested in the question under which conditions a Fisher-Rao curve determined by (8) is also an absolutely continuous curve in the Wasserstein space. For later purposes, note that for Boltzmann densities (7), the second summand of the continuity equation (3) reads as

$$\nabla \cdot (\mu_t v_t) = \nabla \cdot (\rho_t v_t) = \left(-\langle \nabla f_t, v_t\rangle + \nabla \cdot v_t\right)\rho_t. \tag{10}$$

Nevertheless, using (8), for our Fisher-Rao curves the continuity equation becomes

$$\left(\partial_t f_t - \mathbb{E}_{\rho_t}[\partial_t f_t]\right)\rho_t = \nabla \cdot (\rho_t v_t), \tag{11}$$

and in the ideal case that we can find a vector field $v_t = \nabla s_t$, compare (5), this can be rewritten as the family of Poisson equations

$$\left(\partial_t f_t - \mathbb{E}_{\rho_t}[\partial_t f_t]\right)\rho_t = \nabla \cdot (\rho_t \nabla s_t), \tag{12}$$

which needs to be solved for s_t. For *fixed time* t, the solution of the Poisson equation has already been examined, e.g., in Laugesen et al. (2014) and for more recent work, see also Reich (2011); Taghvaei & Mehta (2023). However, the pointwise solution for each fixed t does not ensure that $(\rho_t, \nabla s_t)$ gives rise to an absolutely continuous Wasserstein curve, since by definition this requires additional properties of $\nabla_x s : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, globally in t: i) $\nabla_x s$ is Borel measurable on $[0,1] \times \mathbb{R}^d$, and ii) $t \mapsto \|\nabla_x s(t,\cdot)\|^2_{L_2(\rho_t,\mathbb{R}^d)} \in L_1([0,1])$. If also, iii) $\nabla_x s(t,\cdot) \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d)$ for a.e. $t \in [0,1]$, then it is known that the velocity field $\nabla_x s$ is the *optimal one* in the sense of (4). To our best knowledge, the following two theorems, which are detailed in Appendix A, tackle a *rigorous* solution of this problem for the first time.

Theorem 1. *Assume that $\boldsymbol{\rho}$ determined by $\int_{[0,1]\times\mathbb{R}^d} f(t,x)\,\mathrm{d}\boldsymbol{\rho} = \int_0^1\int_{\mathbb{R}^d} f(t,x)\,\mathrm{d}\rho_t\mathrm{d}t$ satisfies a so-called partial Poincare inequality, see Definition A.7, for some $K > 0$, and that $\frac{1}{\boldsymbol{\rho}}$ is locally integrable. Furthermore, suppose that $\alpha_t - \overline{\alpha_t} \in L_2([0,1]\times\mathbb{R}^d, \boldsymbol{\rho})$. Then, there exists a unique weak solution $\boldsymbol{s}$ in a certain Hilbert space $\mathcal{X}_0^1$ of the problem*

$$\int_0^1\int_{\mathbb{R}^d} \langle \nabla \boldsymbol{s}_t, \nabla\psi_t\rangle\,\mathrm{d}\rho_t\mathrm{d}t = \int_0^1\int_{\mathbb{R}^d} \psi_t\left(\alpha_t - \overline{\alpha_t}\right)\,\mathrm{d}\rho_t\mathrm{d}t, \quad \textit{for all } \psi \in \mathcal{X}_0^1. \tag{13}$$

Theorem 2. *Let $\rho_0, \rho_1 \in \mathcal{P}(\mathbb{R}^d)$ and assume that ρ_t is a Fisher-Rao curve defined by* (9)*, Let $\boldsymbol{\rho}$ and α satisfy the assumptions of Theorem 1. Then, ρ_t is a Wasserstein absolutely continuous curve with vector field $\nabla_x \boldsymbol{s}$, where $\boldsymbol{s} \in \mathcal{X}_0^1$ is the unique weak solution of* (13)*. If ρ_t is bounded from above, it holds that $\nabla_x \boldsymbol{s}_t \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d)$ for a.e. $t \in [0,1]$. Furthermore, it holds that*

$$\int_0^1\int_{\mathbb{R}^d} \|\nabla_x \boldsymbol{s}_t\|^2\,\mathrm{d}\rho_t\mathrm{d}t \le K\int_0^1\int_{\mathbb{R}^d} \left(\alpha_t - \overline{\alpha_t}\right)^2\,\mathrm{d}\rho_t\mathrm{d}t. \tag{14}$$

Note that by the Benamou-Brenier formula, see (Ambrosio et al., 2005, Equation 8.0.3), the left-hand side of (14) is an upper bound for the Wasserstein distance $W_2^2(\rho_0,\rho_1)$, while the double-integral on the right-hand side defines the Fisher-Rao *action* of ρ_t.

3 Neural Sampling from Boltzmann Densities

Our aim is to sample from a Boltzmann measure $\mu_D = \rho_D\,\mathrm{d}x$, where we only have access to $f_D : \mathbb{R}^d \to \mathbb{R}$, but not to the normalization constant Z_D, by following a measure flow starting in a simple to sample measure $\mu_Z = \rho_Z\,\mathrm{d}x$ like the standard Gaussian one. We apply the continuity equation (11) of Fisher-Rao curves, which can be rewritten by (10) and division by $\rho_t > 0$ as

$$\partial_t f_t - \underbrace{\mathbb{E}_{\rho_t}[\partial_t f_t]}_{C_t} = -\langle \nabla f_t, v_t\rangle + \nabla\cdot v_t. \tag{15}$$

Indeed every spatially constant function C_t which fulfills the above equation must be of the form $C_t = \mathbb{E}_{\rho_t}[\partial_t f_t]$, see Máté & Fleuret (2023). This enables us to learn C_t, so that (15) does no longer depend on normalization constants. We will deal with the following parameterizations of f_t and v_t and how to learn them:

1. linear interpolation: $f_t \coloneqq (1-t)f_Z + tf_D \;\hookrightarrow$ learn v_t,
2. learned interpolation by Máté and Fleuret: $f_t \coloneqq (1-t)f_Z + tf_D + t(1-t)\psi_t \;\hookrightarrow$ learn (ψ_t, v_t),
3. gradient flow interpolation: $f_t \coloneqq \frac{T-t}{T}f_D + t\psi_t$, $v_t \coloneqq \nabla(f_t - f_Z) \;\hookrightarrow$ learn ψ_t.

While by definition the first two settings interpolate between ρ_Z and ρ_D and $\rho_0 = \rho_Z$, $\rho_1 = \rho_D$, we will see that the newly proposed third variant interpolates between ρ_D and ρ_Z by the choice of v_t, i.e., $\rho_0 = \rho_D$ and $\lim_{T\to\infty}\rho_T = \rho_Z$. In practice we reparametrize the latter curve into unit time, see Remark 4.

3.1 Linear & Learned Interpolation

Linear interpolation. The simplest interpolation is the linear one $f_t \coloneqq (1-t)f_Z + tf_D$, i.e., $\rho_t \propto \rho_0^{1-t}\rho_1^t$. Under suitable conditions on the end points this density fulfills the assumptions of Theorem 2, see Corollary A.17 in the Appendix, and is therefore a Wasserstein curve. By (8), it also defines a Fisher-Rao curve, and the corresponding Fisher-Rao flow equation (9) is given by

$$\begin{aligned}\partial_t\rho_t &= \left(f_0 - f_1 - \mathbb{E}_{\rho_t}[f_0 - f_1]\right)\rho_t \\ &= -\Big(\log(\frac{\rho_0}{\rho_1}) - \mathbb{E}_{\rho_t}[\log(\frac{\rho_0}{\rho_1})]\Big)\rho_t.\end{aligned}$$

Note that, ρ_t is even a Fisher-Rao *gradient flow* of the negative log likelihood, see also Maurais & Marzouk (2024), given by

$$\mathcal{E}_{\mathrm{NLL}}(\rho) \coloneqq -\mathbb{E}_\rho\left[\log(\frac{\rho_1}{\rho_0})\right].$$

There are a variety of approaches for approximating the vector field corresponding to the linear interpolation. In Maurais & Marzouk (2024); Nüsken (2024) a kernel-based method for solving (12) was proposed. In Máté & Fleuret (2023) they propose to solve (15) which in case of the linear interpolation becomes

$$f_1 - f_0 - C_t + \langle \nabla f_t, v_t \rangle - \nabla \cdot v_t = 0. \tag{16}$$

This equation has many solutions v_t for the same ρ_t, and in order to enforce uniqueness we could demand that $v_t = \nabla s_t$.

This approach leads to learning neural networks $(v_t^{\theta_1}, C_t^{\theta_2})$ by minimizing the loss function

$$L(\theta) := \mathbb{E}_{t \in \mathcal{U}[0,1], x \in \mathcal{U}[a,b]^d} \left[\mathcal{E}(\theta, x, t) \right], \tag{17}$$
$$\mathcal{E}(\theta, x, t) := |f_1 - f_0 - C_t^{\theta_2} + \langle \nabla f_t, v_t^{\theta_1} \rangle - \nabla \cdot v_t^{\theta_1}|^2, \quad \theta := (\theta_1, \theta_2).$$

Remark 3. *Since the approximation of high dimensional integrals is difficult Máté & Fleuret (2023) propose to use an iterative approach by solving the ODE* $\partial_t \varphi_t^{\theta'} = v_t^{\theta'}(\varphi_t^{\theta'})$, $\varphi_0^{\theta'} = \mathrm{Id}$ *for* θ' *from the previous minimization and to solve*

$$L(\theta) := \mathbb{E}_{t \in \mathcal{U}[0,1], z \sim \rho_0} [\mathcal{E}(\theta, \varphi_t^{\theta'}(z), t)].$$

Heuristically this penalizes $\mathcal{E}(\theta, x, t)$ *stronger in regions where the intermediate solution has more mass. Finally, sampling can be done – assuming sufficient regularity of the learned* $v_t^{\theta_i}$ *– by simulating the ODE* (6)*. We compare both methods in our experiments in Section 4.*

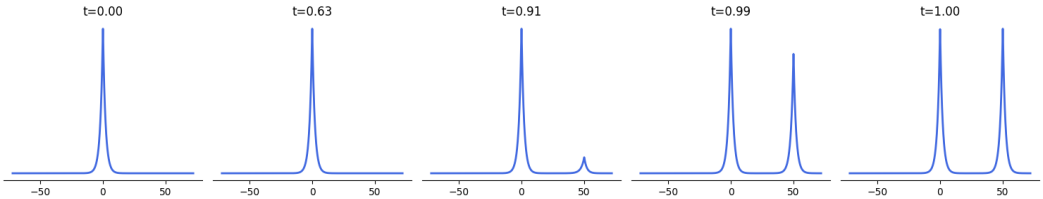


Figure 1: Evolution of the probability densities $\rho_t \propto \rho_0^{1-t}\rho_1^t$, where $\rho_0 \propto e^{-|x|}$ and $\rho_1 \propto e^{-2\min\{|x|,|x-m|\}}$ for $m = 50$. For different values of m see 5.

