Journal of Machine Learning Research 24 (2023) 1-37 Submitted 3/21; Revised 9/22; Published 1/23

Density estimation on low-dimensional manifolds: an inflation-deflation approach

Christian Horvat CHRISTIAN.HORVAT@UNIBE.CH
Jean-Pascal Pfister JEANPASCAL.PFISTER@UNIBE.CH
Department of Physiology
University of Bern
Bern, Switzerland

Editor: Shakir Mohamed

Abstract

Normalizing flows (NFs) are universal density estimators based on neural networks. However, this universality is limited: the density's support needs to be diffeomorphic to a Euclidean space. In this paper, we propose a novel method to overcome this limitation without sacrificing universality. The proposed method inflates the data manifold by adding noise in the normal space, trains an NF on this inflated manifold, and, finally, deflates the learned density. Our main result provides sufficient conditions on the manifold and the specific choice of noise under which the corresponding estimator is exact. Our method has the same computational complexity as NFs and does not require computing an inverse flow. We also demonstrate theoretically (under certain conditions) and empirically (on a wide range of toy examples) that noise in the normal space can be well approximated by Gaussian noise. This allows using our method for approximating arbitrary densities on unknown manifolds provided that the manifold dimension is known.

Keywords: Normalizing flow, density estimation, low-dimensional manifolds, normal space, noise, inflation, deflation, unsupervised learning

1. Introduction

Many modern problems involving high-dimensional data are formulated probabilistically. Key concepts, such as Bayesian classification, denoising, or anomaly detection, rely on the data generating density $p^*(x)$. Therefore, a main research area and of crucial importance is learning this data generating density $p^*(x)$ from samples.

For the case where the corresponding random variable X with values in $\mathbb{R}^D$ takes values on a manifold diffeomorphic to $\mathbb{R}^D$, a normalizing flow (NF) can be used to learn $p^*(x)$ exactly (Huang et al. (2018)). However, in practice, many real-world applications such as predicting protein structures in molecular biology (Hamelryck et al. (2006)), learning motions in robotics (Feiten et al. (2013)), or predicting earthquake patterns in geology (Geller (1997)) are modeled on low-dimensional manifolds, and therefore gave rise to the manifold hypothesis which states that high-dimensional datasets, such as high-resolution images, live close to a low-dimensional manifold (see Fefferman et al. (2016) and the references therein). As a consequence, few attempts have been made to use NFs to learn densities on low-dimensional manifolds, overcoming their topological constraint. To do so, these methods either need to know the manifold beforehand (that is the manifold's chart is given) (Gemici et al. (2016), Rezende et al. (2020), Mathieu and Nickel (2020), Lou

©2023 Christian Horvat and Jean-Pascal Pfister.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/21-0235.html.

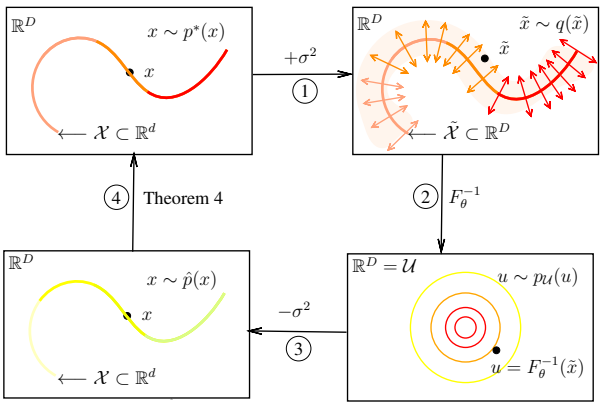

Figure 1: Schematic overview of our method. 1. A density $p^*(x)$ with support on a d-dimensional manifold $\mathcal{X}$ (top left) is inflated by adding noise σ^2 in the normal space (top right). 2. We have an NF $F_\theta^{-1}(x)$ learn this inflated density $q(\tilde{x})$ using a well-known reference measure $p_{\mathcal{U}}(u)$. 3. We deflate the learned density to obtain an estimate $\hat{p}(x)$ for $p^*(x)$. 4. Our main result provides sufficient conditions for the manifold $\mathcal{X}$ and the choice of noise such that $\hat{p}(x) = p^*(x)$.

et al. (2020)), or sacrifice the directness of the estimate (Beitler et al. (2018), Kim et al. (2020), Cunningham et al. (2020), Brehmer and Cranmer (2020)).

Our goal in this paper is to overcome both the aforementioned limitations of using NFs for density estimation on Riemannian manifolds. Given data points from a $d-$dimensional Riemannian manifold denoted as $\mathcal{X}$ embedded in $\mathbb{R}^D$, $d < D$, we first inflate the manifold by adding a specific noise in the normal space direction of the manifold, then train an NF on this inflated manifold, and, finally, deflate the trained density by exploiting the choice of noise and the geometry of the manifold. See figure 1 for a schematic overview of these points. It should be noted that adding Gaussian noise to data is an old trick in machine learning serving various functions. For instance, Vincent et al. (2008) showed that by adding Gaussian noise before training an autoencoder the learned latent representations are more robust to corruption of the inputs. Similarly, dropout introduced by Srivastava et al. (2014) prevents deep neural networks from overfitting by adding noise to their hidden units. More recently, Kim et al. (2020) add noise to smoothen the density and prevent degeneracy problems when training NFs on manifold valued data. However, different from existing approaches, our motivation is to approximate noise restricted to the manifold's normal space and learn the true data generating density $p^*(x)$.

Our main theorem states sufficient conditions on the manifold and the type of noise we use for the inflation step such that the deflation becomes exact. To guarantee the exactness, we do need to know the manifold as in, for example, Rezende et al. (2020) because we need to be able to sample in the manifold's normal space. However, we will show theoretically and empirically that the usual isotropic Gaussian noise serves as a good approximation for a Gaussian restricted to the normal space for a wide range of noise levels. This allows using our method for approximating arbitrary densities on Riemannian manifolds provided that the manifold dimension is known. In addition, our method is based on a single NF without the necessity to invert it. Hence, we don't add any additional complexity to the training procedure of NFs such that autoregressive flows (which are typically D-times slower to invert) can be used. To the best of our knowledge, this is the first

theoretical study that provides sufficient conditions for the learnability of a density with support on a low-dimensional manifold using NFs.

Notations: We denote the Lebesgue measure in $\mathbb{R}^n$ as λ_n. Random variables will be denoted with a capital letter, X, and their corresponding state spaces with the calligraphic version, $\mathcal{X}$. Small letters correspond to vectors with dimensionality given by context. The letters d, D, n, and N are always natural numbers.

2. Background and problem statement

Let X be a random variable that takes values on a $d-$dimensional manifold $\mathcal{X}$ embedded in $\mathbb{R}^D$, that is $\mathcal{X} \subset \mathbb{R}^D$, and let X be generated by an unobserved random variable $U \in \mathcal{U} \subset \mathbb{R}^d$ with density $\pi_u(u)$, where $d < D$. Therefore, from a generative perspective, a sample x from the random variable X is obtained in the following way:

1. sampling from the prior: $u \sim \pi_u(u)$,
2. mapping to the manifold: $x = f(u)$.

