BUILDING NORMALIZING FLOWS WITH STOCHASTIC INTERPOLANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

A simple generative model based on a continuous-time normalizing flow between any pair of base and target distributions is proposed. The velocity field of this flow is inferred from the probability current of a time-dependent distribution that interpolates between the base and the target in finite time. Unlike conventional normalizing flow inference methods based the maximum likelihood principle, which require costly backpropagation through ODE solvers, our interpolant approach leads to a simple quadratic loss for the velocity itself which is expressed in terms of expectations that are readily amenable to empirical estimation. The flow can be used to generate samples from either the base or target, and can be used to estimate the likelihood at any time along the interpolant. The approach is contextualized in its relation to diffusions. In particular, in situations where the base is a Gaussian distribution, we show that the velocity of our normalizing flow can also be used to construct a diffusion model to sample the target as well as estimating its score. This allows one to map methods based on stochastic differential equations to those of ordinary differential equations, simplifying the mechanics of the model, but capturing equivalent dynamics. Benchmarking on density estimation tasks illustrates that the learned flow can match and surpass maximum likelihood continuous flows at a fraction of the conventional ODE training costs.

1 INTRODUCTION

Contemporary generative models have primarily been designed around the construction of a map between two probability distributions that transform samples from the first into samples from the second. While progress has been from various angles with tools such as implicit maps (Goodfellow et al., 2014; Brock et al., 2019), and autoregressive maps (Menick & Kalchbrenner, 2019; Razavi et al., 2019; Lee et al., 2022), we focus on the case where the map has a clear associated *probability flow*. Advances in this domain, namely from flow and diffusion models, have been governed by the introduction of algorithms or inductive biases that make learning this map, and the Jacobian of the associated change of variables, more tractable. The challenge is to choose what structure to impose on the transport to best reach a complex target distribution from a simple one used as base, while maintaining computational efficiency.

In the continuous time perspective, this problem can be framed as the design of a time-dependent map, $X_t(x)$ with $t \in [0,1]$, which functions as the push-forward of the base distribution at time $t = 0$ onto some time-dependent distribution that reaches the target at time $t = 1$. Assuming that these distributions have densities supported on $\Omega \subseteq \mathbb{R}^d$, say ρ_0 for the base and ρ_1 for the target, this amounts to constructing $X_t : \Omega \to \Omega$ such that

$$\text{if} \quad x \sim \rho_0 \quad \text{then} \quad X_t(x) \sim \rho_t \quad \text{with} \quad \rho_{t=0} = \rho_0 \quad \text{and} \quad \rho_{t=1} = \rho_1 \tag{1}$$

One convenient way to represent this time-continuous map is to define it as the flow associated with the ordinary differential equation (ODE)

$$\dot{X}_t(x) = v_t(X_t(x)), \qquad X_{t=0}(x) = x \tag{2}$$

where $v_t(x)$ is the *velocity field* governing the transport. This is equivalent to saying that the density $\rho_t(x)$ defined as the pushforward of the base ρ_0 by the map X_t satisfies the continuity equation (see e.g. (Villani, 2009; Santambrogio, 2015) and Appendix A)

$$\partial_t \rho_t(x) + \nabla \cdot (v_t(x)\rho_t(x)) = 0 \qquad \text{with} \quad \rho_{t=0} = \rho_0 \quad \text{and} \quad \rho_{t=1} = \rho_1, \tag{3}$$

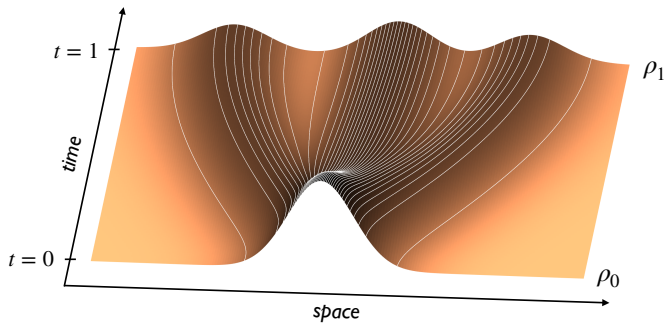

Figure 1: The density $\rho_t(x)$ produced by the stochastic interpolant based on (5) between a standard Gaussian density and a Gaussian mixture density with three modes. Also shows in white are the flow lines of the map $X_t(x)$ our method produces.

and the inference problem becomes to estimate a velocity field such that 3 holds.

In this paper we propose a simple solution to this problem based on introducing a time-differentiable interpolant

$$I_t : \Omega \times \Omega \to \Omega \quad \text{such that} \quad I_{t=0}(x_0, x_1) = x_0 \quad \text{and} \quad I_{t=1}(x_0, x_1) = x_1 \tag{4}$$

A useful instance of such an interpolant that we will employ is

$$I_t(x_0, x_1) = \cos(\tfrac{1}{2}\pi t)x_0 + \sin(\tfrac{1}{2}\pi t)x_1, \tag{5}$$

though we stress the framework we propose applies to any $I_t(x_0, x_1)$ satisfying (4) under mild additional assumptions on ρ_0, ρ_1, and I_t specified below. Given this interpolant, we then construct the stochastic process x_t by sampling independently x_0 from ρ_0 and x_1 from ρ_1 and passing them through I_t:

$$x_t = I_t(x_0, x_1), \qquad x_0 \sim \rho_0, \quad x_1 \sim \rho_1 \tag{6}$$

We refer to the process x_t as a *stochastic interpolant*. Under this paradigm, we make the following key observations as our **main contributions** in this work:

- The probability density $\rho_t(x)$ of x_t, henceforth referred to as the *interpolant density*, satisfies (3) with a velocity $v_t(x)$ which is the unique minimizer of a simple quadratic objective. This result is the content of Proposition 1 below, and it can be leveraged to estimate the velocity $v_t(x)$ in a parametric class (e.g. using deep neural networks) to construct a generative model through the solution of the probability flow equation (2).
- By specifying an interpolant density, the method therefore separates the tasks of minimizing the objective from discovering a path between the base and target densities. This is in contrast with conventional maximum likelihood (MLE) training of flows where one is forced to couple the choice of path in the space of measures to maximizing the objective.
- We show that the Wasserstein-2 (W_2) distance between the target density ρ_1 and the density $\hat{\rho}_1$ obtained by transporting ρ_0 using an approximate velocity $\hat{v}_t$ in (2) is controlled by our objective function. We also show that the value of the objective on $\hat{v}_t$ during training can be used to check convergence of this learned velocity field towards the exact v_t.
- We show that by choosing ρ_0 to be a Gaussian density and using (5) as interpolant, the score of the interpolant density, $\nabla \log \rho_t$, can be explicitly related to the velocity field v_t. This allows us to draw connection between our approach and score-based diffusion models, providing theoretical groundwork for future exploration of this duality.
- We demonstrate the feasibility of the method on toy and high dimensional tabular datasets, and show that the method matches or supersedes conventional ODE flows at lower cost, as it avoids the need to backpropagate through ODE solves.

1.1 Related works

Early works on exploiting transport maps for generative modeling go back at least to Chen & Gopinath (2000), which focuses on normalizing a dataset to infer its likelihood. This idea was brought closer to contemporary use cases through the work of (Tabak & Vanden-Eijnden, 2010;

Tabak & Turner, 2013), which devised to expressly map between densities using simple transport functions inferred through maximum likelihood estimation (MLE). These transformations were learned in sequence via a greedy procedure. We detail below how this paradigm has evolved in the case where the map is represented by a neural network and optimized accordingly.