A teleportation issue with the linear interpolation. Máté & Fleuret (2023) gave examples, where for an asymmetric constellation of target and prior measure, this path transports mass into distant modes only at very late times which may cause problems while learning the corresponding vector field. Mathematically, this corresponds to the fact that the vector field $v_t(x)$ develops a singularity in $x \in \mathbb{R}^d$ for late times t, see (Máté & Fleuret, 2023, Figure 6). In this work, we demonstrate that this phenomenon is *more severe*: even when averaging over the space x, the norm $\|v_t\|_{L_2(\rho_t)}$ of the optimal field (4) explodes for late times t. For an example in $\mathbb{R}^1$ involving a Laplacian $\rho_0 \propto e^{-|x|}$ centered at 0 and a Laplacian-like distribution $\rho_1 \propto e^{-2\min\{|x|,|x-m|\}}$ with a mode at 0 and a second mode at $m > 0$, see Figure 1, we are able to give an explicit and analytic formula for the calculation of $\|v_t\|_{L_2(\rho_t)}$ in the Appendix B. Using this formula (36), we demonstrate that $\|v_t\|_{L_2(\rho_t)}$ explodes when t is close to 1, and that this explosion happens more severely and at later times, the further the second mode $m > 0$ is away from 0, see Figure 2.

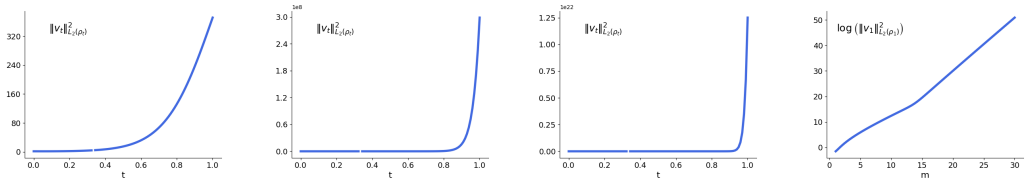
Learned Interpolation. To cope with the above disadvantages of the linear interpolation, Máté & Fleuret (2023) proposed to parameterize also the flow by $f_t := (1-t)f_0 + tf_1 + t(1-t)\psi_t$ with a neural network $\psi_t^{\theta_1}$ and to minimize according to (15) the above loss with

$$\mathcal{E}(\theta, x, t) := |\partial_t f_t^{\theta_1}(x) - C_t^{\theta_3} + \langle \nabla f_t^{\theta_1}, v_t^{\theta_2} \rangle - \nabla \cdot v_t^{\theta_2}|^2, \quad \theta := (\theta_1, \theta_2, \theta_3). \tag{18}$$

and

$$L(\theta) := \mathbb{E}_{t \in \mathcal{U}[0,1], z \sim \rho_0} [\mathcal{E}(\theta, \varphi_t^{\theta'}(z), t)].$$

where $\varphi_t^{\theta'}(z)$ is as in Remark 3 a trajectory of the intermediate solution. Note that in general there are infinitely many solutions of (16) and neither the parameterized curve ρ_t nor the vector field v_t is unique. Further $v_t^{\theta_2}$ obtained in this way may not be minimal in any sense. The proof of the validity of the interpolation with respect to Theorem 2 is future work.



$m = 5$ $m = 15$ $m = 30$ $t = 1$

Figure 2: In the first three figures, the evolution of $\|v_t\|^2_{L_2(\rho_t)}$ belonging to the path in Figure 1 is depicted for different values of the second mode $m > 0$. For larger m, the norm $\|v_t\|^2_{L_2(\rho_t)}$ approaches the limit $\|v_1\|^2_{L_2(\rho_1)}$ at later times, hence with a steeper slope. The last figure shows the *log scale* of $\|v_1\|^2_{L_2(\rho_1)}$ depending on m. It demonstrates that $\|v_1\|_{L_2(\rho_1)}$ roughly grows *exponentially* with m.

3.2 Gradient Flow Interpolation

Recall that the linear interpolation of the energies f_0, f_1 leads to a multiplicative interpolation of densities $\rho_t(x) \propto \rho_Z(x)^{1-t}\rho_D(x)^t$. Note that this does not take account of any *local* behaviour of ρ_Z, ρ_D in a neighbourhood of x, resulting in irregular behavior of the vector field as shown in Figure 1 and 2. In contrast, widely used paths in generative modelling Song et al. (2021); Lipman et al. (2023) originate from an interpolation of densities via (spatial) convolutions $\rho_t \sim \rho_Z(\frac{\cdot}{a_t}) * \rho_D(\frac{\cdot}{b_t})$ for time schedules a_t, b_t, corresponding to an interpolation $X_t = a_t Z + b_t X_0$ of associated independent random variables $Z \sim \rho_Z, X_0 \sim \rho_D$ on the *spatial level*. We aim to learn such a curve and an associated vector field.

In this paper, we propose to use $f_t := \frac{T-t}{T} f_D + t\psi_t$ and $v_t := \nabla(f_t - f_Z)$ and to parameterize ψ_t by a neural network. Then, equation (15) becomes

$$\partial_t f_t - \mathbb{E}_{\rho_t}[\partial_t f_t] + \langle \nabla f_t, \nabla(f_t - f_Z)\rangle - \Delta(f_t - f_Z) = 0. \tag{19}$$

Indeed, we will show in the paragraph below that this equation makes perfect sense, in particular, it has a solution for $\rho_Z \sim \mathcal{N}(0,1)$. Then we can compute $(\psi_t^{\theta_1}, C_t^{\theta_2})$ for $f_t^{\theta_1} := \frac{T-t}{T} f_D + t\psi_t^{\theta_1}$ by minimizing the loss L in (17) with the function

$$\mathcal{E}(\theta, x, t) := |\partial_t f_t^{\theta_1} - C_t^{\theta_2} + \langle \nabla f_t^{\theta_1}, \nabla(f_t^{\theta_1} - f_Z)\rangle - \Delta(f_t^{\theta_1} - f_Z)|^2, \quad \theta := (\theta_1, \theta_2).$$

Finally, to sample from the target distribution, we simulate the ODE (6) backwards on $[0, T]$ to obtain $\tilde{\varphi} : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ satisfying

$$\partial_t \tilde{\varphi}(t, x) = -v_t(\tilde{\varphi}(t, x)), \quad \tilde{\varphi}(T, x) = x$$

which fulfills $\mu_t = \tilde{\varphi}_{T-t,\sharp}\mu_Z$.

Relation to Wasserstein gradient flows. Equation (19) is well-known in connection with Wasserstein gradient flows. A *Wasserstein gradient flow* is an absolutely continuous curve whose vector field $v_t = -\text{grad}_{\rho_t} F \in T_{\rho_t}(\mathcal{P}_2(\mathbb{R}^d))$ is given by the negative Wasserstein gradient (or more generally, *subdifferential*) of a function $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$. Then, the continuity equation

$$\partial_t \rho_t - \nabla \cdot \left(\rho_t \text{grad}_{\rho_t} F\right) = 0 \tag{20}$$

produces a flow towards a minimizer of F. A prominent example, see, e.g. (Ambrosio et al., 2005, Chapter 10.4), is $F(\rho) := \text{KL}(\rho|\rho_Z)$ arising from the Kullback-Leibler divergence which has the unique minimizer ρ_Z. Since it is easy to show, see (Ambrosio et al., 2005, Lemma 10.4.1), that $\text{grad}_{\rho_t} F = \nabla \log \frac{\rho_t}{\rho_Z}$, we can be rewrite (20) as

$$\partial_t \rho_t = -\nabla \cdot (\rho_t(\nabla \log \rho_Z - \nabla \log \rho_t)), \quad \rho_0 = \rho_D, \tag{21}$$

which is also known as the Fokker-Planck equation. Inserting $\rho_t := e^{-f_t}/Z_t$ results exactly in (19). The existence and uniqueness of a solution of (21) is shown in (Ambrosio & Savaré, 2007, Theorem 6.6). However, it is not clear if Boltzmann densities stay Boltzmann densities i.e. if the densities ρ_t are strictly positive for all t. But in the special case, when ρ_Z is a Gaussian, this is the case as we

will see next.
There is a close relation between (21) and SDEs, namely the Langevin SDE

$$\mathrm{d}X_t = \nabla \log \rho_Z(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \tag{22}$$

corresponds to the Fokker-Planck equation (21), see e.g., (Chewi et al., 2024, Chapter 6.2), i.e., $X_t \sim \rho_t$. For $\rho_Z \propto \mathcal{N}(0, I_d)$, (22) becomes $\mathrm{d}X_t = -X_t\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$. This is an Ornstein-Uhlenbeck process, which has the closed form solution

$$X_t = e^{-t}Z + e^{-t}X_0, \quad Z \sim \mathcal{N}\left(0, (e^{2t}-1)I_d\right), \quad X_0 \sim \rho_D,$$

where Z is independent from X_0, see (Herrmann & Massin, 2019, Eq. (2.4)). From the closed form solution it follows that $\rho_t \sim \mathcal{N}\left(0, (1-e^{-2t})I_d\right) * \rho_D\left(e^t\cdot\right) > 0$, and thus $\rho_t \propto e^{-f_t}$ which implies that (19) has a solution.

Remark 4. *The dynamics of both the linear and learned interpolations are on the unit interval. To obtain dynamics on the unit interval for the gradient flow interpolation we apply an approach from the score based diffusion literature Song et al. (2021). We use a linear time schedule* $\beta(t) := \beta_{min} + t\left(\beta_{max} - \beta_{min}\right)$ *and the associated SDE*

$$\mathrm{d}X_t = -\frac{1}{2}\beta(t)X_t\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t \quad X_0 \sim \rho_1. \tag{23}$$

The SDE (23) *has a closed form solution given by*

$$X_t = \sqrt{1-e^{-g(t)}}Z + e^{\frac{-g(t)}{2}}X_0, \tag{24}$$

where $g(t) := \int_0^t \beta(s)\,\mathrm{d}s$ *and* $Z \sim \mathcal{N}(0, I_d)$*. As done in Song et al. (2021) we choose* $\beta_{min} = 0.1$ *and* $\beta_{max} = 20$*. These choices ensure that the distribution of* X_1 *is approximately* $\mathcal{N}(0, I_d)$*. We then aim to learn the solution of the associated Fokker-Planck equation, this corresponds to setting* $v_t := \frac{\beta(t)}{2}\nabla(f_t - \frac{\|\cdot\|^2}{2})$.

The following example illustrates the behavior of the three interpolations, when learning the corresponding vector fields.

Example 5. *Motivated by an example in Máté & Fleuret (2023) and incorporating the idea of our teleportation example with asymmetrically diverging modes, we choose a standard Gaussian source distribution and the target distribution*

$$\frac{1}{3}\mathcal{N}\left([4,4],1\right) + \frac{2}{3}\mathcal{N}\left([-m,-m],1\right).$$

For the linear and learned interpolations, we consider both sampling from uniform domains ("uniform") and sampling from the trajectories ("trj") as described in Remark 3. For the gradient flow interpolation, we sample points uniformly and interpolate them with samples from our latent distribution according to formula (24)*. The results are reported in Figure 3. Note that, in case of the learned interpolation, the mode collapse could be eventually alleviated by choosing a different source distribution with a larger support and/or heavier tails.*

4 Experiments

In this section we apply the different approaches to common sampling problems. We compare the performance of using the linear, learned or gradient flow interpolation.

For both the linear interpolation and the learned interpolation, we can choose to learn the vector field directly, or learn a potential and obtain the vector field as its gradient. In practice, we choose to parameterize it directly as in Máté & Fleuret (2023). Note that in practice for the gradient flow interpolation we apply Remark 4. Furthermore we improve performance by using learned scheduling function $g(t)$, satisfying $g(0) = 1$ and $g(1) = 0$, and parametrizing the corresponding f_t as $f_t := g(t)f_D + t\psi_t$.