If $f : \mathcal{U} \to \mathcal{X}$ is an embedding [1] (as it is the case in Gemici et al. (2016)) the density $p^*(x)$ of X is given by the change of variable formula

$$p^*(x) = |\det G_f(x)|^{-\frac{1}{2}}\, \pi_u(f^{-1}(x)),$$

where we denote the Gram matrix of f evaluated at $f^{-1}(x)$ as

$$G_f(x) := J_f(f^{-1}(x))^T J_f(f^{-1}(x))$$

with J_f^T denoting the transpose of the Jacobian of f. Hence, given an explicit chart f and samples from $p^*(x)$, we can learn the unknown density $\pi_u(u)$ using a standard NF in $\mathbb{R}^d$. However, in general, the generating function f is either unknown or not an embedding creating numerical instabilities for training inputs close to singularity points.

In Brehmer and Cranmer (2020), f and the unknown density π_u are learned simultaneously. Their main idea is to define f as a level set of a usual flow in $\mathbb{R}^D$ and train it together with the flow in $\mathbb{R}^d$ used to learn π_u. To evaluate the density, one needs to calculate $|\det G_f(x)|^{-\frac{1}{2}}$ which computational complexity is $\mathcal{O}(d^2D) + \mathcal{O}(d^3)$. Thus this approach may be slow for high-dimensional data (which we will confirm in section5.3). Besides, to guarantee that f learns the manifold they proposed several ad hoc training strategies. We tie in with the idea to use an NF for learning $p^*(x)$ with unknown f and study the following problem.

Problem 1 *Let $\mathcal{X}$ be a $d-$dimensional manifold embedded in $\mathbb{R}^D$. Let X be a random variable with values in $\mathcal{X}$. Given N samples from $p^*(x)$ as described above, find an estimator $\hat{p}$ of p^* such that in the limit of infinitely many samples we have that $\hat{p}(x) = p^*(x)$, $\mathbb{P}_X$-almost surely.*

The universality of standard NFs: Formally, a standard NF is a diffeomorphism $F_\theta : \mathcal{Z} \subseteq \mathbb{R}^D \to \mathcal{X} \subseteq \mathbb{R}^D$ and induces a density on $\mathcal{X}$ through $p_\theta(x) = |\det G_{F_\theta}(x))|^{-\frac{1}{2}}\, p_{\mathcal{Z}}(F_\theta^{-1}(x))$ where $p_{\mathcal{Z}}$

1. Thus, a regular continuously differentiable mapping (called immersion) which is, restricted to its image, a homeomorphism.

is a known density. The parameters θ are updated such that the KL-divergence between $p^*(x)$ and $p_\theta(x)$,

$$D_{KL}(p^*(x)||p_\theta(x)) = -\mathbb{E}_{x\sim p^*(x)}[\log p_\theta(x)] + const.$$

is minimized. For certain flow architectures, F_θ is expressive enough such that in the limit of infinitely hidden layers n, every $p^*(x)$ with support on $\mathbb{R}^D$ can be learned exactly, see Huang et al. (2018, 2020) for a rigorous mathematical description. However, this universality depends on the architecture and is not true for all flow types, see Zhang et al. (2020).

Remark 2

(i) Note that $p^*(x)$ is uniquely determined by the pair (π, f). For another embedding $f' = f \circ \phi$ with ϕ being a diffeomorphism, the pair (π', f') with $\pi' = \pi \circ \phi^{-1}$ induces the same density $p^*(x)$. Hence, $p^*(x)$ does not depend on the specific embedding.

(ii) The density $p^*(x)$ is with respect to the volume form $dV(x) = \sqrt{|\det G_f(x)|}du$, that is one can calculate probabilities such as $\mathbb{P}_X(A)$ for measurable $A \subset \mathcal{X}$ as follows: $\mathbb{P}(X \in A) = \int_{f^{-1}(A)} \pi_u(u)du = \int_A p^*(x)dV(x)$. Viewing $p^*(x)dV$ as a differential $d-$form, we may say that the volume form dV is induced by the Euclidean metric in $\mathbb{R}^D$.

3. Methods

To solve problem 1, we want to exploit the universality of NFs. We want to inflate $\mathcal{X}$ such that the inflated manifold $\tilde{\mathcal{X}}$ becomes diffeomorphic to a set $\mathcal{U}$ on which a simple density exists. By doing so, this allows us to learn the inflated density $q(\tilde{x}), \tilde{x} \in \mathbb{R}^D$ exactly using a single NF, see section2. Then, given such an estimator for the modified density, we approximate $p^*(x)$ and give sufficient conditions when this approximation is exact.

3.1 The Inflation step

Given a sample x of X, if we add some noise $\mathcal{E} \in \mathbb{R}^D$ to it, the resulting new random variable $\tilde{X} = X + \mathcal{E}$ has the following density

$$q(\tilde{x}) = \int_{\mathcal{X}} q(\tilde{x}|x)d\mathbb{P}_X(x), \tag{1}$$

where $q(\tilde{x}|x)$ is the noise density. Denote the tangent space in x as T_x and the normal space as N_x. By definition, N_x is the orthogonal complement of T_x. Therefore, we can decompose the noise $\mathcal{E}$ into its tangent and normal component, $\mathcal{E} = \mathcal{E}_\mathrm{t} + \mathcal{E}_\mathrm{n}$. In the following, we consider noise in the normal space only, that is $\mathcal{E}_\mathrm{t} = 0$, and denote the density of the resulting random variable as $q_\mathrm{n}(\tilde{x})$. The corresponding noise density $q_\mathrm{n}(\tilde{x}|x)$ has mean x and domain N_x. We denote the support of $q_\mathrm{n}(\cdot|x)$ by $N_{q_\mathrm{n}(\cdot|x)}$. The random variable $\tilde{X} = X + \mathcal{E}_\mathrm{n}$ is now defined on $\widetilde{\mathcal{X}} = \bigcup_{x\in\mathcal{X}} N_{q_\mathrm{n}(\cdot|x)}$. We want $\widetilde{\mathcal{X}}$ to be diffeomorphic to a set $\mathcal{U}$ on which a known density can be defined.

From a generative perspective, a sample $\tilde{x}$ from the random variable $\widetilde{\mathcal{X}}$ is obtained in the following way:

1. sampling from the prior: $u \sim \pi_u(u)$ and $v \sim \pi_v(v)$,
2. mapping to the inflated manifold: $\tilde{x}_\mathrm{n} = x + A_u v$,

where π_v is the noise generating latent density in $\mathbb{R}^{D-d}$, and $A_u \in \mathbb{R}^D \times \mathbb{R}^{D-d}$ is the matrix with columns consisting of normal vectors spanning the normal space in $x = f(u)$. Without loss of generality, we can choose an orthonormal basis for N_x such that $\det A_u^T A_u = 1$.