Discrete and continuous time flows. Instead of working with the continuous map $X_t(x)$ as described in Section 1, the first success of normalizing flows with neural network parametrizations follow the work of (Tabak & Turner, 2013) with a finite set of steps along the map. By imposing structure on the transformation so that it remains an efficiently invertible diffeomorphism, the models of (Rezende & Mohamed, 2015; Dinh et al., 2017; Huang et al., 2018; Durkan et al., 2019) can be optimized through maximum likelihood estimation. However, this comes at the cost of limiting the expressive power of the representation, as the Jacobian of the map must be kept simple to calculate the likelihood. Extending this to the continuous case allowed the Jacobian to be unstructured yet still estimable through trace estimation techniques Chen et al. (2018); Grathwohl et al. (2019); Hutchinson (1989). However learning this map through MLE requires costly backpropagation through numerical integration. The research of (Finlay et al., 2020; Onken et al., 2021) shows that the number of solver calls can be reduced with the addition of some path optimality regularization, though this does not alleviate the main structural challenge of the optimization. Our work uses a continuous map $X_t(x)$ as well but allows for direct estimation of the underlying velocity.

Score-based flows. Adjacent research has made use of diffusion processes, commonly the Ornstein-Uhlenbeck (OU) process, to connect the target ρ_1 to the base ρ_0. In this case the transport is governed by a stochastic differential equation (SDE) that is evolved for infinite time, and the challenge of learning a generative model can be framed as fitting the reverse time evolution of the SDE from Gaussian noise back to ρ_1 (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021b). Doing so indirectly learns the velocity field by means of learning the *score function* $\nabla \log \rho_t(x)$, instead of through MLE. While this approach has shown great promise to model high dimensional distributions (Rombach et al., 2022; Hoogeboom et al., 2022), particularly in the case of text-to-image generation (Ramesh et al., 2022; Saharia et al., 2022), there is an absence of theoretical motivation for the SDE-flow framework and the complexity it induces. Namely, the SDE must evolve for infinite time to connect the distributions, the parameterization of the time steps remains heuristic Feller (1949); Xiao et al. (2022), and the criticality of noise, as well as the score, is not absolutely apparent Bansal et al. (2022); Lu et al. (2022). This motivates further research into effective, ODE-driven, approaches to learning the map. Indeed, under certain conditions, the primary SDE in score-based diffusion can already be associated to its ODE probability flow through a lower bound (Song et al., 2021a). Our approach can be viewed as an alternative to score-based diffusion models in which the ODE velocity is learned through the stochastic interpolant x_t rather than the solution of an OU process.

Interpolants. Data interpolants were recently used in (Choi et al., 2022) for the purposes of density ratio estimation. Here we make use of them to directly discern a velocity field behind a generative model obeying the continuity equation.

1.2 Notations and assumptions

We assume that the base and the target distribution are supported on a nonsingular set $\Omega \subseteq \mathbb{R}^d$ and are both absolutely continuous with respect to the Lesbesgue measure on $\mathbb{R}^d$, with densities ρ_0 and ρ_1, respectively. We do not require these densities to be positive everywhere on Ω, but we assume that $\rho_0(x)$ and $\rho_1(x)$ are continuously differentiable in x. Given any function $f_t(x_0, x_1)$ we denote

$$\mathbb{E}[f_t(x_0, x_1)] = \int_0^1 \int_{\mathbb{R}^d \times \mathbb{R}^d} f_t(x_0, x_1)\rho_0(x_0)\rho_1(x_1)dx_0dx_1dt \tag{7}$$

its expectation over t, x_0, and x_1 drawn independently from the uniform density on $[0, 1]$, ρ_0, and ρ_1, respectively. The gradient $\nabla(\cdot)$ is the gradient with respect to x unless otherwise stated.

Regarding the interpolant $I_t : \Omega \times \Omega$, we assume that it satisfies (4), that $I_t(x_0, x_1)$ is continuously differentiable in (t, x_0, x_1), and that it is such that

$$\mathbb{E}\big[|\partial_t I_t(x_0, x_1)|^2\big] < \infty \tag{8}$$

A few additional technical assumptions on ρ_0, ρ_1 and I_t are listed in Appendix B.

2 STOCHASTIC INTERPOLANTS AND ASSOCIATED FLOWS

Our main theoretical result can be phrased as follows:

Proposition 1 *The stochastic interpolant x_t defined in* (6) *with $I_t(x_0, x_1)$ satisfying* (4) *has a probability density $\rho_t(x)$ which satisfies the continuity equation* (3) *with the velocity field $v_t(x)$ that is the unique minimizer of*

$$G(\hat{v}) = \mathbb{E}\left[|\hat{v}_t(I_t(x_0, x_1))|^2 - 2\partial_t I_t(x_0, x_1) \cdot \hat{v}_t(I_t(x_0, x_1))\right] \tag{9}$$

In addition the minimum value of this objective is given by

$$G(v) = -\mathbb{E}\big[|v_t(I_t(x_0, x_1))|^2\big] = -\int_0^1 \int_{\mathbb{R}^d} |v_t(x)|^2 \rho_t(x) dx dt > -\infty \tag{10}$$

Proposition 1 is proven in Appendix B under Assumption B.1. We now describe some primary facts resulting from this proposition, itemized for clarity:

- The objective $G(\hat{v})$ is given in terms of an expectation that is amenable to empirical estimation given samples t, x_0, and x_1 drawn from ρ_0, ρ_1 and $U([0, 1])$. Below, we will exploit this property to propose a numerical scheme to perform the minimization of $G(\hat{v})$.
- The statement of the proposition remains true if the expectation t is performed using any probability density $\omega(t) > 0$, which may prove useful in practice.
- (10) implies that a necessary (albeit not sufficient) condition for $\hat{v} = v$ is

$$\tilde{G}(\hat{v}) = G(\hat{v}) + \mathbb{E}\big[|\hat{v}_t(I_t(x_0, x_1))|^2\big] = 0. \tag{11}$$

  In our numerical experiments we will monitor this quantity.
- The last bound in (10) implies that the W_2 length of the interpolant density $\rho_t(x)$ is finite.

Let us now provide some intuition about the statements of Proposition 1:

Continuity equation. By definition of the stochastic interpolant x_t we can express its density $\rho_t(x)$ using the Dirac delta distribution as

$$\rho_t(x) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \delta\left(x - I_t(x_0, x_1)\right) \rho_0(x_0) \rho_1(x_1) dx_0 dx_1. \tag{12}$$

Differentiating (12) in time using the chain rule gives

$$\begin{aligned} \partial_t \rho_t(x) &= -\int_{\mathbb{R}^d \times \mathbb{R}^d} \partial_t I_t(x_0, x_1) \cdot \nabla \delta\left(x - I_t(x_0, x_1)\right) \rho_0(x_0) \rho_1(x_1) dx_0 dx_1 \\ &\equiv \nabla \cdot j_t(x) \end{aligned} \tag{13}$$

where we defined the probability current

$$j_t(x) = \int_{\mathbb{R}^d \times \mathbb{R}^d} \partial_t I_t(x_0, x_1) \delta\left(x - I_t(x_0, x_1)\right) \rho_0(x_0) \rho_1(x_1) dx_0 dx_1. \tag{14}$$

Therefore if we introduce the velocity $v_t(x)$ via

$$v_t(x) = \begin{cases} j_t(x)/\rho_t(x) & \text{if } \rho_t(x) > 0, \\ 0 & \text{else} \end{cases} \tag{15}$$

we see that we can write (13) as the continuity equation in (3); since $I_{t=0}(x_0, x_1) = x_0$ and $I_{t=0}(x_0, x_1) = x_1$ by definition, we have $\rho_{t=0} = \rho_0$ and $\rho_{t=1} = \rho_1$, which means that $\rho_t(x)$ also satisfies the boundary conditions at $t = 0, 1$ in (3).