A major design choice is at which points to evaluate the pointwise error $\mathcal{E}(\theta, x, t)$. For the gradient flow interpolation, we chose the sampling strategy described in Example 5. For the linear and learned interpolation we experimented both with sampling from uniform domains, or sampling along the

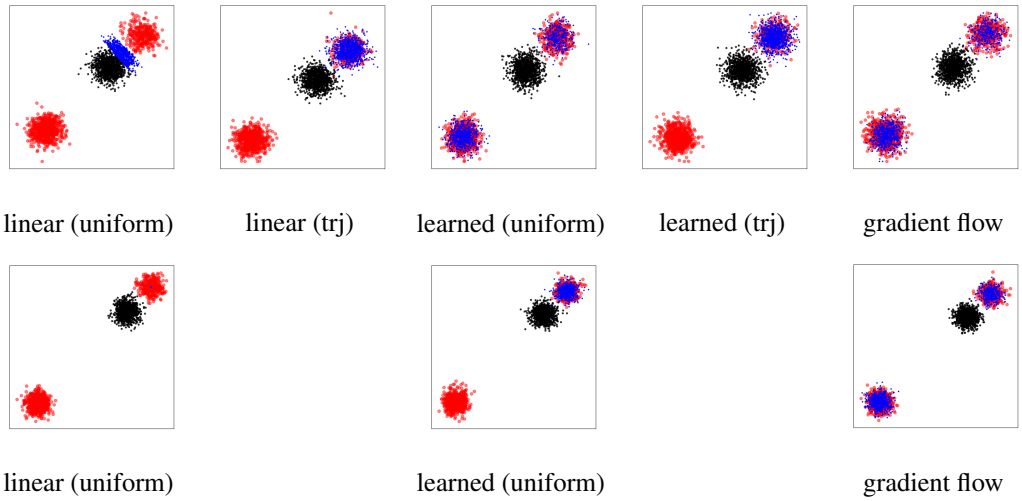


Figure 3: Sampling results for $m = 8$ (top) and $m = 15$ (bottom): source (black), target (red) and estimated (blue). We cannot learn a meaningful vector field for the linear interpolation for both sampling strategies. Sampling along the trajectories leads to mode collapse for the learned interpolation. While the learned interpolation with uniform sampling covers the modes for $m = 8$, increasing m leads to mode collapse. In contrast the gradient flow interpolation does not mode collapse for both m.

trajectory as in Máté & Fleuret (2023). We found in our experiments that the method employed in Máté & Fleuret (2023) led to better performance.

In the following, we evaluate the methods using the effective sample size, negative log likelihood and the energy distance Székely & Rizzo (2013). See D for further details on the implementation and evaluation.

Gaussian Mixture Model. We use the experimental setup from Midgley et al. (2023). The target distribution consists of a mixture of 40 evenly weighted Gaussians in 2 dimensions. The means are distributed uniformly over $[-40, 40]^2$. In Figure 4 we plot the countour lines of the target distribution

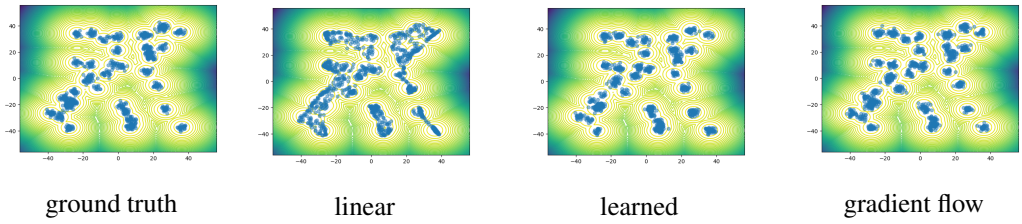


Figure 4: Results for a Gaussian Mixture Model with 40 modes. For the linear and learned interpolation we showed the results for $\sigma = 30$, for the gradient flow interpretation we used $\sigma = 1$.

as well as 1000 samples from the simulated flow and the ground truth, respectively.

For this experiment an important factor is the choice of latent distribution. We restricted ourselves to using a Gaussian latent distribution. However as outlined in Section 3.1 the linear interpolation faces difficulties when some modes are far away from the latent while others are close. In order to mitigate this problem, we experimented with using varying standard deviations σ; the impact of this choice is visualized in D. We found that using a mode covering Gaussian led to the best results. In contrast, the gradient flow interpolation works well when starting in a standard Gaussian which is what we do.

	ESS ↑	NLL ↓	Energy ↓
linear ($\sigma = 30$)	0.183± 0.05	8.657 ± 0.012	0.247± 0.016
learned ($\sigma = 20$)	0.605 ± 0.316	6.894± 0.005	0.058± 0.006
learned ($\sigma = 30$)	0.922± 0.002	6.913 ± 0.003	0.121± 0.009
gradient flow ($\sigma = 1$)	0.992 ± 5e-5	6.864 ± 0.005	0.0021 ± 0.001

Table 1: Comparison of effective sample size (ESS), negative log likelihood (NLL), and the energy distance for different interpolations. Arrows indicate if ascending/descending values are better.

Many Well Distribution. As target distributions we consider 8 and 16-dimensional many well distribution similarly to the implementation of the target energy from Midgley et al. (2023). More precisely let

$$\log p_{m,d}(x) = \sum_{i=1}^{m} \left(-x_i^4 + 6x_i^2 + \frac{1}{2}x_i \right) + \sum_{i=m+1}^{d} -\frac{1}{2}x_i^2 + \text{constant}.$$

The term $-x_i^4 + 6x_i^2 + \frac{1}{2}x_i$ creates two modes with unequal weights and the number of overall modes of $p_{m,d}$ is 2^m. Then our 8 dimensional many well distribution is $p_{4,8}$ and the 16 dimensional target distribution is $p_{4,16}$. The results in Table 2 show that both the learned and gradient flow interpolations significantly outperform the linear interpolation for both dimensions. While the learned and gradient flow interpolations are competitive in $d = 8$, the learned interpolation works better in $d = 16$. The latter behavior might be due to greater flexibility, since both the curve and the vector field are parametrized separately. Here the choice of latent distribution is not as relevant as starting in a standard Gaussian is sufficient for all methods, since the modes of the many well distribution are concentrated close to zero.

	8-dimensional - $p_{4,8}$			16-dimensional - $p_{4,16}$		
Interpolation	ESS ↑	NLL ↓	Energy ↓	ESS ↑	NLL ↓	Energy ↓
Linear	0.985± 0.016	9.689 ± 0.01	0.278 ± 0.003	0.998±3e-5	21.85 ± 0.01	0.201 ± 0.002
Learned	0.999± 1e-5	6.876 ± 0.01	9e-5 ± 3e-5	0.998± 3e-5	18.24± 0.01	1e-4 ± 3e-5
Gradient Flow	0.991± 3e-4	6.989 ± 0.01	8e-4 ± 1e-4	0.544± 0.005	20.53 ± 0.17	0.009 ± 4e-4

Table 2: Comparison of effective sample size (ESS), negative log likelihood (NLL), and energy distance for different interpolations, evaluated for 8-dimensional and 16-dimensional experiments with $m = 4$. Note that the ESS is only meaningful in relation with the Energy.

5 Conclusions

Discussion. We studied different paths in the probability space in order to sample from unnormalized densities and compared their behaviour and computibility. The linear interpolation of energies suffers from an exploding vector field, and learning the vector field *and* energy path simultaneously (*learned interpolation*) in practice solves this problem. However, there is no theoretical knowledge about the path which we want to approximate and in turn no conclusions about the regularity of the corresponding vector field can be drawn. In contrast, the proposed gradient flow approach allows for a thorough mathematical description of the velocity field and the path itself, opening the doors for a rigorous analysis via Wasserstein gradient flows and Fokker-Planck equations. We show numerical performance on sampling from Gaussian mixture models and a many well distribution. Furthermore we close a gap in the literature by showing that a large class of Fisher Rao curves and in particular curves of Boltzmann densities are in fact absolutely continuous curves in the Wasserstein geometry. Hence the existence of a corresponding vector field driving the evolution is implied, justifying sampling from the target density via solving an ODE.

Limitations and Outlook. The methods proposed require knowledge of the domain of the target distribution. Furthermore solving PDEs in high dimensional settings remains challenging. These issues can be addressed by developing more efficient strategies to sample points at which to evaluate the loss $\mathcal{E}(\theta, x, t)$, such as combining uniform sampling and trajectory based sampling. From a theoretical point of view further investigation of the behaviour and regularity of the gradient flow path is needed, especially the behaviour of the norm $\|v_t\|_{L_2(\rho_t)}$ would be of interest.

REFERENCES

R. Adams and J. Fournier. *Sobolev Spaces*. Academic Press, 2003.

T. Akhound-Sadegh, J. Rector-Brooks, A. J. Bose, S. Mittal, P. Lemos, C.-H. Liu, M. Sendera, S. Ravanbakhsh, G. Gidel, Y. Bengio, et al. Iterated denoising energy matching for sampling from Boltzmann densities. *arXiv preprint arXiv:2402.06121*, 2024.

Michael S Albergo and Eric Vanden-Eijnden. Nets: A non-equilibrium transport sampler. *arXiv preprint arXiv:2410.02711*, 2024.

L. Ambrosio and G. Savaré. Gradient flows of probability measures. In *Handbook of differential equations: evolutionary equations*, volume 3, pp. 1–136. Elsevier, 2007.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

D. Bakry, F. Barthe, P. Cattiaux, and A. Guillin. A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability*, 13:60–66, 2008.

J. Berner, L. Richter, and K. Ullrich. An optimal control perspective on diffusion-based generative modeling. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=oYIjw37pTP.

Nicholas M Boffi and Eric Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, 2023.

O. Chehab and A. Korba. A practical diffusion path for sampling, 2024. URL https://arxiv.org/abs/2406.14040.

R. T. Q. Chen. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq.

Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart. Gradient flows for sampling: Mean-field models, Gaussian approximations and affine invariance. *arXiv preprint arXiv:2302.11024*, 2023.

Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart. Efficient, multimodal, and derivative-free Bayesian inference with Fisher-Rao gradient flows, 2024. URL https://arxiv.org/abs/2406.17263.

S. Chewi, J. Niles-Weed, and P. Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

R. N. Dhara and A. Kałamajska. On equivalent conditions for the validity of Poincaré inequality on weighted Sobolev space with applications to the solvability of degenerated PDEs involving p-Laplacian. *Journal of Mathematical Analysis and Applications*, 432(1):463–483, 2015.

R. Duong, V. Stein, R. Beinert, J. Hertrich, and G. Steidl. Wasserstein gradient flows of MMD functionals with distance kernel and Cauchy problems on quantile functions. *arXiv preprint arXiv:2408.07498*, 2024.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouve, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690, 2019.

T. O. Gallouët and L. Monsaingeon. A JKO splitting scheme for Kantorovich–Fisher–Rao gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1100–1130, 2017.

W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Springer, 1995.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:123—-214, 2011.

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

S. Herrmann and N. Massin. Exit problem for Ornstein-Uhlenbeck processes: a random walk approach; to appear in Discrete Contin. *Dyn. Syst. Ser. B, hal-02143409*, 2019.

J. Hertrich and R. Gruhlke. Importance corrected neural JKO sampling. *arXiv preprint arXiv:2407.20444*, 2024.

M. D. Hoffman and A. Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

A. Kufner and B. Opic. How to define reasonably weighted Sobolev spaces. *Commentationes Mathematicae Universitatis Carolinae*, 25(3):537–554, 1984.

R. S. Laugesen, P. G. Mehta, S. P. Meyn, and M. Raginsky. Poisson's equation in nonlinear filtering. *arXiv preprint arXiv:1412.5845*, 2014.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=PqvMRDCJT9t`.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

A. Maurais and Y. Marzouk. Sampling in unit time with kernel Fisher-Rao flow. *arXiv preprint arXiv:2401.03892*, 2024.

L. I. Midgley, V. Stimper, G. N. C. Simm, B. Schölkopf, and J. M. Hernández-Lobato. Flow annealed importance sampling bootstrap. *arXiv preprint arXiv:2208.01893*, 2023.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2006.