Example 1

(a) Let $\mathcal{X} = S^1 = \{x \in \mathbb{R}^2 \mid ||x||_2 = 1\}$ be the unit circle where $||\cdot||_2$ denotes the L_2-norm. For each $x \in S^1$ there exists $u \in [0, 2\pi)$ such that $x = e_r(x) = (\cos(u), \sin(u))^T$. To sample a point $\tilde{x}$ in N_x, which is spanned by $e_r(x)$, we sample a scalar value v and set $\tilde{x} = x + ve_r(x)$. With $V \sim \mathrm{Uniform}[-1, 1)$, we have that

$$\widetilde{\mathcal{X}} = \bigcup_{x \in \mathcal{X}} \{x + ve_r(x) | v \in [-1, 1)\} = \{x \in \mathbb{R}^2 \mid ||x||_2 < 2\}$$

which is the open disk with radius 2. The open disk is diffeomorphic to $(0, 1) \times (0, 1)$. Thus, $q_{\mathrm{n}}(\tilde{x})$ can be learned by a single NF denoted as F^{-1} and $p_{\mathcal{Z}}(u) = \mathrm{Uniform}\left((0, 1) \times (0, 1)\right)$ as reference.

(b) As in (a), we consider the unit circle. Now we set V to be a shifted χ^2- distribution with support $[-1, \infty)$. Then,

$$\widetilde{\mathcal{X}} = \bigcup_{x \in \mathcal{X}} \{x + ve_r(x) | v \in [-1, \infty)\} = \mathbb{R}^2.$$

Thus, $q_{\mathrm{n}}(\tilde{x})$ can be learned by a single NF denoted as F^{-1} and $p_{\mathcal{Z}}(z) = \mathcal{N}(z; 0, I_D)$ as reference.

Both cases can be analogously extended to higher dimensions.

3.2 The Deflation step

Equation (1) defines the density of the random variable $\tilde{X} = X + \mathcal{E}$. However, if the noise $\mathcal{E}$ is added in the normal space such that for each realization $\tilde{x}$ there exist only one x, we show that

$$q_{\mathrm{n}}(x) = q_{\mathrm{n}}(x|x)p^*(x). \tag{2}$$

If the estimator $\hat{q}_{\mathrm{n}}(\tilde{x})$ is exact, that is $\hat{q}_{\mathrm{n}}(\tilde{x}) = q_{\mathrm{n}}(\tilde{x})$ for $\mathbb{P}_{\tilde{X}}-$almost all $\tilde{x} \in \widetilde{\mathcal{X}}$, we have for $\tilde{x} = x$ that $p^*(x) = \hat{q}_{\mathrm{n}}(x)/q_{\mathrm{n}}(x|x)$ and therefore $p^*(x)$ can be computed from an NF and a known scaling factor.

For equation (2) to be true, we need to guarantee that almost every $\tilde{x}$ corresponds to only one $x \in \mathcal{X}$. This is certainly the case whenever all the normal spaces have no intersections at all (think of a simple line in $\mathbb{R}^2$). We can relax this assumption by allowing null-set intersections. Moreover, only those subsets of the normal spaces are of interest which are generated by the specific choice of noise $q_{\mathrm{n}}(\tilde{x}|x)$. Thus, only the support of $q_{\mathrm{n}}(\tilde{x}|x)$, denoted by $N_{q_{\mathrm{n}}(\cdot|x)}$, matters. The key concept for our main result is expressed in the following definition:

Definition 3 *Let $\mathcal{X}$ be a $d-$dimensional manifold and N_x the normal space in $x \in \mathcal{X}$. Let $q_{\mathrm{n}}(\cdot|x)$ be a density defined on N_x with support denoted by $N_{q_{\mathrm{n}}(\cdot|x)} \subseteq N_x$. Denote the collection of all such densities as $Q := \{q_{\mathrm{n}}(\cdot|x)\}_{x \in \mathcal{X}}$. For $\tilde{x} \in \widetilde{\mathcal{X}}$, we define the set of all possible generators of $\tilde{x}$ as $\mathcal{A}(\tilde{x}) = \{x' \in \mathcal{X} | N_{q_{\mathrm{n}}(\cdot|x')} \ni \tilde{x}\}$. We say $\mathcal{X}$ is $Q-$**normally reachable** if for all $x \in \mathcal{X}$, it holds that $\mathbb{P}_{\tilde{X}|X=x}\left(\tilde{x} \in N_x | \#\mathcal{A}(\tilde{x}) > 1\right) = 0$ where $\#\mathcal{A}(\tilde{x})$ is the cardinality of the set $\mathcal{A}(\tilde{x})$. In other words, every $\tilde{x} \in N_x$ is $\mathbb{P}_{\tilde{X}|X=x}$-almost surely determined by x.*

To familiarize with this concept, consider figure 2 and the following example:

Example 2 *For the circle in example 1, we chose $\mathcal{E}_n$ to be uniformly distributed on the half-open interval $[-1, 1)$. The point $(0,0)^T$ is contained in $N_{q_n(\cdot|x)}$ for all $x \in \mathcal{X}$, and thus $N_{q_n(\cdot|x')} \cap N_{q_n(\cdot|x)} = \{(0,0)^T\}$ for all $x \neq x'$, see figure 2 (middle). Hence, for any given $\tilde{x} \in N_x$ we have that $\mathcal{A}(\tilde{x}) = \mathcal{X}$ if $\tilde{x} = (0,0)^T$ and $\mathcal{A}(\tilde{x}) = x$ otherwise. Therefore, $\#\mathcal{A}(\tilde{x}) = \infty$ if $\tilde{x} = (0,0)^T$ and $\#\mathcal{A}(\tilde{x}) = 1$ else.*
Thus, $\mathbb{P}_{\tilde{X}|X=x}\left[\tilde{x} \in \tilde{\mathcal{X}} | \#\mathcal{A}(\tilde{x}) > 1\right] = \mathbb{P}_{\tilde{X}|X=x}\left[\tilde{x} = (0,0)^T\right] = 0$ for all $x \in \mathcal{X}$. What follows is that $\mathcal{X}$ is $Q-$normally reachable.

If we were to choose $\mathcal{E}_n$ to be uniformly distributed on $[-1.5, 1)$, see figure 2 (right), the normal spaces would overlap and we would have that $\mathbb{P}_{\tilde{X}|X=x}\left[\tilde{x} \in \tilde{\mathcal{X}} | \#\mathcal{A}(\tilde{x}) > 1\right] > 0$. In this case, $\mathcal{X}$ would not be $Q-$normally reachable.