Variational formulation. To derive the objective in (9) for the velocity, consider the alternative objective

$$H(\hat{v}) = \int_0^1 \int_{\mathbb{R}^d} |\hat{v}_t(x) - v_t(x)|^2 \rho_t(x) dx dt = \mathbb{E}[|\hat{v}_t(I_t(x_0, x_1)) - v_t(I_t(x_0, x_1))|^2]. \tag{16}$$

Expanding the square gives

$$\begin{aligned} H(\hat{v}) &= \int_0^1 \int_{\mathbb{R}^d} \left(|\hat{v}_t(x)|^2 - 2\hat{v}_t(x) \cdot v_t(x) + |v_t(x)|^2 \right) \rho_t(x) dx dt \\ &= \int_0^1 \int_{\mathbb{R}^d} \left(|v_t(x)|^2 \rho_t(x) + |\hat{v}_t(x)|^2 \rho_t(x) - 2\hat{v}_t(x) \cdot j_t(x) \right) dx dt \end{aligned} \tag{17}$$

where $j_t(x)$ is the probability current defined in (B.7). Using the expressions in (12) and (14) for $\rho_t(x)$ and $j_t(x)$ we deduce that

$$H(\hat{v}) = G(\hat{v}) + \int_0^1 \int_{\mathbb{R}^d} |v_t(x)|^2 \rho_t(x) dx dt = G(\hat{v}) + \mathbb{E}|v_t(I_t(x_0, x_1))|^2 \tag{18}$$

with $G(\hat{v})$ given in (9). Completing the square in $G(\hat{v})$ we can also write (18) as

$$\begin{aligned} H(\hat{v}) &= \mathbb{E}\big[|\hat{v}_t(I_t(x_0, x_1)) - \partial_t I_t(x_0, x_1)|^2\big] - \mathbb{E}\big[|\partial_t I_t(x_0, x_1)|^2\big] + \mathbb{E}\big[|v_t(I_t(x_0, x_1))|^2\big] \\ &\geq -\mathbb{E}\big[|\partial_t I_t(x_0, x_1)|^2\big] + \mathbb{E}\big[|v_t(I_t(x_0, x_1))|^2\big] \end{aligned} \tag{19}$$

Since we can always find a $\hat{v}_t(x)$ such that $H(\hat{v})$ is finite, using (8) we deduce that $\mathbb{E}[|v_t(I_t(x_0, x_1))|^2] < \infty$. This implies that the minimizer of $H(\hat{v})$ is $v_t(x)$ and $H(v) = 0$. It also implies that the second term at the right hand side of (18) is a finite constant independent on $\hat{v}_t(x)$, which shows that $G(\hat{v})$ and $H(\hat{v})$ have the same minimizer $v_t(x)$.

Minimum value of the the objective. The equality in (10) follows by evaluating (18) at $\hat{v} = v$ using $H(v) = 0$. To get the lower bound in (10) notice that by completing the square in $G(\hat{v})$ we can optimize the inequality in (19) by evaluating it at $\hat{v} = v$ using again $H(v) = 0$, to deduce

$$\mathbb{E}\big[|v_t(I_t(x_0, x_1))|^2\big] \leq \mathbb{E}\big[|\partial_t I_t(x_0, x_1)|^2\big] < \infty \tag{20}$$

which shows that $\mathbb{E}\big[|v_t(I_t(x_0, x_1))|^2\big] = \int_0^1 \int_{\mathbb{R}^d} |v_t(x)|^2 \rho_t(x) dx dt < \infty$ and implies the lower bound in (10). We stress however that the first inequality in (20) is not saturated, i.e. $\mathbb{E}|v_t(I_t(x_0, x_1))|^2 \neq \mathbb{E}|\partial_t I_t(x_0, x_1)|^2$, in general. Hence, we can use $\mathbb{E}|\hat{v}_t(I_t(x_0, x_1)) - \partial_t I_t(x_0, x_1)|^2$ as alternative objective (its first variation in $\hat{v}$ coincides with that of $G(\hat{v})$), but this objective is non-zero at $\hat{v} = v$ in general.

2.1 Wasserstein bounds

The following result shows that the objective in (16) controls the Wasserstein distance between the target density ρ_1 and the the density $\hat{\rho}_1$ obtained as the pushforwrad of the base density ρ_0 by the map $\hat{X}_{t=1}$ associated with the velocity $\hat{v}_t$:

Proposition 2 *Let $\rho_t(x)$ be the exact interpolant density defined in* (B.5) *and, given a velocity field $\hat{v}_t(x)$, let us define $\hat{\rho}_t(x)$ as the solution of the initial value problem*

$$\partial_t \hat{\rho}_t(x) + \nabla \cdot (\hat{v}_t(x) \hat{\rho}_t(x)) = 0, \qquad \hat{\rho}_0(x) = \rho_0(x) \tag{21}$$

Assume that $\hat{v}_t(x)$ is Lipschitz in x uniformily on $(t, x) \in [0, 1] \times \mathbb{R}^d$ with Lipschitz constant $\hat{K}$ Then the square of the W_2 distance between ρ_1 and $\hat{\rho}_1$ is bounded by

$$W_2^2(\rho_1, \hat{\rho}_1) \leq e^{1+2\hat{K}} H(\hat{v}) \tag{22}$$

where $H(\hat{v})$ is the objective function defined in (16)*.*

The proof of Proposition 2 is given in Appendix C: it leverages the following bound on the square of W-2 distance

$$W_2^2(\rho_1, \hat{\rho}_1) \leq \int_0^1 \int_{\mathbb{R}^d} |X_{t=1}(x) - \hat{X}_{t=1}(x)|^2 \rho_0(x) dx dt \tag{23}$$

where X_t is the flow map solution of (2) with the exact $v_t(x)$ defined in (15) and $\hat{X}_t$ is the flow map obtained by solving (2) with $v_t(x)$ replaced by $\hat{v}_t(x)$. Note that the proposition suggests to regularize $G(\hat{v})$ using e.g.

$$G_\lambda(\hat{v}) = G(\hat{v}) + \lambda \int_0^1 \int_{\mathbb{R}^d} \|\nabla \hat{v}_t(x)\|^2 \rho_t(x) dx dt = G(\hat{v}) + \lambda \mathbb{E}\Big[\|\nabla \hat{v}_t(I_t(x_0, x_1)\|^2\Big] \tag{24}$$

with some small $\lambda > 0$. In the numerical results presented below no such regularization was included, and is the subject of future work.