B. Máté and F. Fleuret. Learning interpolations between Boltzmann densities. *arXiv preprint arXiv:2301.07388*, 2023.

R. M. Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

F. Noé, S. Olsson, J. Köhler, and H. Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 2019.

N. Nüsken, F. Vargas, S. Padhy, and D. Blessing. Transport meets variational inference: Controlled Monte Carlo diffusions. In *The Twelfth International Conference on Learning Representations: ICLR 2024*, 2024.

N. Nüsken. Stein transport for Bayesian inference. *arXiv preprint arXiv:2409.01464*, 2024. URL `https://arxiv.org/abs/2409.01464`.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: an imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51:235–249, 2011.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2021. URL `https://arxiv.org/abs/2011.13456`.

J. Sun, J. Berner, L. Richter, M. Zeinhofer, J. Müller, K. Azizzadenesheli, and A. Anandkumar. Dynamical measure transport and neural PDE solvers for sampling. *arXiv:2407.07873*, 2024. URL `https://arxiv.org/abs/2407.07873`.

Jingtong Sun, Julius Berner, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural networks for sampling. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*.

G. J. Székely and M. L. Rizzo. Energy statistics: a class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

A. Taghvaei and P. G. Mehta. A survey of feedback particle filter and related controlled interacting particle systems. *Annual Reviews in Control*, 55:356–378, 2023.

F. Vargas, W. S. Grathwohl, and A. Doucet. Denoising diffusion samplers. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=8pvnfTAbu1f`.

Y. Wang and W. Li. Accelerated information gradient flow. *Journal of Scientific Computing*, 90: 1–47, 2022.

H. Wu, J. Köhler, and F. Noé. Stochastic normalizing flows. *Advances in Neural Information Processing Systems*, 33:5933–5944, 2020.

C. Xu, X. Cheng, and Y. Xie. Normalizing flow neural networks by JKO scheme. *Advances in Neural Information Processing Systems*, 36, 2024.

Q. Zhang and Y. Chen. Path integral sampler: A stochastic control approach for sampling. *Proceedings of Machine Learning Research*, 2022.

A Wasserstein meets Fisher-Rao: Solution of a family of Poisson Equations

We will show that under mild assumptions on f, the Fisher-Rao curve ρ_t from (8) is also a Wasserstein absolutely continuous curve with a tangent vector field ∇s_t[2]. In order to do that, we will describe s_t as the unique solution of a Poisson equation in a suitable Hilbert space $\mathcal{X}_0^1$. More precisely, we show that there exists a unique $s \in L_2([0,1] \times \mathbb{R}^d; \boldsymbol{\rho})$ (with $\boldsymbol{\rho}$ given by Definition A.1) such that $\nabla s_t := \nabla s(t, \cdot)$ fulfills the Poisson equation

$$\nabla \cdot (\rho_t \nabla s_t) = \Big(\partial_t f - \mathbb{E}_{\rho_t}\left[\partial_t f\right]\Big)\rho_t \tag{25}$$

in a distributional sense in $\mathcal{X}_0^1$. For fixed time t, this has already been done, e.g., in Laugesen et al. (2014) and Nüsken et al. (2024). But note that a pointwise-in-time treatment as in those papers does not allow us to conclude *Wasserstein absolute continuity* of our curve ρ_t. Therefore, an important contribution is that we generalize their methods to a *global-in-time* ansatz; in particular, under suitable assumptions on f, we find a unique weak solution s of (25) such that

1. $\nabla_x s$ is Borel measurable on $[0,1] \times \mathbb{R}^d$,
2. $t \mapsto \|\nabla_x s(t,\cdot)\|^2_{L_2(\rho_t)} \in L_1([0,1])$, i.e., $\int_0^1 \|\nabla_x s(t,\cdot)\|^2_{L_2(\rho_t)} \,\mathrm{d}t < \infty$,
3. $\nabla_x s(t,\cdot) \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d)$ for a.e. $t \in [0,1]$.

These conditions ensure that the Fisher-Rao curve (8) is absolutely continuous in the Wasserstein sense, and that the velocity field $\nabla_x s$ is the *optimal one* fulfilling the continuity equation w.r.t. our curve ρ_t as described in (4).

As a by-product, we obtain the estimate

$$\int_0^1 \int_{\mathbb{R}^d} \|\nabla_x s(t,\cdot)\|^2 \,\mathrm{d}\rho_t \mathrm{d}t \leq K \int_0^1 \int_{\mathbb{R}^d} \left(\alpha_t - \overline{\alpha_t}\right)^2 \,\mathrm{d}\rho_t \mathrm{d}t,$$

[2]Overall, let $\nabla := \nabla_x$ denote differentiation in direction $x \in \mathbb{R}^d$.

linking (an upper bound of) the Wasserstein-2 distance $W_2^2(\rho_0, \rho_1)$ on the one side, and the Fisher-Rao action on the other.

Next we will define the appropriate spaces for our analysis.

A.1 Prerequisites for solving the Poisson equation

Definition A.1. *Let $t \mapsto \rho_t$ be a weakly continuous curve $[0,1] \to \mathcal{P}^{\mathrm{ac}}(\mathbb{R}^d)$. Then, for all Borel measurable $B \subseteq \mathbb{R}^d$, the map $t \mapsto \rho_t(B)$ is measurable: Indeed, it is the pointwise limit of continuous, hence measurable maps $t \mapsto \int_{\mathbb{R}^d} f_k \,\mathrm{d}\rho_t$ with appropriate f_k. Thus, $\boldsymbol{\rho}$ defined by the disintegration $\int_{[0,1]\times\mathbb{R}^d} f(t,x)\,\mathrm{d}\boldsymbol{\rho} = \int_0^1 \int_{\mathbb{R}^d} f(t,x)\,\mathrm{d}\rho_t \mathrm{d}t$, is a probability measure on $[0,1]\times\mathbb{R}^d$, see e.g., (Ambrosio et al., 2005, Chapter 5.3).*

Definition A.2. *Let $\rho\,\mathrm{d}x \in \mathcal{P}^{\mathrm{ac}}(\mathbb{R}^d)$ with density ρ. We consider the Sobolev space*

$$H^1(\rho) := \Big\{ f \in L_2(\mathbb{R}^d, \rho) : \text{ the weak derivative } \partial_{x_i} f \text{ (w.r.t. } \mathcal{L}) \text{ exists and } \partial_{x_i} f \in L_2(\rho),\ \forall i \in \{1,\dots,d\} \Big\}$$

with inner product $\langle f, g\rangle_{H^1(\rho)} := \langle f, g\rangle_{L_2(\rho)} + \langle \nabla f, \nabla g\rangle_{L_2(\rho,\mathbb{R}^d)}$. Furthermore, define the space of Sobolev functions with zero mean by

$$H_0^1(\rho) := \left\{ f \in H^1(\rho) : \int_{\mathbb{R}^d} f \,\mathrm{d}\rho = 0 \right\}.$$

Lemma A.3. *Assume that $\rho\,\mathrm{d}x \in \mathcal{P}^{\mathrm{ac}}(\mathbb{R}^d)$ and $\frac{1}{\rho} \in L_{1,\mathrm{loc}}(\mathbb{R}^d)$. Then, $(H^1(\rho), \langle\cdot,\cdot\rangle_{H^1(\rho)})$ is a Hilbert space and $H_0^1(\rho)$ is a closed subspace.*

Proof. The first claim is (Kufner & Opic, 1984, Theorem 1.11). For the second claim note that for $f_n \to f$ in $H^1(\rho)$ with $f_n \in H_0^1(\rho)$, we have that

$$\begin{aligned}\left|\int_{\mathbb{R}^d} f \,\mathrm{d}\rho\right| &= |\langle f, 1\rangle_{L_2(\rho)}| \le |\langle f - f_n, 1\rangle_{L_2(\rho)}| + |\langle f_n, 1\rangle_{L_2(\rho)}| \\ &\le \|f_n - f\|_{H^1(\rho)} + 0 \to 0.\end{aligned}$$

□

Definition A.4. *Consider a weakly continuous curve $[0,1] \to \mathcal{P}^{\mathrm{ac}}(\mathbb{R}^d),\ t \mapsto \rho_t$, and the associated measure $\boldsymbol{\rho} \in \mathcal{P}^{\mathrm{ac}}([0,1]\times\mathbb{R}^d)$ given by Definition A.1. Define*

$$H_x^1(\boldsymbol{\rho}) := \Big\{ f \in L_2([0,1]\times\mathbb{R}^d, \boldsymbol{\rho}) : \text{ the weak derivative } \partial_{x_i} f \text{ exists and } \partial_{x_i} f \in L_2(\boldsymbol{\rho}),\ \forall i \in \{2,\dots,d+1\} \Big\}$$

with respect to the inner product $\langle f, g\rangle_{H_x^1(\boldsymbol{\rho})} := \langle f, g\rangle_{L_2(\boldsymbol{\rho})} + \langle \nabla_x f, \nabla_x g\rangle_{L_2(\boldsymbol{\rho})}$, where $f = f(t,x)$, $g = g(t,x)$. Now, consider the subspace

$$\mathcal{X}_0^1 := \Big\{ s \in H_x^1(\boldsymbol{\rho}) : s(t,\cdot) \in H_0^1(\rho_t) \text{ for almost every } t \in [0,1] \Big\}.$$

For $s \in \mathcal{X}_0^1$ we will also write s_t for denoting the function $s(t,\cdot)$.

Lemma A.5. *Assume that $\frac{1}{\boldsymbol{\rho}} \in L_{1,\mathrm{loc}}([0,1]\times\mathbb{R}^d)$. Then, the space $H_x^1(\boldsymbol{\rho})$ is a Hilbert space, and $\mathcal{X}_0^1$ is a closed subspace of $H_x^1(\boldsymbol{\rho})$.*

Proof. The Hilbert space property of $H_x^1(\boldsymbol{\rho})$ again follows by (Kufner & Opic, 1984, Theorem 1.11) and we only prove the closedness of $\mathcal{X}_0^1$. Now, let $(s_n) \subset \mathcal{X}_0^1$ be a sequence with $s_n \to s$ in $H_x^1(\boldsymbol{\rho})$. Then,

$$\int_0^1 \int_{\mathbb{R}^d} |s_n(t,\cdot) - s(t,\cdot)|^2 + \|\nabla_x s_n(t,\cdot) - \nabla_x s(t,\cdot)\|^2 \,\mathrm{d}\rho_t \mathrm{d}t \to 0.$$

This implies pointwise convergence for a.e. $t \in [0,1]$ (along a subsequence), i.e.,

$$\int_{\mathbb{R}^d} |s_n(t,\cdot) - s(t,\cdot)|^2 + \|\nabla_x s_n(t,\cdot) - \nabla_x s(t,\cdot)\|^2 \,\mathrm{d}\rho_t \to 0 \quad \text{for a.e. } t \in [0,1].$$

Thus, we have that $s_n(t,\cdot) \to s(t,\cdot)$ converges in $H^1(\rho_t)$ for almost all $t \in [0,1]$. Since $H_0^1(\rho_t)$ is closed, we have that the limit $s(t,\cdot)$ is in $H_0^1(\rho_t)$ for a.e. t. □

Definition A.6. *On $\mathcal{X}_0^1$ we define an bilinear form by setting*

$$\langle f, g\rangle_{\mathcal{X}_0^1} = \langle \nabla_x f, \nabla_x g\rangle_{L_2(\boldsymbol{\rho})}.$$

Definition A.7. *We say that a probability measure $\boldsymbol{\rho}$ fulfills the partial Poincare inequality* (PPI) *if there exists $K > 0$ such that*

$$\|g\|^2_{L_2(\boldsymbol{\rho})} \leq K \|\nabla_x g\|^2_{L_2(\boldsymbol{\rho})} \tag{PPI}$$

for all $g \in \mathcal{X}_0^1$. Note that usually it is formulated as $Var_\mu[g] \leq K\|\nabla_x g\|^2_{L_2(\mu)}$, but since we have $\mathbb{E}_{\rho_t}[g_t] = 0$ for $g \in \mathcal{X}_0^1$ both notions coincide.