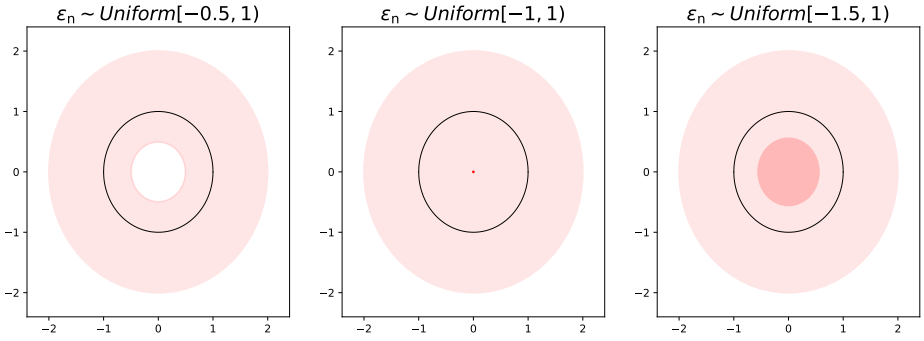

Figure 2: Q-normal reachability for different noise distributions $q_n(\tilde{x}|x)$ used to inflate $\mathcal{X} = S^1$ (black line). **Left:** $\mathcal{X}$ is Q-normally reachable since every point in the inflated space $\widetilde{\mathcal{X}}$ (red shaded area) has a unique generator. **Middle:** $\mathcal{X}$ is Q-normally reachable since $\mathbb{P}_{\tilde{X}}-$almost every point in $\widetilde{\mathcal{X}}$ has a unique generator. **Right:** $\mathcal{X}$ is not Q-normally reachable since every point in the dark shaded area has two generators. Note that the pink area denotes the inflated manifold and not the density.

From a generative perspective, $Q-$normal reachability ensures that the mapping

$$\begin{aligned} \tilde{f} : \mathcal{U} \times \mathcal{V} &\mapsto \widetilde{\mathcal{X}} \\ (u, v) &\mapsto f(u) + A_u v \end{aligned}$$

is bijective (up to a set of measure 0). As f is an embedding by assumption, $\tilde{f}$ is even a diffeomorphism if $||v||_2$ is sufficiently small, as we will show in theorem 4. This, together with the assumption that latent distribution is factorized, that is $\pi(u, v) = \pi_u(u)\pi_v(v)$, implies that the density $q_n(\tilde{x})$ is given by

$$q_n(\tilde{x}) = \left|\det G_{\tilde{f}}(\tilde{x})\right|^{-\frac{1}{2}} \pi_u(u)\pi_v(v), \tag{3}$$

where $(u, v) = \tilde{f}^{-1}(\tilde{x})$. When setting $v = 0$, we have that $\tilde{x} = x$ and indeed equation (2) holds as we will show that $\left|\det G_{\tilde{f}}(x)\right| = |\det G_f(x)|$ and $\pi_v(0) = q_n(x|x)$. Note that our flow of

arguments do not require the manifold $\mathcal{X}$ to be generated by a single chart f. Hence, as long as the manifold is $Q-$normal reachable, equation (3) holds locally for any chart f.

Theorem 4 *Let $\mathcal{X}$ be a $d-$dimensional, C^2 manifold. For each $x \in \mathcal{X}$, let $q_{\mathrm{n}}(\cdot|x)$ denote a continuous distribution with support $N_{q(\cdot|x)}$ in the normal space of x, that is $N_{q(\cdot|x)} \subseteq N_x$, such that $x \in N_{q(\cdot|x)}$. Further, assume that the latent distribution of the inflated random variable $\tilde{X} = X + \mathcal{E}_{\mathrm{n}}$ is factorized. If $\mathcal{X}$ is $Q-$normally reachable where $Q := \{q_{\mathrm{n}}(\cdot|x)\}_{x\in\mathcal{X}}$, then for all $x \in \mathcal{X}$ it holds that $q_{\mathrm{n}}(x) = p^*(x)q_{\mathrm{n}}(x|x)$, thus*

$$p^*(x) = \frac{q_{\mathrm{n}}(x)}{q_{\mathrm{n}}(x|x)}. \tag{4}$$

The proof can be found in appendix A.1. As a consequence of theorem 4 , if the density $q_{\mathrm{n}}(\tilde{x})$ can be learned exactly using a single NF (which is the case whenever the inflated space $\tilde{\mathcal{X}}$ is diffeomorphic to $\mathbb{R}^D$ and the NF is sufficiently expressive), the true density $p^*(x)$ can be retrieved exactly.

Proposition 5 *With the assumptions from theorem 4 , if the inflation is such that $\widetilde{\mathcal{X}}$ is diffeomorphic to $\mathbb{R}^D$, then $q_{\mathrm{n}}(\tilde{x})$ can be learned exactly using a single NF denoted as F, that is $q_{\mathrm{n}}(\tilde{x}) = (\det G_F(F^{-1}(\tilde{x})))^{-\frac{1}{2}}\mathcal{N}(F^{-1}(\tilde{x}); 0, 1)$. Then, using equation (4) the true density $p^*(x)$ can be calculated exactly.*

Remark 6 *It is important to note that the density q_{n} is with respect to the Euclidean metric because this ensures that we can learn it using an NF. If we consider the density with respect to the product metric on $\widetilde{\mathcal{X}}$ denoted as $q_{\mathrm{n}}^{\otimes}$, we can't use a standard NF to learn it. However, we prove in the appendix A.4 that with the assumptions of theorem 4 , we have that $q_{\mathrm{n}}^{\otimes}(\tilde{x}) = p^*(x)q_{\mathrm{n}}(\tilde{x}|x)$ which is based on the fact that $\widetilde{\mathcal{X}}$ isomorphic to $\bigcup_{x\in\mathcal{X}} \left(\{x\} \times N_{q_{\mathrm{n}}(\cdot|x)}\right)$ up to set of measure 0.*

3.3 Gaussian noise as normal noise and the choice of σ^2

Our proposed method depends on some critical points. First, we need to be able to sample in the normal space of $\mathcal{X}$, and we need to determine the magnitude and type of noise. Second, we need to make sure that the conditions of theorem 4 are fulfilled. We address (partially) those points.

1. The inflation must not garble the manifold too much. For instance, adding Gaussian noise with magnitude $\sigma \geq r$ to S^1 will blur the circle. Since the curvature of the circle is $1/r$, intuitively, we want σ to scale with the second derivative of the generating function f. Additionally, we do not want to lose the information of $p^*(x)$ by inflating the manifold. If the generating distribution $\pi(z)$ makes a sharp transition at z_0, $\pi(z_0 - \Delta z_o) \ll \pi(z_0 + \Delta z_o)$ for $|\Delta z_o| \ll 1$, adding too much noise in $x_0 = f(z_0)$ will smooth out that transition. Hence, we want σ to inversely scale with $\pi''(z)$. We formalize these intuitions in proposition 7 and prove it in appendix A.5. We denote the inflated density with Gaussian noise by $q_\sigma(\tilde{x})$ in the following.