2.2 Link with score-based generative models

The following results shows that for a specific choice of base density ρ_0, the velocity $v_t(x)$ can be related to the score of the density $\rho_t(x)$:

Proposition 3 *Assume that the base density $\rho_0(x)$ is a standard Gaussian density $N(0, Id)$ and suppose that the interpolant $I_t(x_0, x_1)$ is given* (5)*. Then the score $\nabla \log \rho_t(x)$ is related to velocity $v_t(x)$ as*

$$\nabla \log \rho_t(x) = \begin{cases} -x - \dfrac{2}{\pi} \tan(\frac{1}{2}\pi t) v_t(x) & \textit{if } t \in [0, 1) \\ -x - \dfrac{4}{\pi^2} \partial_t v_t(x)|_{t=1} & \textit{if } t = 1. \end{cases} \tag{25}$$

The proof of this proposition is given in Appendix D. The first formula for $t \in [0, 1)$ is based on a direct calculation using Gaussian integration by parts; the second formula at $t = 1$ is obtained by taking the limit of the first using $v_{t=1}(x) = 0$ from (B.13) and l'Hôpital's rule. It shows that we can in principle resample ρ_t at any $t \in [0, 1]$ using the stochastic differential equation in artificial time τ whose drift is the score $\hat{s}_t(x)$ obtained by evaluating (25) on the estimated $\hat{v}_t(x)$:

$$dx_\tau = -\hat{s}_t(x_\tau) d\tau + \sqrt{2} dW_\tau. \tag{26}$$

Similarly, the score $\hat{s}_t(x)$ estimated this way could in principle be used in score-based diffusion models, as explained in Appendix D. We stress however that while our velocity is well-behaved for all times in $[0, 1]$, as shown in (B.13), the drift and diffusion coefficient in the associated SDE are singular at $t = 0, 1$.

3 Practical Implementation and Numerical Experiments

The objective detailed in Section 2 is amenable to efficient empirical estimation, which we utilize to experimentally validate the method. Let $\{x_0^i\}_{i=1}^N$ be N samples from the base density ρ_0, $\{x_1^j\}_{i=1}^n$ n samples from the target density ρ_1, and $\{t_k\}_{k=1}^K$ K samples from the uniform density on $[0, 1]$. Then an empirical estimate of the objective function in (9) is given by

$$G_{N,n,K}(\hat{v}) = \frac{1}{KnN} \sum_{k=1}^K \sum_{j=1}^n \sum_{i=1}^N \left|\hat{v}_{t_k}\big(I_{t_k}(x_0^i, x_1^j)\big)\right|^2 - 2\partial_t I_{t_k}(x_0^i, x_1^j) \cdot \hat{v}_{t_k}(I_{t_k}(x_0^i, x_1^j)). \tag{27}$$

This calculation is fortunately parallelizable. Moreover, it is appealing to consider a neural network parameterization of the velocity field. In this case, the parameters of the model $\hat{v}$ can be optimized through stochastic gradient descent (SGD) or its variants, namely the Adam optimizer (Kingma & Ba, 2015). We take this approach in the following experiments to demonstrate the validity and efficacy of the method using very simple network parameterizations, the purpose of which is to state that sensitive fine tuning is not necessary to demonstrate efficacy.

Following the literature in the advent of new recent density estimation methods, we benchmark the method on visualizable yet complicated densities that display multimodality, as well as higher dimensional tabular data initially provided in (Papamakarios et al., 2017) and tested in other works such as (Grathwohl et al., 2019). We take the 2D test case to demonstrate the ability to flow between empirical densities with *no known analytic* form while also maintaining a benchmark of convergence. In both instances, we compare both the computational cost as well as algorithmic efficiency to the conventional means of training ODE flows and show orders of magnitude speed up.

In all cases, numerical integration for sampling is done with the Dormand–Prince, explicit Runge-Kutta of order (4)5 (Dormand & Prince, 1980). The choice of interpolant for experimentation was universally selected to be that of (5), as it is the one used to draw connections to the technique of score based diffusions in Proposition 3.

3.1 2D DENSITY ESTIMATION

An intuitive first test to benchmark the validity of the method is sampling a target density whose analytic form is known or whose density can be visualized for comparison. To this end, we follow the practice of choosing a few complicated 2-dimensional toy datasets, namely those from (Grathwohl et al., 2019), which were selected to differentiate the flexibility of continuous flows from discrete time flows, which cannot fully separate the modes. We consider anisotropic curved densities, a mixture of 8 separated Gaussians, and a checkerboard density.

The velocity field of the interpolant flow is parameterized by a simple feed forward neural network with ReLU Nair & Hinton (2010) activation functions. The network for each model has 3 layers, each of width equal to 256 hidden units. Adam optimizer is performed on $G(\hat{v})$ for 10k epochs. The results are presented in Figure 2. We plot a kernel density estimate over 80k samples from the flow as well as from the true distribution for comparison. The bounds from (11) are saturated up to noise from estimation of the expectation. The interpolant flow effectively manages to capture all the modes the target density without artificial stretching or smearing, evincing a smooth map.

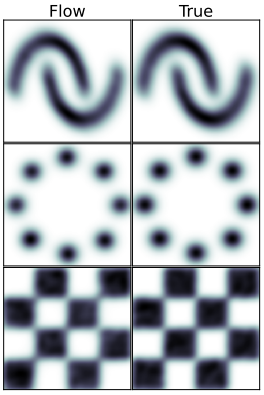

Figure 2: Comparison of density estimation on highly multimodal, structured 2-D distributions.

3.2 DATASET INTERPOLATION

As described in Section 2, the velocity field associated to the flow can be inferred from arbitrary densities ρ_0, ρ_1 – this deviates from the score-based diffusion perspective, in which one distribution must be taken to be Gaussian for the training paradigm to be tractable.

In Figure 3, we illustrate this capacity by learning the velocity field connecting the anisotropic swirls distribution to that of the checkerboard. The interpolant formulation allows us to draw samples from ρ_t at any time $t \in [0, 1]$, which we exploit to check that the velocity field is empirically correct at all times on the interval, rather than just at the end points. This aspect of interpolants is also noted in (Choi et al., 2022), but for the purpose of density ratio estimation. The above observation highlights an *intrinsic difference* of the proposed method compared to MLE training of flows, where the map that is the minimizer of $G(\hat{v})$ is not empirically known. We stress that the evaluation of ρ_0 nor ρ_1 need be possible in the case of this interpolation, though the change of measure between ρ_0 and ρ_1 produced by the model $\hat{v}$ can still be computed as desired.

3.3 TABULAR DATA FOR HIGHER DIMENSIONAL TESTING

A set of tabular datasets introduced by (Papamakarios et al., 2017) has served as a consistent test bed for demonstrating flow-based sampling and its associated density estimation capabilities. We continue that practice here to provide a benchmark of the method on models which provide an

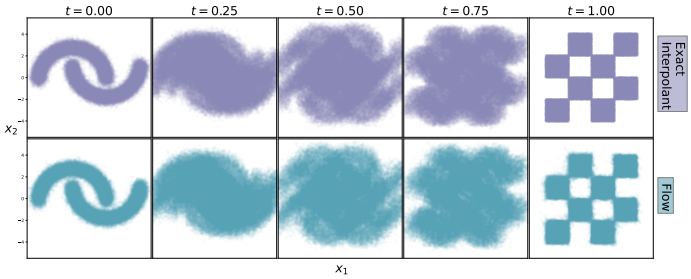

Figure 3: Demonstration of learning a flow map between distributions when neither have analytically known densities. Samples from the intermediate density are still available by the interpolant function $x_t = I_t(x_0, x_1)$, emitting an empirical diagnostic.