Note that the above definition is more general than the pointwise Poincare inequality used in Nüsken et al. (2024), since we can allow single time points t, where the inequality may fail.

Lemma A.8. *Assume that $\boldsymbol{\rho}$ fulfills the* (PPI)*, and that $\frac{1}{\boldsymbol{\rho}} \in L_{1,\mathrm{loc}}([0,1]\times\mathbb{R}^d)$. Then, $(\mathcal{X}_0^1, \langle\cdot,\cdot\rangle_{\mathcal{X}_0^1})$ is a Hilbert space.*

Proof. We have to show that $\langle\cdot,\cdot\rangle_{\mathcal{X}_0^1}$ is positive definite and $\mathcal{X}_0^1$ is complete. The first claim follows directly from the (PPI). The second follows from the fact that $\|\cdot\|_{H_x^1(\boldsymbol{\rho})} \leq \sqrt{K+1}\|\cdot\|_{\mathcal{X}_0^1} \leq \sqrt{K+1}\|\cdot\|_{H_x^1(\boldsymbol{\rho})}$ and $\mathcal{X}_0^1$ is closed in $H_x^1(\boldsymbol{\rho})$. □

This assumption (PPI) is usually fulfilled in our setting as we discuss in the following.

Lemma A.9. *(Bakry et al., 2008, Corollary 1.6) Let $\mu = e^{-V(x)}\,\mathrm{d}x$ be a probability measure on $\mathbb{R}^d$ and assume that $V \in C^2(\mathbb{R}^d)$ such that V is lower bounded. Assume furthermore that there exists $\eta > 0, \tilde{R} \geq 0$ such that for all $\|x\| > \tilde{R}$ we have*

$$\langle x, \nabla V(x)\rangle \geq \eta\|x\|.$$

Then, there exists a constant $C = C\big(d, \eta, \|V\|_{L^\infty(B(R))}, \|\nabla V\|_{L^\infty(B(R))}\big) > 0$, where $R := \max\{\frac{2d}{\eta}, \tilde{R}\}$, such that for all $\phi \in H_0^1(\mu)$, it holds

$$\int_{\mathbb{R}^d} \phi^2(x)\,\mathrm{d}\mu \leq C \int_{\mathbb{R}^d} \|\nabla\phi\|^2\,\mathrm{d}\mu.$$

The constant C can be estimated as follows: Let $\beta := \beta_{V,R} := 2\|V\|_{L^\infty(B(R))}$. There exist $C_{1,\eta,R}, C_{2,\eta,R}, C_{3,\eta,R} > 0$ (not depending on V) and a universal constant $D > 0$ such that

$$C \leq \frac{4d}{\eta^2}\Big(1 + \big(C_{1,\eta,R} + C_{2,\eta,R}\|\nabla V\|_{L^\infty(B(R))} + C_{3,\eta,R}\big)\, D\, R^2 e^{\beta}\Big).$$

Proof. We first show the claim for all smooth $\phi \in C^\infty(\mathbb{R}^d) \cap H^1(\mu)$ with zero mean. Unfortunately in (Bakry et al., 2008, Corollary 1.6), the constants are not explicitly computed which is why we track them here. They consider for $\gamma = \eta/2$ the function given by $W(x) = W_{\eta,R}(x) := e^{\gamma\|x\|}$ for $\|x\| > R$, and smooth continuation for $\|x\| \leq R$. They show that, *if* for $L := \Delta - \langle\nabla V, \nabla\cdot\rangle$, it holds that

$$LW \leq -\theta W(x) + b\chi_{B(R)},$$

then, the inequality $Var_\mu[\phi] \leq C\int_{\mathbb{R}^d}\|\nabla\phi^2\|\,\mathrm{d}\mu$ is true with constant $C = \frac{1}{\theta}(1 + b\kappa_R)$, where the constant κ_R can be estimated by $D\,R^2 e^{\beta}$. We are left to upper bound the constants θ and b. Let $\|x\| \leq R$. Then, we have that

$$|LW(x)| \leq |\Delta W(x)| + |\langle\nabla V(x), \nabla W(x)| \leq \|\Delta W\|_{L^\infty(B(R))} + \|\nabla V\|_{L^\infty(B(R))}\|\nabla W\|_{L^\infty(B(R))},$$

and for $\|x\| > R$, by choice of R,

$$LW(x) = \frac{\eta}{2}\left(\frac{d-1}{\|x\|} + \frac{\eta}{2} - \frac{\langle x, \nabla V\rangle}{\|x\|}\right) W(x) \leq \frac{\eta}{2}\left(\frac{d-1}{R} + \frac{\eta}{2} - \eta\right) W(x)$$
$$\leq \frac{\eta}{2}\left(\frac{\eta(d-1)}{2d} - \frac{\eta d}{2d}\right) W(x) = -\frac{\eta^2}{4d} W(x),$$

and thus the claim is true for $\phi \in C^\infty(\mathbb{R}^d) \cap H^1(\mu)$ with zero mean, using the constants $\theta = \frac{\eta^2}{4d} > 0$ and $b = C_{1,\eta,R} + C_{2,\eta,R}\|\nabla V\|_{L^\infty(B(R))} + C_{3,\eta,R}$, where $C_{1,\eta,R} := \|\Delta W_{\eta,R}\|_{L^\infty(B(R))}$, $C_{2,\eta,R} := \|\nabla W_{\eta,R}\|_{L^\infty(B(R))}$ and $C_{3,\eta,R} := \theta\|W_{\eta,R}\|_{L^\infty(B(R))}$.
In order to extend this result to $\phi \in H^1_0(\mu)$, note that $C_c^\infty(\mathbb{R}^d)$ is dense in $H^1(\mathbb{R}^d)$, see (Adams & Fournier, 2003, Cor. 3.23). Since by assumption, V is lower bounded, i.e., e^{-V} is upper bounded, $C_c^\infty(\mathbb{R}^d)$ is also dense in $H^1(\mu)$ by Lebesgue's dominated convergence theorem. Now take a sequence $(\phi_n) \in C_c^\infty(\mathbb{R}^d)$ converging to ϕ in $H^1(\mu)$. Then, also $\tilde{\phi_n} := \phi_n - \int_{\mathbb{R}^d} \phi_n \,\mathrm{d}\mu$ converges to ϕ in $H^1(\mu)$, and $\tilde{\phi_n} \in C^\infty(\mathbb{R}^d) \cap H^1(\mu)$ has zero mean. The first part now yields the claim. □

For the specific path ρ_t given by the *linear* potential interpolation $V_t = (1-t)f_0 + tf_1$, the disintegration measure $\boldsymbol{\rho}$ fulfills the partial Poincare inequality (PPI) given suitable assumptions on the end points f_0, f_1, as the next proposition shows.

Proposition A.10. *Let $f_0, f_1 \in C^2(\mathbb{R}^d)$ be lower bounded such that there exists $\alpha, \eta > 0$ and $\tilde{R} > 0$ such that for all $\|x\| \geq \tilde{R}$ it holds the* drift condition

$$\langle x, \nabla f_i(x)\rangle \geq \eta\|x\|, \quad i = 0, 1, \tag{26}$$

and the linear growth condition

$$f_i(x) \geq \alpha\|x\|, \quad i = 0, 1. \tag{27}$$

Let $V_t = (1-t)f_0 + tf_1$ and $\rho_t = Z_t^{-1}e^{-V_t}$ for $Z_t = \int_{\mathbb{R}^d} e^{-V_t}\,\mathrm{d}x$, and $\boldsymbol{\rho}$ as in Definition A.1. Then, $\boldsymbol{\rho}$ satisfies the partial Poincare inequality (PPI)*, i.e., there exists $C > 0$ such that for all $\phi \in \mathcal{X}_0^1$ it holds*

$$\int_{[0,1]\times\mathbb{R}^d} \phi^2 \,\mathrm{d}\boldsymbol{\rho} \leq C \int_{[0,1]\times\mathbb{R}^d} \|\nabla_x \phi\|^2 \,\mathrm{d}\boldsymbol{\rho}.$$

Proof. First note that by the lower boundedness and linear growth assumption (27) on f_0, f_1, the normalization constant Z_t is well-defined, and ρ_t is indeed a probability measure for all $t \in [0,1]$. Further note that $\boldsymbol{\rho}$ exists if ρ_t is weakly continuous, see again Definition A.1. We now show the weak continuity of ρ_t:
Let g be a bounded and continuous function on $\mathbb{R}^d$. By the linear growth (27) of f_0, f_1 and dominated convergence, the map $t \mapsto Z_t = \int_{\mathbb{R}^d} e^{-V_t}\,\mathrm{d}x$ is continuous on $[0,1]$. Since it is also strictly positive on $[0,1]$, also its inverse $t \mapsto Z_t^{-1}$ is continuous on $[0,1]$. By the same token, the map $t \mapsto \int_{\mathbb{R}^d} ge^{-V_t}\,\mathrm{d}x$ is continuous, and altogether, $t \mapsto \rho_t$ is weakly continuous.
In order to prove the inequality, we conclude by (26) that $\langle x, \nabla V_t(x)\rangle \geq (1-t)\eta\|x\| + t\eta\|x\| \geq \eta\|x\|$ for all $\|x\| \geq \tilde{R}$. Let $R := \max\{\frac{2d}{\eta}, \tilde{R}\}$ and note that for $V(t,x) := V_t(x) + \ln(Z_t)$, it is now easy to check that $V(t,\cdot) \in C^2(\mathbb{R}^d)$, $\|\nabla_x V\|_{L^\infty([0,1]\times B(R))} < \infty$ and $\beta := 2\|V\|_{L^\infty([0,1]\times B(R))} < \infty$. Then, by Proposition A.9 we know that

$$\int_{\mathbb{R}^d} \phi(t,\cdot)^2 \,\mathrm{d}\rho_t \leq C \int_{\mathbb{R}^d} \|\nabla_x \phi(t,\cdot)\|^2 \,\mathrm{d}\rho_t$$

for all $\phi \in \mathcal{X}_0^1$ and all $t \in [0,1]$, with a constant C which does *not* depend on t, hence the claim. □

Remark A.11. *(i) Gaussian distributions given by $f_i := \frac{1}{2\sigma^2}\|x-m\|^2$ satisfy the assumptions of Proposition A.10.*
(ii) In Proposition A.10, the regularity assumptions on the end points, i.e., $f_0, f_1 \in C^2(\mathbb{R}^d)$ can be relaxed. It is enough to assume that there exist sequences of functions $(f_{0,n}), (f_{1,n}) \subset C^2(\mathbb{R}^d)$, which converge pointwise to f_0 and f_1, respectively, such that the assumptions of Proposition A.10 are satisfied uniformly in n and such that $\|\nabla_x V_n\|_{L^\infty}, \|V_n\|_{L^\infty}$ are bounded in n.
Hereby, also the Fisher-Rao curve given by linear interpolation of $f_0 := \|x\|$ and $f_1 := \min\{\|x\|, \|x-m\|\}$ with $m \in \mathbb{R}^d$, i.e., the interpolation between two Laplacian (type) distributions, satisfies (PPI).