Proposition 7 *Let $X \in \mathbb{R}^D$ be generated by $U \sim \pi_u(u)$ through an embedding $f : \mathbb{R}^d \to \mathbb{R}^D$, that is $f(U) = X$. Let $\pi_u \in C^2(\mathbb{R}^d)$. For $q_\sigma(\tilde{x})$ to approximate well $p^*(x)$, in the sense that*[2]

2. Note that $q_{\mathrm{n}}(x|x)$ also depends on σ.

$q_\sigma(x)/q_{\mathrm{n}}(x|x) \approx p^*(x)$ for $x \in \mathcal{X}$, a necessary condition is that:

$$\sigma^2 \ll \frac{2\pi_u(u_0)}{||\pi_u''(u_0) \odot (J_f^T(u_0)J_f(u_0))^{-1}||_+},$$

where $||A||_+ = \left|\sum_{i,j=1}^d A_{ij}\right|$ for $A \in \mathbb{R}^{d\times d}$, and $\odot$ denotes the elementwise product, and $(\pi''(u_0))_{ij} = \frac{\partial^2 \pi_u(u)}{\partial u_i \partial u_j}|_{u=u_0}$ is the Hessian of π_u evaluated at $u_0 = f^{-1}(x)$. Note that in the limit of small noise variance, we have $\lim_{\sigma^2\to 0} q_\sigma(x)/q_{\mathrm{n}}(x|x) = p^*(x)$.

Intuitively, a second necessary condition is that the noise magnitude should be much smaller than the radius of the curvature of the manifold which directly depends on the second-order derivatives of f. This can be illustrated in the following example:

Example 3 *For the circle in $\mathbb{R}^2$ generated by $f(u) = (\cos(u), \sin(u))^T$ and a von Mises distribution $\pi_u(u) \propto \exp(\kappa\cos(u))$, we get that $\sigma^2 \ll \min\left(\left|\frac{2}{\kappa(\kappa\sin^2(u)-\cos(u))}\right|, 1\right)$ where the first condition comes from proposition 7 and the second one comes from the curvature argument. Note that, if $\kappa \leq \frac{\sqrt{7}}{2}$, we have that $\kappa(\kappa\sin^2(u) - \cos(u)) \leq 2$ for all $u \in [-\pi, \pi]$ and thus $\sigma^2 \ll 1$ in that case. Otherwise, the minimum depends on u. However, we can calculate a uniform upper bound by taking the minimum over all u. In that case, we have that $\sigma^2 \ll \frac{2}{0.25+\kappa^2}$ if $\kappa > \frac{\sqrt{7}}{2}$.*

Even though this bound may not be useful as such in practice when f and π_u are unknown, it can still be used if f and π_u are estimated locally with nearest neighbor statistics. Also, proposition 7 together with theorem 4 tells us that for sufficiently small noise variance σ^2, the standard Gaussian noise can well be used to approximate a Gaussian restricted in the normal space.

From a numerical perspective, inflating a manifold using Gaussian noise circumvents degeneracy problems when training a vanilla NF on low-dimensional manifolds. More precisely, those degeneracy problems occur when the condition number of the flow's Jacobian is too high, see Belsley et al. (2005). Since the condition number is simply the ratio of the largest singular value to the lowest singular value, it is clear that in the limit of small normal noise, the smallest singular value will tend to zero and therefore the condition number will tend to infinity. As a consequence, setting up a maximal condition number will automatically provide a lower bound on the inflation noise. In the appendix A.2, we approximate the condition number of the Jacobian from which we derive a lower bound for the inflation noise.

2. Intuitively, if the curvature of the manifold is not too high and if the manifold is not too entangled, $Q-$normal reachability is satisfied for a sufficiently small magnitude of noise. In the manifold learning literature, the entanglement can be measured by the reach number. Informally, the reach number provides a necessary condition on the manifold such that it is learnable through samples, see chapter 2.3 in Berenfeld and Hoffmann (2019).

Formally, the reach number is the maximum distance $\tau_{\mathcal{X}}$ such that for all $\tilde{x}$ in a $\tau_{\mathcal{X}}-$neighborhood of $\mathcal{X}$, its nearest point on $\mathcal{X}$ is unique. In appendix A.6 we prove theorem 8 which states that any closed manifold $\mathcal{X}$ with $\tau_{\mathcal{X}} > 0$ is $Q-$normally reachable.

Theorem 8 *Let $\mathcal{X} \subset \mathbb{R}^D$ be a closed d-dimensional manifold. If $\mathcal{X}$ has a positive reach number $\tau_{\mathcal{X}}$, then $\mathcal{X}$ is $Q-$normally reachable where $Q := \{q_{\mathrm{n}}(\cdot|x)\}_{x\in\mathcal{X}}$ is the collection of uniform distributions on a ball with radius $\tau_{\mathcal{X}}$, that is $q_{\mathrm{n}}(\tilde{x}|x) = \mathrm{Uniform}(B(x, \tau_{\mathcal{X}}) \cap N_x)$ where*

$B(x, \tau_{\mathcal{X}}) = \{y \in \mathbb{R}^D, s.t.\ ||x - y||_2 < \tau_{\mathcal{X}}\}$ *denotes the* $D-$*dimensional ball with radius* $\tau_{\mathcal{X}}$ *and center* x*.*

To appreciate theorem 8, we refer to the tubular neighborhood theorem, which states that every smooth and compact manifold has positive reach (see Lee (2019) for a proof).

4. Related work

Here, we give an overview of methods based on NFs for density estimation on low-dimensional manifolds. One direction of research concentrates on densities defined on a given manifold, such as spheres, tori or hyperboloids (Rezende and Mohamed, 2015,Rezende et al., 2020, Mathieu and Nickel, 2020, Lou et al., 2020). Orthogonal to that direction, Brehmer and Cranmer, 2020, Beitler et al., 2018, Kim et al., 2020, Cunningham et al., 2020 do not rely on an explicit chart while focusing on improving the generative ability. From the latter works, only Brehmer and Cranmer (2020) learn, in theory, the density on the manifold $p^*(x)$ exactly.

Cunningham et al., 2020 assume that data live on a noisy, that is inflated manifold and propose to learn a stochastic inverse $q(z|\tilde{x})$ of the generator $q(\tilde{x}|z)$. To train the parameters of $q(\tilde{x}|z)$, they rely on variational inference making this approach a special case of a variational autoencoder. Their injective noisy flow improves the sampling quality compared to a baseline NF and, in addition, learns a latent representation. However, by construction, they only learn the inflated distribution $q(\tilde{x})$.

Kim et al., 2020 follow our methodology closely by inflating the manifold so that a usual NF can be used to learn the inflated density. For each sample x, they first draw a value c uniformly on $[0, 0.1]$, and then add a sample ν from $\mathcal{N}(0, c^2 I_D)$ to x, that is $\tilde{x} = x + \nu$. They learn the conditional distribution of the inflated manifold, $q(\tilde{x}|c)$, allowing for sampling on the manifold by setting $c = 0$. Their method does not require any knowledge of the manifold (neither the chart, nor the dimensionality), and improve 3D point cloud generation. However, they don't provide a deflation of the inflated distribution, and thus don't learn $p^*(x)$ exactly.