	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
MADE	3.08	−3.56	20.98	15.59	−148.85
Real NVP	−0.17	−8.33	18.71	13.55	−153.28
Glow	−0.17	−8.15	18.92	11.35	−155.07
CPF	−0.52	−10.36	16.93	10.58	−154.99
NSP	−0.64	−13.09	14.75	9.67	−157.54
FFJORD	−0.46	−8.59	14.92	**10.43**	**−157.40**
OT-Flow	−0.30	−9.20	17.32	10.55	−154.20
InterFlow (Ours)	**−0.57**	**−12.35**	**14.85**	**10.42**	−156.22

Table 1: Table 2: Negative log-likelihood (NLL) computed on test data unseen during training. (lower is better). Values of MADE, Real NVP, and Glow quoted from the FFJORD paper. Values of OT-Flow, CPF, and NSP quoted from their respective publications. Bold numbers represent best results for continuous flows, with ties also bold if the results are not distinguishable.

exact likelihood, separating and comparing to exemplary discrete and continuous flows: MADE (Germain et al., 2015), Real NVP (Dinh et al., 2017), Convex Potential Flows (CPF) (Huang et al., 2021), Neural Spline Flows (NSP) Durkan et al. (2019), Free-form continuous flows (FFJORD) (Grathwohl et al., 2019), and OT-Flow (Finlay et al., 2020). Our primary point of comparison is to other continuous time models, so we sequester them in benchmarking.

We train the interpolant flow model on each target dataset listed in Table 1, choosing the reference distribution of the interpolant ρ_0 to be a Gaussian with mean $\mu = 0$ and variance $\Sigma = I_d$, where d is the data dimension. The architectures and hyperparameters are given in Appendix E. We highlight some of the main characteristics of the models here. All models are feed-forward ReLU networks, except for the network trained on the BSDS300 dataset, which used ELU activations (Clevert et al., 2016). In each case, sampling of the time t was reweighted according to a Beta distribution, with parameters α, β provided in the same appendix. These were selected to sample times closer to the $t = 1$ endpoint of the transport to ensure that the velocity at times close to the target is well-inferred.

Results from the tabular experiments are displayed in Table 1, in which the negative log-likelihood averaged over a test set of held out data is computed. We note that the interpolant flow achieves better or equivalent held out likelihoods on all ODE based models, except BSDS300, in which the FFJORD outperforms the interpolant by $\sim 0.6\%$. We note upwards of 30% improvements compared to baselines. We observed during model selection that increasing batch size generally improved performance, but this was not consistent across datasets, and the improvements were marginal. For example, a large target batch size proved useful on the POWER dataset, but the same did not hold for the BSDS300 dataset, for which there was no clear correlation between batch size and performance above a minimal threshold. We note that these likelihoods are achieved *without* direct optimization of the quantity.

3.4 Computational efficiency and convergence

In the following, we show that the results achieved in the previous section are driven by a model that can train significantly more efficiently than the maximum likelihood approach to ODEs. We also provide an illustration of the convergence requirements on the objective defined in (11), which helps track that the learned model is meeting necessary conditions of the minimizer of $G(v)$.

Figure 4 shows a comparison of both the cost per training epoch and the convergence of the log likelihood across epochs. We take the architecture of the vector field as defined the FFJORD paper for the 2-dimensional Gaussian case, as well as for the MiniBooNE dataset, and use it to define the vector field for the interpolant flow. For the Gaussian case, this is a 3 layer neural network with hidden widths of 64 units; for the 43-dimensional MiniBooNE target, this is a 3 layer neural network with hidden widths of 860 units. The left side of Figure 4 shows that, overall, the cost per iteration on a batch size of 50 is constant for the interpolant flow, while it grows for MLE based approaches as the ODE gets more complicated to solve. Moreover, the two examples show a 40x and 400x speedup for the interpolant flow as compared to FFJORD, suggesting continued gains in higher dimensions. The right figure shows that, under the same training conditions (optimizer, batch size, and learning rate), the interpolant flow can converge faster in number of training

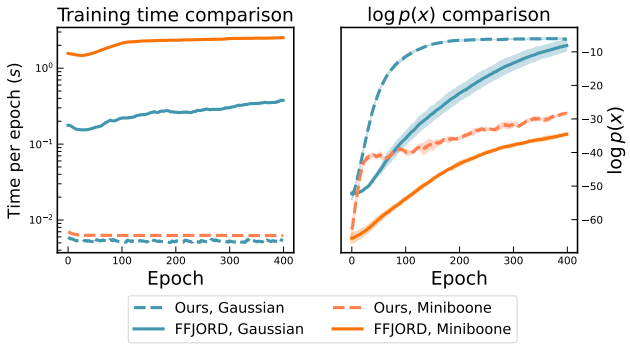

Figure 4: Computational efficiency of proposed method. Left: Training time per epoch for our method versus standard ODE flows on an NVIDIA A100 GPU, showing up to a 400x speedup on the MiniBooNE dataset. Right: log-likelihood evolution per epoch. Our method shows more efficient log likelihood ascent per epoch.

steps, in addition to being cheaper per step. The extent of this benefit is dependent on the set of hyperparameters necessary to achieve good performance, as well as the dataset, so a general statement about convergence speeds is difficult to make. For the sake of this comparison, we averaged over 5 trials for each model and dataset, the variance of which is shown shaded around the curves.

As described in Section 1, the minimizer of $G(v)$ is bounded by the square of the path taken by the map $X_t(x)$. The shifted value of the objective $\tilde{G}(\hat{v})$ in (11) can be tracked to ensure that the model velocity $\hat{v}_t(x)$ meets the requirement of the objective. It must be the case that $\tilde{G}(\hat{v}) = 0$ if $\hat{v}_t(x)$ is taken to be the minimizer of G, so we can look for this signature during the training of the interpolant flow. Figure 5 displays this phenomenon for an interpolant flow trained on the POWER dataset. Here, the shifted loss converges to the target $\tilde{G}(v) = 0$ and remains there throughout training. This suggests that the dynamics of the stochastic optimization of $G(\hat{v})$ are dual to the squared path length of the map.

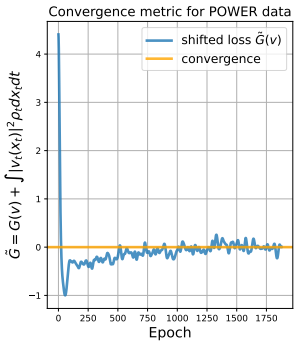

Figure 5: Demonstration of the convergence diagnostic on POWER dataset. Note that this is a necessary but not sufficient condition for convergence. See Section 2 for definition of the shifted loss.

4 Discussion, Challenges, and Future Work

Above, we introduced a continuous time flow method that does not rely on maximum likelihood estimation. The approach has a number of intriguing and appealing characteristics. The training circumvents any backpropagation through ODE solves, and emits a stable and interpretable quadratic objective function. This objective has an easily accessible diagnostic which can verify whether a proposed minimizer of the loss is a valid minimizer.