A.2 Existence of a solution to the Poisson Equation

Now, we consider the Poisson equation (25) in a slightly more abstract manner: Showing that a given Fisher-Rao flow ρ_t is also a Wasserstein absolutely continuous curve requires solving the following (family of) PDEs

$$-\nabla \cdot (\rho_t \nabla s_t) = (\alpha_t - \overline{\alpha_t})\, \rho_t, \quad t \in [0,1], \tag{28}$$

for the vector field $v_t = \nabla s_t$ satisfying a suitable integrability condition. Here, $\alpha \in L_2(\boldsymbol{\rho})$, $\alpha_t := \alpha(t,\cdot)$ and $\overline{\alpha_t} := \mathbb{E}_{\rho_t}[\alpha_t]$ are given.

In order to do that, we look to find a weak solution to (28) in the space $\mathcal{X}_0^1$.

Definition A.12. *An element $\phi \in \mathcal{X}_0^1$ is called a weak solution to* (28) *if it satisfies,*

$$\int_0^1 \int_{\mathbb{R}^d} \langle \nabla \phi_t, \nabla \psi_t \rangle \, \mathrm{d}\rho_t \mathrm{d}t = \int_0^1 \int_{\mathbb{R}^d} \psi_t \, (\alpha_t - \overline{\alpha_t}) \, \mathrm{d}\rho_t \mathrm{d}t, \quad \forall \psi \in \mathcal{X}_0^1. \tag{29}$$

Note that for $\psi \in C_c^\infty((0,1)\times\mathbb{R}^d)$ it holds for $g(t) := \mathbb{E}_{\rho_t}[\psi(t,\cdot)]$ that $\psi - g \in \mathcal{X}_0^1$. Since $\nabla_x(g) = 0 = \int_0^1 \int_{\mathbb{R}^d} g(t)(\alpha_t - \overline{\alpha_t}) \, \mathrm{d}\rho_t \mathrm{d}t$, we have that (29) *is also fulfilled for all $\psi \in C_c^\infty((0,1)\times\mathbb{R}^d)$.*

We can now prove Theorem 1 from Section 2.2, which provides a solution to the Poisson problem (28) globally in time $t \in [0,1]$.

Theorem A.13. *Assume that $\boldsymbol{\rho}$ from Definition A.1 satisfies* (PPI) *for some $K > 0$, and that $\frac{1}{\boldsymbol{\rho}} \in L_{1,\mathrm{loc}}([0,1]\times\mathbb{R}^d)$. Furthermore, assume that $\alpha_t - \overline{\alpha_t} \in L_2(\boldsymbol{\rho})$, i.e., $\|\alpha_t - \overline{\alpha_t}\|^2_{L_2(\boldsymbol{\rho})} \le C < \infty$. Then, there exists a unique weak solution $s \in \mathcal{X}_0^1$ to* (28)*. Moreover, s satisfies the following inequality*

$$\int_0^1 \int_{\mathbb{R}^d} \|\nabla s\|^2 \, \mathrm{d}\rho_t \mathrm{d}t \le K \int_0^1 \int_{\mathbb{R}^d} (\alpha_t - \overline{\alpha_t})^2 \, \mathrm{d}\rho_t \mathrm{d}t. \tag{30}$$

Proof. Let $A : \mathcal{X}_0^1 \to \mathbb{R}$ be defined as

$$A\psi := \int_0^1 \int_{\mathbb{R}^d} \psi_t \, (\alpha_t - \overline{\alpha_t}) \, \mathrm{d}\rho_t \mathrm{d}t.$$

Therefore, we can rewrite our system of PDEs (28) as

$$\langle s, \psi \rangle_{\mathcal{X}_0^1} = A\psi, \quad \forall \psi \in \mathcal{X}_0^1. \tag{31}$$

Clearly, A is a linear operator. Since $\boldsymbol{\rho}$ satisfies (PPI) the following holds:

$$\begin{aligned}
|A\psi|^2 &= \left| \int_0^1 \int_{\mathbb{R}^d} \psi_t \, (\alpha_t - \overline{\alpha_t}) \, \mathrm{d}\rho_t \mathrm{d}t \right|^2 = \left| \int_{[0,1]\times\mathbb{R}^d} \psi_t(x) \, (\alpha_t(x) - \overline{\alpha_t}(x)) \, \mathrm{d}\boldsymbol{\rho}(t,x) \right|^2 \\
&\le \Big(\int (\alpha_t - \overline{\alpha_t})^2 \, \mathrm{d}\boldsymbol{\rho} \Big) \cdot \Big(\int |\psi_t|^2 \, \mathrm{d}\boldsymbol{\rho} \Big) \le C \int_0^1 \int_{\mathbb{R}^d} |\psi_t|^2 \, \mathrm{d}\rho_t \mathrm{d}t \\
&\le C\, K \int_0^1 \int_{\mathbb{R}^d} \|\nabla \psi_t\|^2 \, \mathrm{d}\rho_t \mathrm{d}t = C\, K \|\psi\|^2_{\mathcal{X}_0^1}.
\end{aligned}$$

By Lemma A.8, the space $\mathcal{X}_0^1$ is a Hilbert space. Applying the Riesz Representation theorem, there exists a unique element $s \in \mathcal{X}_0^1$ satisfying (31). □

Remark A.14. *Note that in the case of Fisher-Rao flows, α satisfies*

$$\partial_t \rho_t = (\alpha_t - \overline{\alpha_t}) \, \rho_t, \quad t \in [0,1].$$

Then, the Fisher-Rao action $\mathcal{A}_{\mathrm{FR}}(\rho_t)$ *of ρ_t is given by $\int_0^1 \|\alpha_t - \overline{\alpha_t}\|^2_{L_2(\rho_t)} \, \mathrm{d}t$. Hence, the r.h.s. of* (30) *can be described by $K\, \mathcal{A}_{\mathrm{FR}}(\rho_t)$.*

Lemma A.15. *Let $\boldsymbol{\rho}$ be such that ρ_t is bounded (from above) for a.e. $t \in [0,1]$. Then, for $s \in \mathcal{X}_0^1$ we have that $\nabla s_t \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d)$ for a.e. $t \in [0,1]$.*

Proof. For a.e. t, since ρ_t is bounded from above by assumption, $C_c^\infty(\mathbb{R}^d)$ is dense in $H^1(\rho_t)$. Now, $s_t \in H^1(\rho_t)$ immediately yields the claim $\nabla s_t \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\phi : \phi \in C_c^\infty(\mathbb{R}^d)\}}^{L_2(\rho_t)}$. □

It follows the proof of Theorem 2 from Section 2.2, which says that certain Fisher-Rao curves are indeed also Wasserstein absolutely continuous curves.

Theorem A.16. *Let ρ_0, ρ_1 be probability densities on $\mathbb{R}^d$ and assume that ρ_t is a Fisher-Rao curve defined by*

$$\partial_t \rho_t = (\alpha_t - \overline{\alpha_t})\rho_t,$$

such that $\boldsymbol{\rho}$ and α satisfy the assumptions of Theorem A.13.

Then, ρ_t is a Wasserstein absolutely continuous curve with vector field $\nabla_x s$, where $s \in \mathcal{X}_0^1$ is the unique weak solution of (29)*. If ρ_t is bounded for a.e. t, we have that $\nabla_x s_t \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d)$ for a.e. $t \in [0,1]$, i.e., for every other vector field v_t satisfying the continuity equation for ρ_t, we have that $\|v_t\|_{L_2(\rho_t)} \geq \|\nabla_x s_t\|_{L_2(\rho_t)}$ for a.e. $t \in [0,1]$.*

Furthermore, it holds that

$$\int_0^1 \int_{\mathbb{R}^d} \|\nabla_x s_t\|^2 \,\mathrm{d}\rho_t \mathrm{d}t \leq K \int_0^1 \int_{\mathbb{R}^d} (\alpha_t - \overline{\alpha_t})^2 \,\mathrm{d}\rho_t \mathrm{d}t. \tag{32}$$

Note that by the *Benamou-Brenier formula*, see (Ambrosio et al., 2005, Equation 8.0.3), the left-hand side of (32) is an upper bound for the Wasserstein distance $W_2^2(\rho_0, \rho_1)$, while the double-integral on the right-hand side defines the Fisher-Rao action $\mathcal{A}_{\mathrm{FR}}$ of ρ_t.

Proof. Using $(\alpha_t - \overline{\alpha_t})\rho_t = \partial_t \rho_t$ and Theorem A.13, we can conclude that ρ_t satisfies the continuity equation

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0 \tag{33}$$

with a Borel vector field $v_t := \nabla s_t$ in the sense of distributions. Note that by construction, $s \in \mathcal{X}_0^1$, and hence, the gradient $\nabla_x s$ is already a Borel-measurable vector field on $[0,1] \times \mathbb{R}^d$ such that $\nabla_x s_t \in L_2(\rho_t)$ for a.e. $t \in [0,1]$. Further, (30) implies that $t \mapsto \|\nabla s_t\|_{L_2(\rho_t)} \in L^1([0,T])$.

Together with the fact that ρ_t is weakly continuous, we infer by (Ambrosio et al., 2005, Theorem 8.3.1) that ρ_t is Wasserstein absolutely continuous. By Lemma A.15, it holds that $\nabla s_t \in T_{\rho_t}\mathcal{P}_2(\mathbb{R}^d)$, if ρ_t is bounded for a.e. t. In this case, (Ambrosio et al., 2005, Proposition 8.4.5) immediately implies that ∇s_t has minimal norm $\|\nabla s_t\|_{L_2(\rho_t)}$ among all Borel vector fields v_t fulfilling (33). □

Let us apply the above result to the curve given by linear interpolation of the energy functions f_0, f_1.

Corollary A.17. *Let $f_0, f_1 \in C^2(\mathbb{R}^d)$ be lower bounded by some $\gamma \in \mathbb{R}$ such that the conditions* (26) *and* (27) *of Propostion A.10 are satisfied for some $\alpha, \eta > 0$. Further, for simplicity, assume that f_0, f_1 have at most polynomial growth outside a ball, i.e., there exists $R > 0$ such that*

$$|f_i(x)| \leq M\|x\|^k \quad \textit{for all } \|x\| > R,\ i = 0, 1. \tag{34}$$

Then, Theorem A.16 can be applied to the Fisher-Rao curve ρ_t given by $V_t = (1-t)f_0 + tf_1$ and $\rho_t = Z_t^{-1}e^{-V_t}$ with $Z_t = \int_{\mathbb{R}^d} e^{-V_t}\,\mathrm{d}x$.

Proof. First, by the drift condition (26) and linear growth (27), the measure $\boldsymbol{\rho}$ satisfies (PPI) by Proposition A.10. Furthermore, in Proposition A.10 we showed that $t \mapsto Z_t^{-1}$ is continuous on $[0,1]$, hence $(t,x) \mapsto \rho_t(x) = Z_t^{-1}e^{-V_t(x)}$ is strictly positive and continuous on $[0,1] \times \mathbb{R}^d$. Therefore, it immediately follows that $\frac{1}{\boldsymbol{\rho}} \in L_{1,\mathrm{loc}}([0,1] \times \mathbb{R}^d)$.