Beitler et al., 2018 propose to use different reference measures for the flow to encode the relevant manifold and irrelevant off-manifold directions. They propose to model the first d latent variables, say u, of the flow as standard Gaussian and the remaining $D - d$ variables, say v, as a diagonal Gaussian with small variance. The hope is that maximum likelihood training is sufficient to encode the manifold in the first d components, so that a sampling procedure where the remaining $D - d$ components are set to 0, that is $v = 0$, would produce samples on the manifold. The gist is very similar to our idea expressed in equation (2). However, in general, this does not lead to the right density on the manifold, as explained in a footnote on page 4 in Brehmer and Cranmer, 2020, which justifies the name Pseudo-invertible encoder (PIE). Nevertheless, as noted by Brehmer and Cranmer, 2020, it is surprising that "somehow in practice learning dynamics and the inductive bias of the model seem to couple in a way that favor an alignment of the level set $v = 0$ with the data manifold. Understanding these dynamics better would be an interesting research goal." Our work gives a theoretical explanation of why the PIE-model favors that alignment: When adding noise with small magnitude to the dataset (for instance, dequantization for images), the resulting density can be well approximated by a product of $p^*(x)$ and the noise distribution $q(\tilde{x}|x)$, such that treating the latent variables u and v differently, and thus having a product of two different measures as reference measure, biases the flow to learn this product form. A further interesting future direction would be

to make this bias more explicit by constructing a flow for which the Jacobian determinant is in such a product form as well.

In Brehmer and Cranmer, 2020, the generating chart $f : \mathbb{R}^d \to \mathbb{R}^D$ is learned simultaneously with $p^*(x)$. They first transform x using a usual flow on $\mathbb{R}^D$, and then project to the first d components which is their proposal for f^{-1}. They then use another flow to learn the latent density π_u. To avoid calculating the Gram determinant of f, which is computationally expensive, especially for $D \gg d$, they propose to train the parameters of f using the mean squared error while updating the parameters of π_u using maximum likelihood. They call the former manifold learning phase and the latter density learning phase. Different learning schemes (alternating and sequential) are proposed to ensure that f encodes the manifold and π captures the density. For the alternating scheme, they alternate for every epoch between a manifold training phase (updating the parameters of f), and the density training phase (updating the parameters for learning π_u). The experiments conducted by Brehmer and Cranmer, 2020 seem to verify that, indeed, $p^*(x)$ is learned exactly. Nevertheless, the ad hoc training procedures without a unified maximum likelihood objective requires some further experimental verification.[3]

State-of-the-art methods for image generation based on NF dequantize the training data as a preprocessing step, see Kingma and Dhariwal, 2018. This dequantization is essentially an inflation of the data-manifold and is typically based on uniform noise. For images, it is generally assumed that $D \gg d$, and thus a dequantization based on Gaussian noise allows us to interpret the dequantization as a thickening of the data-manifold in the normal direction.

5. Results

We have three goals in this section: first, we numerically confirm the scaling factor in equation (4) for different manifolds. Second, we verify that Gaussian noise can be used to approximate a Gaussian noise restricted to the normal space. Third, we numerically test the bounds for σ^2 derived in section3.3. For training details, we refer to appendix B.2. The code for our experiments can be found at `https://github.com/chrvt/Inflation-Deflation`.

The *standard procedure* for our experiments and for evaluating the learned density is the following:

0. **Data generation:** We sample latent variables $u \sim \pi_u(u)$ for a given $\pi_u(u)$, and generate points x on the manifold using a mapping f, that is $x = f(u)$.

1. **Inflation:** We add noise ε to x, $\tilde{x} = x + \varepsilon$, either in the normal space N_x or in the full ambient space. As an acronym for our inflation-deflation method, we use ID. In particular, when the inflation is performed in the normal space, we call the method normal inflation-deflation (NID) and when the inflation is isotropic, we call it isotropic inflation-deflation (IID).

2. **Training:** We learn the inflated distribution, that is $q_\sigma(\tilde{x})$ in case of isotropic noise or $q_{\mathrm{n}}(\tilde{x})$ in case of normal noise, using a block neural autoregressive flow (BNAF) introduced in De Cao et al. (2020).

3. **Deflation:** Given an estimator $\hat{q}_{\mathrm{n}}(\tilde{x})$, we use equation (4) to calculate $p^*(x)$. For a $d-$dimensional manifold embedded in $\mathbb{R}^D$, the scaling factor when using Gaussian noise is $q_{\mathrm{n}}(x|x) = (2\pi\sigma^2)^{\frac{d-D}{2}}$.

3. We further motivate this requirement in appendix B.4.

4. **Quantitative evaluation:** To quantify the quality of the learned density beyond visual similarity, we use the estimate of $p^*(x)$ to approximate $\pi_u(u)$. These densities are related through the Gram determinant of the generating mapping f, $\det G_f$, see section2 . For that, we calculate the KS statistics between this estimate $\hat{\pi}$ and the ground truth π. The KS statistic is defined as

$$KS = \sup_{u \in \mathcal{U}} |F(u) - G(u)|,$$

where F and G are the cumulative distribution functions associated with the probability densities $\pi_u(u)$ and $\hat{\pi}_u(u)$, respectively. By definition, $KS \in [0, 1]$ and $KS = 0$ if and only if $\pi_u(u)$ is equal to $\hat{\pi}_u(u)$ for almost every $u \in \mathcal{U}$. Note that, if our estimate does not yield a density on the manifold (that is it is not normalized to 1), the KS statistics still serves as a relative performance measure as the KS value will be lower bounded by a strictly positive number in this case (1 minus the corresponding normalization constant).
In 2D, comparing two random variables based on $\mathbb{P}(X_1 \leq x_1, X_2 \leq x_2)$ or based on $\mathbb{P}(X_1 \leq x_1, X_2 \geq x_2)$ (or any of the other two combinations) may lead to different results. Hence, for the KS value in 2D, we need to calculate the KS statistics based on all possible orderings and then take the maximum.[4]

5. σ^2-**bounds:** In proposition 7, we derived a necessary condition in form of an upper bound σ^2_{Prop} for σ^2 such that NID can be well approximated by IID. In addtion, we argued that σ^2 should not exceed the curvature radius, see example 3. For $d = 1$, this curvature radius is straightforwardly computed using the first and second derivatives of the generating mapping. For $d = 2$, we use the inverse of the maximum principal curvature, that is the curvature in direction of the greatest change of the normal space. This corresponds to the greatest eigenvalue κ_1 of the Weingarten map (or the shape operator), see Do Carmo (2016). Therefore, to estimate the upper bound for σ^2 uniformly for all points, we sample 10^4 points from the target distribution, calculate $\min\{\sigma^2_{\text{Prop}}, 1/\kappa_1^2\}$ for each point. and then take the 10th percentile of the distribution of those minima in order to have a robust upper bound. As a lower bound σ^2_{LB}, we proposed to choose σ such that the flow's average condition number (which depends on the maximum singular value of J_f) is upper bounded, see appendix A.2 for more details. We estimate the average of the condition number using 10^4 samples from the target distribution.