While our method is computationally more efficient, the new objective introduces an additional sum over time slices, which increases memory requirements. However, as shown in the experiments, these batch sizes need not be prohibitively large when stochastically optimizing $G(v)$. This additional time sum also appears in score-based diffusion approaches and demonstrably poses no issue.

One salient feature of the proposed method is that choosing an interpolant $I_t(x_0, x_1)$ decouples the optimization problem from that of also choosing a transport path, which is implicit in the MLE approach. This separation offers the opportunity of having explicit control on both. In future work, it would be worthwhile to consider learning the interpolant I_t in addition to minimizing $G(\hat{v})$. This could be approached from the perspective of path regularization as suggested by (24) and contextualized in the optimal transport literature (Villani, 2009; Santambrogio, 2015) or through directly parameterizing I_t under the necessary boundary conditions of (4).

Finally, the intrinsic connection to score-based diffusion presented in Proposition 3 may be fruitful ground for understanding the benefits and tradeoffs of SDE vs ODE approaches to generative modeling. Exploring this relation is already underway (Lu et al., 2022; Boffi & Vanden-Eijnden, 2022), and can hopefully provide theoretical insight into designing more effective models.

REFERENCES

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022. URL https://arxiv.org/abs/2208.09392.

Nicholas M. Boffi and Eric Vanden-Eijnden. Probability flow solution of the fokker-planck equation, 2022. URL https://arxiv.org/abs/2206.04642.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Scott Chen and Ramesh Gopinath. Gaussianization. In T. Leen, T. Dietterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000. URL https://proceedings.neurips.cc/paper/2000/file/3c947bc2f7ff007b86a9428b74654de5-Paper.pdf.

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pp. 2552–2573. PMLR, 2022. URL https://proceedings.mlr.press/v151/choi22a.html.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.07289.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HkpbnH9lx.

J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. ISSN 0377-0427. doi: https://doi.org/10.1016/0771-050X(80)90013-3. URL https://www.sciencedirect.com/science/article/pii/0771050X80900133.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. 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. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf.

William Feller. On the theory of stochastic processes, with particular reference to applications. In *Proceedings of the [First] Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1949.

Chris Finlay, Joern-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ODE: the world of Jacobian and kinetic regularization. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3154–3164. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/finlay20a.html.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 881–889, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/germain15.html.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, and David Duvenaud. Scalable reversible generative models with free-form continuous dynamics. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxgknCcK7.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Emiel Hoogeboom, Víctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8867–8887. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/hoogeboom22a.html.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2078–2087. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/huang18d.html.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=te7PVH1sPxJ.

M.F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics - Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806. URL https://doi.org/10.1080/03610918908812806.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. *CoRR*, abs/2203.01941, 2022. doi: 10.48550/arXiv.2203.01941. URL https://doi.org/10.48550/arXiv.2203.01941.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion ODEs by high order denoising score matching. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14429–14460. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/lu22f.html.

Jacob Menick and Nal Kalchbrenner. GENERATING HIGH FIDELITY IMAGES WITH SUBSCALE PIXEL NETWORKS AND MULTIDIMENSIONAL UPSCALING. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HylzTiC5Km.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9223–9232, May 2021. doi: 10.1609/aaai.v35i10.17113. URL https://ojs.aaai.org/index.php/AAAI/article/view/17113.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 2335–2344, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. 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. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/rezende15.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/0a9fdbb17feb6ccb7ec405cfb85222c4-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

E. G. Tabak and Cristina V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. doi: https://doi.org/10.1002/cpa.21423. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21423.

Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217 – 233, 2010. doi: cms/1266935020. URL https://doi.org/.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=JprM0p-q0Co.

A BACKGROUND ON TRANSPORT MAPS AND THE CONTINUITY EQUATION

The following result is standard and can be found e.g. in (Villani, 2009; Santambrogio, 2015)

Proposition A.1 *Let $\rho_t(x)$ satisfy the continuity equation*

$$\partial_t \rho_t(x) + \nabla \cdot (v_t(x)\rho_t(x)) = 0. \tag{A.1}$$

Assume that $v_t(x)$ is C^2 in both t and x for $t \geq 0$ and globally Lipschitz in x. Then, given any $t, t' \geq 0$, the solution of (A.1) *satisfies*

$$\rho_t(x) = \rho_{t'}(X_{t,t'}(x)) \exp\left(-\int_{t'}^{t} \nabla \cdot v_s(X_{t,s}(x))ds\right) \tag{A.2}$$

where $X_{s,t}$ is the probability flow solution to

$$\frac{d}{dt} X_{s,t}(x) = v_t(X_{s,t}(x)), \qquad X_{s,s}(x) = x. \tag{A.3}$$

In addition, given any test function $\phi : \Omega \to \mathbb{R}$, we have

$$\int_\Omega \phi(x)\rho_t(x)dx = \int_\Omega \phi(X_{t',t}(x))\rho_{t'}(x)dx. \tag{A.4}$$

In words, Lemma A.1 states that an evaluation of the PDF ρ_t at a given point x may be obtained by evolving the probability flow equation (2) backwards to some earlier time t' to find the point x' that evolves to x at time t, assuming that $\rho_{t'}(x')$ is available. In particular, for $t' = 0$, we obtain

$$\rho_t(x) = \rho_0(X_{t,0}(x)) \exp\left(-\int_0^t \nabla \cdot v_s(X_{t,s}(x))ds\right), \tag{A.5}$$

and

$$\int_\Omega \phi(x)\rho_t(x)dx = \int_\Omega \phi(X_{0,t}(x))\rho_0(x)dx. \tag{A.6}$$

Proof: The assumed C^2 and globally Lipschitz conditions on v_t guarantee global existence (on $t \geq 0$) and uniqueness of the solution to (2). Differentiating $\rho_t(X_{t',t}(x))$ with respect to t and using (2) and (A.1) we deduce

$$\begin{aligned} \frac{d}{dt}\rho_t(X_{t',t}(x)) &= \partial_t \rho_t(X_{t',t}(x)) + \frac{d}{dt} X_{t',t}(x) \cdot \nabla \rho_t(X_{t',t}(x)) \\ &= \partial_t \rho_t(X_{t',t}(x)) + v_t(X_{t',t}(x)) \cdot \nabla \rho_t(X_{t',t}(x)) \\ &= -\nabla \cdot v_t(X_{t',t}(x))\, \rho_t(X_{t',t}(x)) \end{aligned} \tag{A.7}$$

Integrating this equation in t from $t = t'$ to $t = t$ gives

$$\rho_t(X_{t',t}(x)) = \rho_{t'}(x) \exp\left(-\int_{t'}^{t} \nabla \cdot v_\tau(X_{t',\tau}(x))d\tau\right) \tag{A.8}$$

Evaluating this expression at $x = X_{t,t'}(x)$ and using the group properties (i) $X_{t',t}(X_{t,t'}(x)) = x$ and (ii) $X_{t',\tau}(X_{t,t'}(x)) = X_{t,\tau}(x)$ gives (A.2). Equation (A.4) can be derived by using (A.2) to express $\rho_t(x)$ in the integral at the left hand-side, changing integration variable $x \to X_{t',t}(x)$ and noting that the factor $\exp\left(-\int_{t'}^{t} \nabla \cdot v_\tau(X_{t,\tau}(x))\right)$ is precisely the Jacobian of this change of variable. The result is the integral at the right hand-side of (A.4). □