We are left to check the L_2-condition on $\alpha_t = -\partial_t V_t = f_0 - f_1$. First, since $t \mapsto Z_t^{-1}$ is continuous on $[0,1]$, hence bounded, it follows together with (27) and (34) that

$$\begin{aligned}
|\overline{\alpha_t}| &\leq Z_t^{-1} \int_{\mathbb{R}^d} |f_0 - f_1| e^{-V_t} \,\mathrm{d}x \\
&\leq Z_t^{-1} \int_{\|x\| \leq R} |f_0 - f_1| e^{-V_t} \,\mathrm{d}x \; + Z_t^{-1} \int_{\|x\| > R} |f_0 - f_1| e^{-V_t} \,\mathrm{d}x \\
&\leq C_1 \int_{\|x\| \leq R} |f_0 - f_1| e^{-\gamma} \,\mathrm{d}x \; + C_2\, M \int_{\|x\| > R} \|x\|^k e^{-\alpha \|x\|} \,\mathrm{d}x \\
&\leq C < \infty,
\end{aligned}$$

with a constant $C > 0$ not depending on t. Hence, $\overline{\alpha_t} \in L_2(\boldsymbol{\rho})$, and we are left to prove $\alpha_t \in L_2(\boldsymbol{\rho})$. But this follows by the similar arguments

$$\begin{aligned}
Z_t^{-1} \int_{\mathbb{R}^d} |f_0 - f_1|^2 e^{-V_t} \,\mathrm{d}x &\leq Z_t^{-1} \int_{\|x\| \leq R} |f_0 - f_1|^2 e^{-V_t} \,\mathrm{d}x \; + Z_t^{-1} \int_{\|x\| > R} |f_0 - f_1|^2 e^{-V_t} \,\mathrm{d}x \\
&\leq C_1 \int_{\|x\| \leq R} |f_0 - f_1|^2 e^{-\gamma} \,\mathrm{d}x \; + C_2\, M^2 \int_{\|x\| > R} \|x\|^{2k} e^{-\alpha \|x\|} \,\mathrm{d}x \\
&\leq C < \infty,
\end{aligned}$$

again, with a constant $C > 0$ not depending t. Integrating over $t \in [0,1]$ yields $\alpha_t \in L_2(\boldsymbol{\rho})$, and hence, all assumptions of Theorem A.16 are fulfilled. □

Remark A.18. *Hence, the Fisher-Rao curves defined by linear interpolation $V_t = (1-t)f_0 + tf_1$ are absolutely continuous in the Wasserstein sense, e.g., in the cases (see also Remark A.11)*

- *Gaussian distributions given by $f_i := \frac{1}{2\sigma^2}\|x - m\|^2$,*
- *Laplacian distributions given by $f_i := \frac{1}{\sigma}\|x - m\|$,*
- *Laplacian type distributions given by $f_i := \min\{\|x\|, \|x - m\|\}$.*

B TELEPORTATION ISSUE WITH THE LINEAR INTERPOLATION

In this section we want to demonstrate that the norm $\|v_t\|_{L_2(\rho_t)}$ of the (optimal) velocity field (4) belonging to the Fisher-Rao path given by linear interpolation $f_t := (1-t)f_0 + tf_1$, explodes for times t close to 1, when the measures ρ_0 and ρ_1 are asymmetrical to each other.

In the following, we restrict ourselves to the one-dimensional setting $\mathbb{R}^1$, in order to exploit the isometry from the Wasserstein space $\mathcal{P}_2(\mathbb{R}^1)$ into the convex cone $\mathcal{C} \subset L_2(0,1)$ of *quantile functions*, also see (Duong et al., 2024, Section 3). For the sake of self-containedness, we give a short introduction of the necessary tools: For $\mu \in \mathcal{P}_2(\mathbb{R})$ we define its cumulative distribution function (CDF) by

$$F_\mu(x) := \mu(-\infty, x], \quad x \in \mathbb{R},$$

and its left-continuous quantile function

$$Q_\mu(s) := \min\{x \in \mathbb{R} : F(x) \geq s\}, \quad s \in (0,1).$$

Then, it holds the isometric property

$$W_2^2(\mu, \nu) = \int_0^1 |Q_\mu(s) - Q_\nu(s)|^2 \,\mathrm{d}s \quad \text{for all } \mu, \nu \in \mathcal{P}_2(\mathbb{R}). \tag{35}$$

Now, by (Ambrosio et al., 2005, Theorem 8.3.1), the norm $\|v_t\|_{L_2(\rho_t)}$ of the optimal field v_t belonging to an absolutely continuous Wasserstein curve ρ_t can be characterized by the *metric derivative* of the curve via

$$\|v_t\|_{L_2(\rho_t)} = |\rho'(t)| := \lim_{s \to t} \frac{W_2(\rho_t, \rho_s)}{|t - s|} \quad \text{for a.e. } t \in [0,1].$$

Using the aforementioned isometry (35) and Komura's theorem, it holds for a.e. $t \in [0,1]$

$$\lim_{s\to t} \frac{W_2(\rho_t,\rho_s)}{|t-s|} = \lim_{s\to t} \frac{\|Q_{\rho_t} - Q_{\rho_s}\|_{L_2(0,1)}}{|t-s|} = \|\partial_t(Q_{\rho_t})\|_{L_2(0,1)}.$$

Hence, if we can explicitly compute the quantiles Q_{ρ_t} (and its time derivative), we get an analytic representation of $\|v_t\|_{L_2(\rho_t)}$ via

$$\|v_t\|^2_{L_2(\rho_t)} = \int_0^1 \Big(\partial_t(Q_{\rho_t})(s)\Big)^2 \,\mathrm{d}s. \tag{36}$$

In order to do so, we choose the prior and target measures the following way: Let $f_0(x) := |x|$ and $f_1(x) := 2\min\{|x|, |x-m|\}$, where the second mode $m > 0$ is fixed for the moment. By Corollary A.17 and Remark A.18, the Fisher-Rao curve ρ_t given by linear interpolation of f_0, f_1 is Wasserstein absolutely continuous. Now, it is straightforward (but tedious) to explicitly calculate the CDF and quantile function of ρ_t:

First, since $\rho_t \propto \rho_0^{1-t}\rho_1^t$ and $2\min\{|x|,|x-m|\} = |x| + |x-m| - \big||x| - |x-m|\big|$, we obtain

$$\rho_t \propto e^{-|x|-t|x-m|+t||x|-|x-m||} \quad \text{for all } x \in \mathbb{R},\ t \in [0,1],$$

up to a normalization constant Z_t which we calculate below.

Calculating the CDF of ρ_t. By dissolving the absolute values on suitable subintervals, we obtain for the CDF

$$\begin{aligned}
Z_t \cdot F_{\rho_t}(x) &= \frac{1}{1+t} e^{x(1+t)}, \quad x \le 0,\\
Z_t \cdot F_{\rho_t}(x) &= \frac{2}{1+t} - \frac{1}{1+t} e^{-x(1+t)}, \quad 0 \le x \le \frac{m}{2},\\
Z_t \cdot F_{\rho_t}(x) &= \frac{2}{1+t} - \Big(\frac{1}{1+t} + \frac{1}{-1+3t}\Big) e^{-\frac{m}{2}(1+t)} + \frac{1}{-1+3t} e^{x(-1+3t)-2tm}, \quad \frac{m}{2} \le x \le m,\\
Z_t \cdot F_{\rho_t}(x) &= \frac{2}{1+t} - \Big(\frac{1}{1+t} + \frac{1}{-1+3t}\Big) e^{-\frac{m}{2}(1+t)}\\
&\quad + \Big(\frac{1}{1+t} + \frac{1}{-1+3t}\Big) e^{-m(1-t)} - \frac{1}{1+t} e^{-x(1+t)+2tm}, \quad x \ge m,
\end{aligned}$$

where the normalization constant is hence given by

$$\begin{aligned}
Z_t = Z_t \cdot F_{\rho_t}(\infty) &= \frac{2}{1+t} - \Big(\frac{1}{1+t} + \frac{1}{-1+3t}\Big) e^{-\frac{m}{2}(1+t)}\\
&\quad + \Big(\frac{1}{1+t} + \frac{1}{-1+3t}\Big) e^{-m(1-t)}.
\end{aligned}$$

Calculating the quantile function of ρ_t. Since $F_{\rho_t}(\cdot)$ is continuous and strictly increasing, its inverse is given by $F_{\rho_t}^{-1} = Q_{\rho_t}$. A simple calculation now yields

$$\begin{aligned}
Q_{\rho_t}(s) &= \frac{1}{1+t} \ln\big(Z_t(1+t)s\big), \quad 0 < s \le s_1,\\
Q_{\rho_t}(s) &= -\frac{1}{1+t} \ln\big(-Z_t(1+t)s + 2\big), \quad s_1 \le s \le s_2,\\
Q_{\rho_t}(s) &= \frac{\ln\big((-1+3t)(sZ_t - \frac{2}{1+t} + (\frac{1}{1+t} + \frac{1}{-1+3t})e^{-\frac{m}{2}(1+t)})\big) + 2tm}{-1+3t}, \quad s_2 \le s \le s_3,\\
Q_{\rho_t}(s) &= \frac{-\ln\big(Z_t(1+t)(1-s)\big) + 2tm}{1+t}, \quad s_3 \le s < 1.
\end{aligned}$$

Here, the subintervals are given by

$$\begin{aligned}
s_1 &= \frac{1}{Z_t(1+t)},\\
s_2 &= \frac{1}{Z_t}\Big(\frac{2}{1+t} - \frac{1}{1+t} e^{-\frac{m}{2}(1+t)}\Big),\\
s_3 &= \frac{1}{Z_t}\Big(\frac{2}{1+t} - (\frac{1}{1+t} + \frac{1}{-1+3t}) e^{-\frac{m}{2}(1+t)} + \frac{1}{-1+3t} e^{-m(1-t)}\Big).
\end{aligned}$$

Calculating the time derivative $\partial_t Q_{\rho_t}$ and the norm $\|v_t\|^2_{L_2(\rho_t)}$. Using standard differentiation rules, the time derivative of Q_{ρ_t} can be computed separately on each subinterval (s_i, s_{i+1}). Since the resulting derivatives become unbearably long, we only include them in the Supplementary Material. Nevertheless, they can be computed analytically as elementary functions, and be inserted into the desired integral (36). By numerical integration, we can finally approximate the norm $\|v_t\|^2_{L_2(\rho_t)}$ arbitrarily close. The computed results are depicted in Figure 2 for different values of the mode $m > 0$; roughly speaking, for late times $t \sim 1$, the norm of the velocity field grows *exponentially* with m.

On a side note, the integration of $\partial_t Q_{\rho_t}$ *only* on the subintervals (s_1, s_2) and (s_2, s_3) contributes to the explosion of $\|v_t\|^2_{L_2(\rho_t)}$. Intuitively, this corresponds to the fact that mass gets shifted very abruptly *mainly* in the areas $x \in (0, \frac{m}{2})$ and $(\frac{m}{2}, m)$ for late times t.