6. **Benchmarking:** For a known manifold consisting of a single chart f, Gemici et al. (2016) used f to encode the manifold into the corresponding latent space $\mathbb{R}^d$, and then learn the latent density using a standard NF. However, manifolds such as spheres or tori, cannot be described using a single chart. In Brehmer and Cranmer (2020), such degeneracy problems were avoided numerically by simply moving points that are close to singularities away from them. As a consequence, the density close to these singularities cannot be learned exactly (we illustrate this in the first experiment on $\mathbb{S}^1$). In Brehmer and Cranmer (2020), this method is named Flow on manifolds (FOM) and we stick to this notation in the following. In our case, as we are evaluating the qualitative performance using the KS-statistics on the latent densities, we simply

4. Note that we are using the KS statistics in a somewhat unusual way. Indeed, the standard KS statistics compares an empirical distribution with an explicit distribution while we compare here the ground truth density π_u with the estimated density $\hat{\pi}_u$. The supremum is computed as a maximum over evenly spaced points over $\mathcal{U}$.

train a standard NF directly on the latent space for the remaining experiments (thus avoiding potential degeneracy problems altogether). [5]

Remark 9 *For our qualitative evaluation using the KS-statistics, we rely on being able to relate the density in the data-space $p^*(x)$ with the latent distribution $\pi_u(u)$ via the Gram determinant of the manifold generating mapping f. If f consists of singularities, the KS-statistics is still well-defined if these singularities have 0 measure (as it is the case for spheres or tori). However, note that our method does not rely on a specific embedding and thus avoids degeneracy problems during training. The IID model does not even need any explicit knowledge of the manifold except its dimensionality for the right scaling factor. We validate the generality of our method by learning a manifold that cannot be described by a single chart covering all the points up to a set of measure 0. For that, we glue a half sphere with the positive part of a hyperboloid (compactly denoted as $(\mathbb{HS})^2$), see table 2 .*

5.1 Proof of concept: $\mathbb{S}^1$

We start with a circle of radius 3, a $1-$dimensional manifold embedded in $\mathbb{R}^2$, see table 1 .

feature	$\mathcal{U}$	$f(u)$	$\det G_f(x)$
closed	$[-\frac{\pi}{2}, \frac{\pi}{2}]$	$3\begin{pmatrix}\cos(u)\\ \sin(u)\end{pmatrix}$	3

Table 1: Characteristics of $\mathbb{S}^1$.

We let $\pi_u(u) \propto \exp(8\cos(u))$ be a von Mises distribution.

Inflation and Deflation: We inflate $\mathcal{X}$ using 3 types of noise: Gaussian in the normal space (NID), Gaussian in the full ambient space (IID), and χ_2-noise in the normal space as described in example 1(b) with scale parameter 3. Technically, Gaussian noise violates the $Q-$normal reachability assumption. However, if σ^2 is small and the scale parameter for the von Mises distribution is large enough, this is practically fulfilled. Given an estimator for $q_{\mathrm{n}}(\tilde{x})$, we use equation (4) to calculate $p^*(x)$. For the IID and NID methods, we have that $q_{\mathrm{n}}(x|x) = 1/\sqrt{2\pi\sigma^2}$ and for the normal χ^2-noise is $q_{\mathrm{n}}(x|x) = \sqrt{3}e^{-3/2}/(\sqrt{8}\Gamma(\frac{3}{2}))$.

5.1.1 FULL GAUSSIAN VS. NORMAL SPACE NOISE

In figure 3, we show the results for $\sigma^2 = 0.01$ and $\sigma^2 = 1$. In the respective plot, the first row shows training samples from the inflated distributions $q_\sigma(\tilde{x})$ (left), and $q_{\mathrm{n}}(\tilde{x})$ (middle), respectively. We color code a sample $\tilde{x} = x + \varepsilon$ according to $p^*(x)$ to illustrate the impact of noise on the inflated density. Note that the FOM model (top right) does not need any inflation and therefore is trained on samples from $p^*(x)$ only. In the respective plot, the second row shows the learned density for the different models and compares it to the ground truth von Mises distribution $\pi_u(u)$ depicted in black.

5. For the case where the latent dimension is 1, we use a Gaussian-Kernel density estimator to estimate the latent density. Otherwise, we use BNAF.

As we can see, for $\sigma^2 = 0.01$ all models perform very well, although the FOM model slightly fails to capture $p(u)$ for u close to 0 which corresponds to the chosen singularity point (see point 5. in the standard procedure description). For $\sigma^2 = 1$, we see a significant drop in the performance of the Gaussian model. Although the manifold is significantly disturbed, the normal noise model still learns the density almost perfectly [6], so does the normal χ^2-noise model, as predicted by theorem 4.