B PROOF OF PROPOSITION 1

We will work under the following assumption:

Assumption B.1 *The densities $\rho_0(x)$ and $\rho_1(x)$ are continuously differentiable in x, $I_t(x_0, x_1)$ is continuously differentiable in (t, x_0, x_1) and*

$$\begin{aligned} &\int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{ik \cdot I_t(x_0,x_1)} \rho_0(x_0)\rho_1(x_1) dx_0 dx_1 \right| dk < \infty \\ &\int_{\mathbb{R}^d} \left| \int_{\mathbb{R}^d \times \mathbb{R}^d} \partial_t I_t(x_0,x_1) e^{ik \cdot I_t(x_0,x_1)} \rho_0(x_0)\rho_1(x_1) dx_0 dx_1 \right| dk < \infty \end{aligned} \tag{B.1}$$

This structural assumption guarantees that the stochastic interpolant x_t has a probability density and a probability current. It is satisfied e.g. if ρ_0 and ρ_1 are Gaussian mixture densities and

$$I_t(x_0, x_1) = a_t x_0 + b_t x_1, \tag{B.2}$$

where a_t and b_t are C^1 function of $t \in [0, 1]$ satisfying

$$a_0 = 1, \quad a_1 = 0, \quad b_0 = 0, \quad b_1 = 1. \tag{B.3}$$

The interpolant (4) is in this class for the choice

$$a_t = \cos(\tfrac{1}{2}\pi t), \qquad b_t = \sin(\tfrac{1}{2}\pi t) \tag{B.4}$$

Our proof of Proposition 1 will rely on the following result that quantifies the probability density and the probability current of the stochastic interpolant x_t defined in (6):

Proposition B.1 *Assume that I_t satisfies* (4) *and* (8) *and that Assumption B.1 holds. Then the stochastic interpolant x_t defined in* (6) *has a probability density function $\rho_t(x)$ given by*

$$\rho_t(x) = (2\pi)^{-d} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} e^{ik \cdot (x - I_t(x_0,x_1))} \rho_0(x_0)\rho_1(x_1) dx_0 dx_1 dk \tag{B.5}$$

and it satisfies the continuity equation

$$\partial_t \rho_t(x) + \nabla \cdot j_t(x) = 0, \qquad \rho_{t=0}(x) = \rho_0(x), \quad \rho_{t=1}(x) = \rho_1(x) \tag{B.6}$$

withe the probability current $j_t(x)$ given by

$$j_t(x) = (2\pi)^{-d} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \partial_t I_t(x_0,x_1) e^{ik \cdot (x - I_t(x_0,x_1))} \rho_0(x_0)\rho_1(x_1) dx_0 dx_1 dk \tag{B.7}$$

In addition the action of $\rho_t(x)$ and $j_t(x)$ against any test function $\phi : \mathbb{R}^d \to \mathbb{R}$ can be expressed as

$$\int_{\mathbb{R}^d} \phi(x)\rho_t(x) dx = \int_{\mathbb{R}^d \times \mathbb{R}^d} \phi(I_t(x_0,x_1))\rho_0(x_0)\rho_1(x_1) dx_0 dx_1 \tag{B.8}$$

$$\int_{\mathbb{R}^d} \phi(x) j_t(x) dx = \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} \partial_t I_t(x_0,x_1)\phi(I_t(x_0,x_1))\rho_0(x_0)\rho_1(x_1) dx_0 dx_1 \tag{B.9}$$

Note that (B.8) and (B.9) can be formally rewritten as (12) and (14) using the Dirac delta distribution.

Proof: By definition of x_t in (6), the characteristic function of this random variable is

$$\mathbb{E}\big[\exp(ik \cdot x_t)\big] = \int_{\mathbb{R}^d \times \mathbb{R}^d} e^{ik \cdot I_t(x_0,x_1)} \rho_0(x_0)\rho_1(x_1) dx_0 dx_1 \tag{B.10}$$

Under Assumption B.1, the Fourier inversion theorem implies that x_t has a density $\rho_t(x)$ given by (B.5). Taking the time derivative of this density gives

$$\begin{aligned} \partial_t \rho_t(x) &= -(2\pi)^{-d} \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} ik \cdot \partial_t I_t(x_0,x_1) e^{ik \cdot (x - I_t(x_0,x_1))} \rho_0(x_0)\rho_1(x_1) dx_0 dx_1 dk \\ &= -\nabla \cdot j_t(x) \end{aligned} \tag{B.11}$$

with $j_t(x)$ given by (B.7). $\square$

Proposition B.1 shows that $\rho_t(x)$ satisfies the continuity equation (3) with the velocity field defined in (15). For the rest of the proof of Proposition 1 we can then follow the argument in Sec. 2, starting with equation (16)

Finally, notice that while we do not have an explicit expression for $v_t(x)$ when $t \in (0,1)$ in general, we can always evaluate $v_{t=0}(x)$ and $v_{t=1}(x)$. For example, with the choice in (4) for $I_t(x_0, x_1)$, we have $\partial_t I_t(x_0, x_1) = \frac{1}{2}\pi x_1$ at $t = 0$ and $\partial_t I_t(x_0, x_1) = -\frac{1}{2}\pi x_0$ at $t = 1$, and as a result

$$j_{t=0}(x) = \tfrac{1}{2}\pi\rho_0(x)\int_{\mathbb{R}^d} x_1\rho_1(x_1)dx_1, \qquad j_{t=1}(x) = -\tfrac{1}{2}\pi\rho_1(x)\int_{\mathbb{R}^d} x_0\rho_0(x_0)dx_0. \tag{B.12}$$

From (12), this implies

$$v_{t=0}(x) = \tfrac{1}{2}\pi\int_{\mathbb{R}^d} x_1\rho_1(x_1)dx_1, \qquad v_{t=1}(x) = -\tfrac{1}{2}\pi\int_{\mathbb{R}^d} x_0\rho_0(x_0)dx_0. \tag{B.13}$$

C Proof of Proposition 2

To use the bound in (23), let us consider the evolution of

$$Q_t = \int_{\mathbb{R}^d} |X_t(x) - \hat{X}_t(x)|^2\rho_0(x)dx \tag{C.1}$$

Using $\dot{X}_t(x) = v_t(X_t(x))$ and $\dot{\hat{X}}_t(x) = \hat{v}_t(\hat{X}_t(x))$, we deduce

$$\begin{aligned}\dot{Q}_t &= 2\int_{\mathbb{R}^d} (X_t(x) - \hat{X}_t(x))\cdot(v_t(X_t(x)) - \hat{v}_t(\hat{X}_t(x)))\rho_0(x)dx \\ &= 2\int_{\mathbb{R}^d} (X_t(x) - \hat{X}_t(x))\cdot(v_t(X_t(x)) - \hat{v}_t(X_t(x)))\rho_0(x)dx \\ &+ 2\int_{\mathbb{R}^d} (X_t(x) - \hat{X}_t(x))\cdot(\hat{v}_t(X_t(x)) - \hat{v}_t(\hat{X}_t(x)))\rho_0(x)dx\end{aligned} \tag{C.2}$$