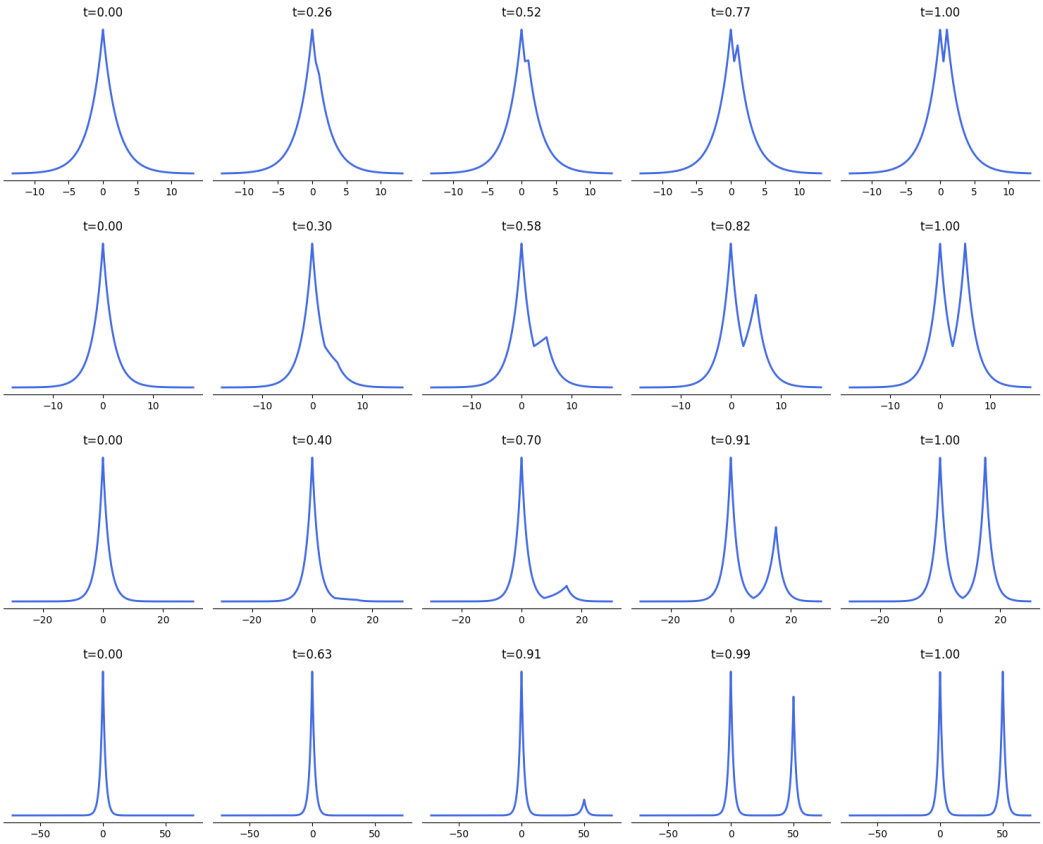


Figure 5: Evolution of probability densities ρ_t for $m \in \{1, 5, 15, 50\}$.

C ALGORITHMS

Algorithm 1 Learning v_t^θ, C_t^θ in (17) for x^i sampled from trajectories.

Parameters and Networks: $\theta = (\theta_1, \theta_2)$ and $v_t^{\theta_1}, C_t^{\theta_2}$.
Functions : $f_t = (1-t)f_0 + tf_1,\ v_t^{\theta_1},\ C_t^{\theta_2}$.
Loss: $\mathcal{E}(\theta, x, t) := |f_1 - f_0 - C_t^{\theta_2} + \langle \nabla f_t, v_t^{\theta_1} \rangle - \nabla \cdot v_t^{\theta_1}|^2$
Algorithm:
for $i = 1, ..., N$ **do**
  grad = 0
  for $l = 1, ..., n$ **do**
    Draw $x^j \sim \rho_Z, 1 \leq j \leq m$
    ODEsolve $x_{t_l}^j$ as solution of $\partial x_t^j = v_t^\theta(x_t^j), x_0^j = x^j$ at time points $t_l \in [0, 1]$
    Detach $x_{t_l}^j$ from computational graph
    grad+= $\nabla_\theta \sum_{i=1}^n \sum_{j=1}^m \mathcal{E}(\theta, x_{t_l}^j, t_l)$
  end for
  Update $v_t^{\theta_1}, C_t^{\theta_2}$ using grad
end for
Output: $v_t^{\theta_1}, C_t^{\theta_2}$.

Algorithm 2 Learning $f_t^{\theta_1}, v_t^{\theta_2}, C_t^{\theta_3}$ in (18) for x^i sampled from trajectories.

Parameters and Networks: $\theta = (\theta_1, \theta_2, \theta_3)$ and $\psi_t^{\theta_1}, v_t^{\theta_2}, C_t^{\theta_3}$.
Functions: $f_t^{\theta_1} = (1-t)f_0 + tf_1 + (1-t)t\psi_t^{\theta_1},\ v_t^{\theta_2},\ C_t^{\theta_3}$.
Loss: $\mathcal{E}(\theta, x, t) := |\partial_t f_t^{\theta_1}(x) - C_t^{\theta_3} + \langle \nabla f_t^{\theta_1}, v_t^{\theta_2} \rangle - \nabla \cdot v_t^{\theta_2}|^2$
Algorithm:
for $i = 1, ..., N$ **do**
  grad = 0
  for $l = 1, ..., n$ **do**
    Draw $x^j \sim \rho_Z, 1 \leq j \leq m$
    ODEsolve $x_{t_l}^j$ as solution of $\partial x_t^j = v_t^\theta(x_t^j), x_0^j = x^j$ at time points $t_l \in [0, 1]$
    Detach $x_{t_l}^j$ from computational graph
    grad+= $\nabla_\theta \sum_{i=1}^n \sum_{j=1}^m \mathcal{E}(\theta, x_{t_l}^j, t_l)$
  end for
  Update $f_t^{\theta_1}, v_t^{\theta_2}, C_t^{\theta_3}$ using grad
end for
Output: $f_t^{\theta_1}, v_t^{\theta_2}, C_t^{\theta_3}$.

Algorithm 3 Learning $\psi_t^{\theta_1}, C_t^{\theta_2}$ in (19) for x^i uniformly sampled.

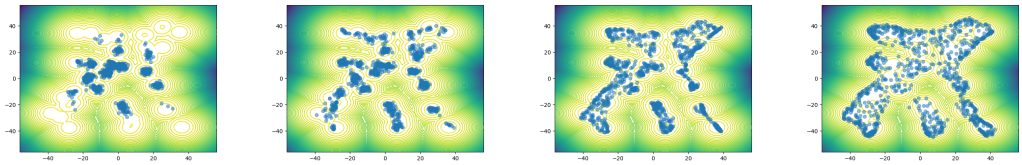
Networks: $\psi_t^{\theta_1}, C_t^{\theta_2}$
Parameters and functions: $\theta = (\theta_1, \theta_2)$ and $f_t = (1-t)f_1 + t\psi_t^{\theta_1},\ C_t^{\theta_2}$.
Loss: $\mathcal{E}(\theta, x, t) := |\partial_t f_t^{\theta_1} - C_t^{\theta_2} + \langle \nabla f_t^{\theta_1}, \nabla(f_t^{\theta_1} - f_Z) \rangle - \Delta(f_t^{\theta_1} - f_Z)|^2$
Algorithm:
for $i = 1, ..., N$ **do**
  Draw $t_i \in [0, 1], 1 \leq i \leq n$ uniformly and $x_{t_i}^{j_k} \in \mathcal{U}_{a,b}, 1 \leq k \leq m$ uniformly
  Compute $\nabla_\theta \sum_{i=1}^n \sum_{j=1}^m \mathcal{E}(\theta, x_{t_i}^{j_k}, t_i)$
  Update θ
end for
Output: $\psi_t^{\theta_1}, C_t^{\theta_2}$.

D ADDITIONAL DETAILS ON THE EXPERIMENTS

Implementation and evaluation details. We implement the algorithms in Pytorch Paszke et al. (2019). For both experiments we use standard MLPs with "swish" activation functions. For the linear and learned interpolations we parameterize the vector fields directly. The learned interpolation also uses the same network for ψ. The linear interpolation uses the same network for the time dependent function C_t. For the learned and gradient flow interpolations we implement C_t using the same network with less parameters. The gradient flow interpolation also uses a smaller network for the time scheduling function g as defined in Section 4. In case of normalized energy functions we can omit the parametrization of C_t for the learned and gradient flow interpolations. We simulate the corresponding ODEs using the torchdiffeq Chen (2018) package with the Runge-Kutta adaptive step size solver ("dopri5"). For the linear and learned interpolation we use 50 time steps along which the loss is computed and the gradients are accumulated with a batch size of 256. For the gradient flow interpolation we sample 4096 particles at random uniform time points and therefore do not accumulate gradients. For the GMM example we train the linear and learned interpolations for $5 * 10^4$ iterations. For the many well example we train the linear interpolation for $5 * 10^4$ and the learned for $4 * 10^4$ iterations. The gradient flow interpolation is trained for $4 * 10^5$ iterations in both experiments.
We compute log weights of the generated samples by using the continuous change of variables formula and the corresponding implementation from Chen (2018). We evaluate the methods by generating $5 * 10^4$ samples and log weight and computing the effective sample size, negative log likelihood and energy distance. We report mean and standard deviation over 10 evaluation runs. The effective sample size is computed as in Midgley et al. (2023). We compute the energy distance using the GeomLoss library Feydy et al. (2019). For the target energy functions and plotting utilities we use the code provided in Midgley et al. (2023).

The gradient flow interpolation faced some difficulty in the many well experiment when, while sampling, particles entered regions of very low density. This could severely affect sampling times when using the adaptive step size solver. Therefore we used an Euler scheme and resampled a given batch if a NaN error is raised. This happens approximately for 0.01% of the particles.

Choice of latent distribution. The linear and learned interpolations depend significantly on the choice of initial distribution. In the following we visualize and compare numerically the results when starting in Gaussians with different standard deviations. Note that when using a small standard deviation not all modes are recovered. On the other hand choosing a very large standard deviation leads to the modes close to 0 not receiving enough mass.

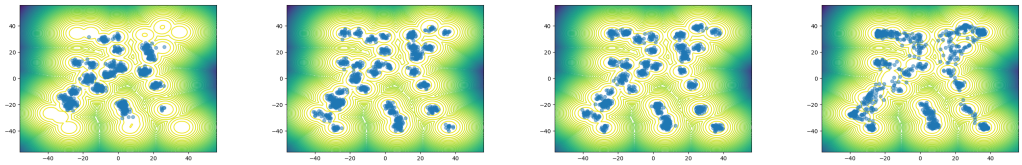


$\sigma = 10$ $\sigma = 20$ $\sigma = 30$ $\sigma = 40$

Figure 6: Results for the linear interpolation for different values of σ.

	ESS ↑	NLL ↓	Energy ↓
$\sigma = 10$	0.002 ± 0.003	7.071 ± 0.005	2.887 ± 0.021
$\sigma = 20$	0.063 ± 0.07	7.362 ± 0.005	0.592 ± 0.005
$\sigma = 30$	0.183± 0.05	8.657 ± 0.012	0.247± 0.016
$\sigma = 40$	0.092 ± 0.001	12.909 ± 0.023	0.554± 0.008

Table 3: Comparison of effective sample size, negative log likelihood, and the energy distance for the linear interpolation with initial distribution being a Gaussian with standard deviation σ.



$\sigma = 10$ $\sigma = 20$ $\sigma = 30$ $\sigma = 40$

Figure 7: Results for the learned interpolation for different values of σ.

	ESS ↑	NLL ↓	Energy ↓
$\sigma = 10$	0.009 ± 0.011	7.275 ± 0.005	1.253 ± 0.017
$\sigma = 20$	0.605 ± 0.316	6.894± 0.005	0.058± 0.006
$\sigma = 30$	0.922± 0.002	6.913 ± 0.003	0.121± 0.009
$\sigma = 40$	0.139 ± 0.052	7.779 ± 0.012	0.531± 0.014

Table 4: Comparison of effective sample size, negative log likelihood, and the energy distance for the learned interpolation with initial distribution being a Gaussian with standard deviation σ.