Now use

$$2(X_t - \hat{X}_t)\cdot(v_t(X_t) - \hat{v}_t(X_t)) \le |X_t - \hat{X}_t|^2 + |v_t(X_t) - \hat{v}_t(X_t)|^2 \tag{C.3}$$

and

$$2(X_t - \hat{X}_t)\cdot(\hat{v}_t(X_t) - \hat{v}_t(\hat{X}_t)) \le 2\hat{K}|X_t - \hat{X}_t|^2 \tag{C.4}$$

to obtain

$$\dot{Q}_t \le (1 + 2\hat{K})Q_t + \int_{\mathbb{R}^d} |v_t(X_t(x)) - \hat{v}_t(X_t(x))|^2\rho_0(x)dx \tag{C.5}$$

Therefore, by Gronwall's inequality and since $Q_0 = 0$ we deduce

$$Q_1 \le e^{1+2\hat{K}}\int_0^1\int_{\mathbb{R}^d} |v_t(X_t(x)) - \hat{v}_t(X_t(x))|^2\rho_0(x)dxdt = e^{1+2\hat{K}}H(\hat{v}). \tag{C.6}$$

Since $W_2^2(\rho_1, \hat{\rho}_1) \le Q_1$ by (23), we are done. $\square$

D Proof of Proposition 3 and link with score-based diffusion models

Assume that the interpolant is of the type (B.2) so that $\partial_t I_t(x_0, x_1) = \dot{a}_t x_0 + \dot{b}_t x_1$. Assume also that $a_t > 0$ except at $t = 0$ and $b_t > 0$ except at $t = 1$. For $t \in (0,1)$ let us write expression (14) for the probability current as

$$\begin{aligned}j_t(x) &= \int_{\mathbb{R}^d\times\mathbb{R}^d} (\dot{a}_t x_0 + \dot{b}_t x_1)\delta(x - a_t x_0 + b_t x_1)\rho_0(x_0)\rho_1(x_1)dx_0dx_1 \\ &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(\frac{\dot{b}_t}{b_t}(a_t x_0 + b_t x_1) + \Big(\dot{a}_t - \frac{\dot{b}_t}{b_t}a_t\Big)x_0\right)\delta(x - a_t x_0 + b_t x_1)\rho_0(x_0)\rho_1(x_1)dx_0dx_1 \\ &= \frac{\dot{b}_t}{b_t}x\rho_t(x) + \Big(\dot{a}_t - \frac{\dot{b}_t}{b_t}a_t\Big)\int_{\mathbb{R}^d\times\mathbb{R}^d} x_0\delta(x - a_t x_0 + b_t x_1)\rho_0(x_0)\rho_1(x_1)dx_0dx_1\end{aligned} \tag{D.1}$$

If $\rho_0(x_0) = (2\pi)^{-d/2} e^{-\frac{1}{2}|x_0|^2}$, we have the identity $x_0 \rho_0(x_0) = -\nabla_{x_0} \rho_0(x_0)$. Inserting this equality in the last integral in (D.1) and integrating by part using

$$\nabla_{x_0} \delta(x - a_t x_0 + b_t x_1) = -a_t \nabla_x \delta(x - a_t x_0 + b_t x_1) \tag{D.2}$$

gives

$$j_t(x) = \frac{\dot{b}_t}{b_t} x \rho_t(x) - a_t \Big(\dot{a}_t - \frac{\dot{b}_t}{b_t} a_t \Big) \nabla \rho_t(x) \tag{D.3}$$

This means that

$$v_t(x) = \frac{\dot{b}_t}{b_t} x - a_t \Big(\dot{a}_t - \frac{\dot{b}_t}{b_t} a_t \Big) \nabla \log \rho_t(x) \tag{D.4}$$

Solving this expression in $\nabla \log \rho_t(x)$ and specializing it to the trigonometric interpolant with a_t, b_t given in (B.4) gives the first equation in (25). The second one can be obtained by taking the limit of this first equation using $v_{t=1}(x) = 0$ from (B.13) and l'Hôpital's rule. □

Note that (D.3) shows that, when the interpolant is of the type (B.2) and $\rho_0(x_0) = (2\pi)^{-d/2} e^{-\frac{1}{2}|x_0|^2}$, the continuity equation (3) can also be written as the diffusion equation

$$\partial_t \rho_t(x) + \frac{\dot{b}_t}{b_t} \nabla \cdot (x \rho_t(x)) = a_t \Big(\dot{a}_t - \frac{\dot{b}_t}{b_t} a_t \Big) \Delta \rho_t(x) \tag{D.5}$$

If we assume that $\dot{a}_t \leq 0$ and $\dot{b}_t \geq 0$ (which is consistent with the end-point conditions on these factors) the diffusion coefficient in this equation is negative

$$a_t \Big(\dot{a}_t - \frac{\dot{b}_t}{b_t} \Big) \leq 0 \tag{D.6}$$

This means that (D.5) is well-posed backward in time, i.e. it corresponds to backward diffusion from $\rho_{t=1} = \rho_1$ to $\rho_{t=0} = \rho_0 = (2\pi)^{-d/2} e^{-\frac{1}{2}|x_0|^2}$. Therefore, reversing this backward diffusion, similar to what is done in score-based diffusion models, gives an SDE that transforms samples from ρ_0 to ρ_1. Interestingly, these forward and backward diffusion processes arise on the finite time interval $t \in [0, 1]$; notice however that both the drift and the diffusion coefficient are singular at $t = 1$. This is unlike the velocity $v_t(x)$ which is finite at $t = 0, 1$ and is given by

$$v_0(x) = \dot{a}_0 x + \dot{b}_0 \int_{\mathbb{R}^d} x_1 \rho_1(x_1) dx_1, \qquad v_1(x) = \dot{b}_1 x + \dot{a}_1 \int_{\mathbb{R}^d} x_0 \rho_0(x_0) dx_0 \tag{D.7}$$

These expressions reduce to (B.13) when a_t, b_t are given by (B.4) (trigonometric interpolant).

E Implementation Details for Numerical Experiments

	POWER	GAS	HEPMASS	MINIBOONE	BSDS300
Dimension	6	8	21	43	63
# Training point	$1,615,917$	852,174	315,123	29,556	$1,000,000$
Target batch size	800	800	800	800	300
Base batch size	150	150	150	150	200
Time batch size	10	10	20	10	20
Training Steps	10^5	10^5	10^5	10^5	10^5
Learning Rate (LR)	0.003	0.003	0.003	0.003	0.002
LR decay (4k epochs)	0.8	0.8	0.8	0.8	0.8
Hidden layer width	512	512	512	512	1024
# Hidden layers	4	5	4	4	5
Inner activation functions	ReLU	ReLU	ReLU	ReLU	ELU
Beta α, β, time samples	(1.0,0.5)	(1.0,0.5)	(1.0,0.5)	(1.0,1.0)	(1.0,0.7)

Table 2: Hyperparameters and architecture for tabular datasets.

The architectural information and hyperparameters of the models for the resulting likelihoods in Table 1 is presented in Table 2. ReLU (Nair & Hinton, 2010) activations were used throughout,

barring the BSDS300 dataset, where ELU (Clevert et al., 2016) was used. Table formatting based on (Durkan et al., 2019).

In addition, reweighting of the uniform sampling of time values in the empirical calculation of (27) was done using a Beta distribution under the heuristic that the flow should be well trained near the target. This is in line with the statements under Proposition B.1 that any weight $w(t)$ maintains the same minimizer.