6. Note that our method still depends on how well an NF can learn the inflated distribution.

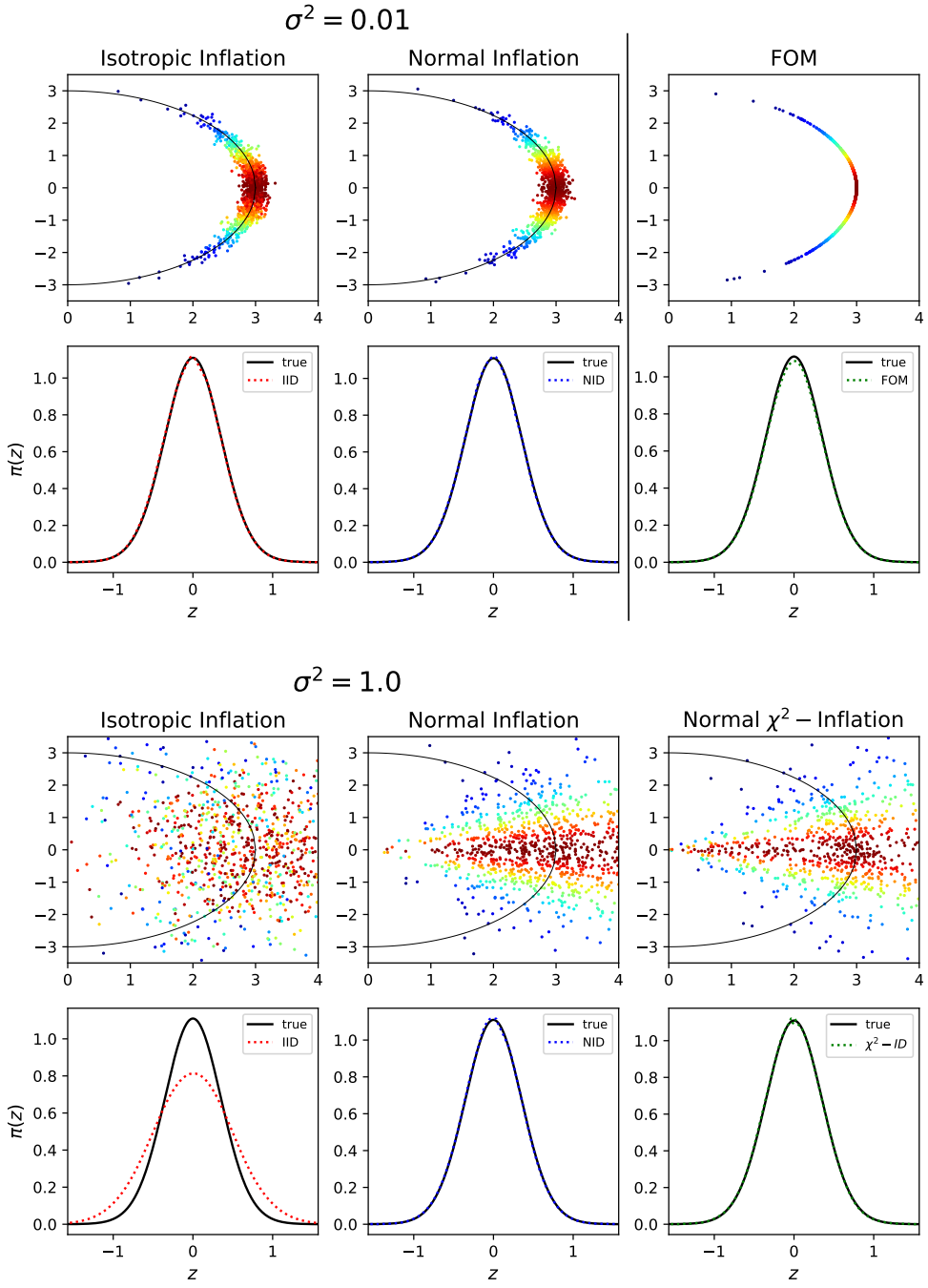

Figure 3: Learned densities for $\sigma^2 = 0.01$ (above) and $\sigma^2 = 1$ (below), respectively. **First row:** Samples used for training the respective model: IID (left), NID (middle), FOM/ χ^2 (right). The black line depicts the manifold $\mathcal{X}$ (a circle with radius 3) and the colors code the value of $p^*(x)$. **Second row:** Colored line: Learned density $\hat{\pi}_u(u)$ according to equation (4) multiplied by 3. Black-line: ground truth von Mises distribution.

5.1.2 NOISE DEPENDENCE AND HIGHER EMBEDDING DIMENSIONS

To measure the dependence of our method on the magnitude of noise, we iterate this experiment for various values of σ^2 and estimate the Kolmogorov-Smirnov (KS) statistics. In figure 4, we display the KS values depending on different levels of noise, for the NID (blue) and IID (orange) methods compared with the ground truth von Mises distribution. Also, we embed the circle into higher dimensions $D = 5, 10, 15, 20$ and repeat this experiment. The result for $D = 2$ and $D = 20$ are shown in the first row (left and right).[7] We add the performance of the FOM model (which is independent of σ^2) horizontally. Besides, we depict the lower and upper bound for σ^2 from section3.3 with dashed vertical lines. In the lower-left image, we show the optimal KS values obtained for both models depending on D. The lower-right image shows the corresponding σ^2 for those optimal KS. In bright, the optimal average σ^2 is shown whereas the dark regions are the minimum respectively maximum values for σ^2 such that we outperformed the FOM benchmark. We note that for both cases, the averaged optimal σ^2 is within the predicted bounds for σ^2 (depicted as dashed black horizontal lines).

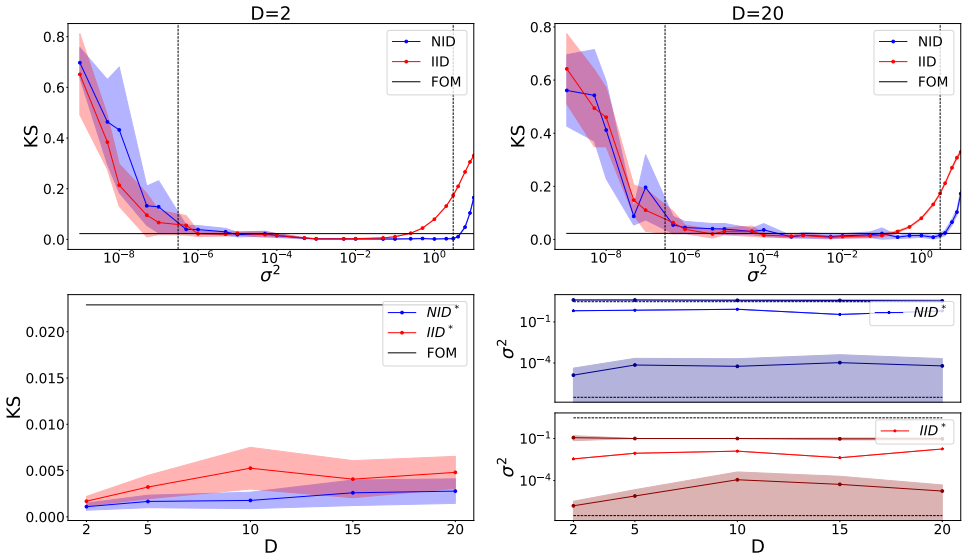

Figure 4: KS values for the NID- (**blue**) and IID-noise method (**orange**) depending on $\sigma^2 \in [10^{-9}, 10]$ and the embedding dimension $D = 5, 10, 15, 20$ in log-scale. For $D = 2$ (top left) and $D = 20$ (top right), the two vertical lines represent the lower and upper bound for σ^2 estimated according to section 3.3 with 10K samples. We plot horizontally the KS value obtained from FOM. **Bottom left:** Optimal KS values depending on D. **Bottom right:** Optimal averaged σ^2 such that optimal KS is obtained (bright). The maximum and minimum σ^2 such that the FOM benchmark is outperformed (dark). The dashed horizontal lines are again the theoretical bounds. We used 10 seeds for the error bars and plot in log-scale.

7. Note that the scaling factor depends on D, $q_{\mathrm{n}}(x|x) = 1/(2\pi\sigma^2)^{\frac{D-d}{2}}$.

The optimal KS values do not change much depending on D, and the NID and IID models approach each other, as predicted. For increasing σ^2, $\tilde{\mathcal{X}}$ resembles more and more a double cone which is not diffeomorphic to $\mathbb{R}^2$ and thus the NF used to train the inflated distribution may not be able to capture the density close to the circle's center correctly. Also, the $Q-$normal reachability is more and more violated with an increasing σ^2.

5.2 Densities on surfaces

We show that we can learn different distributions on different manifolds, see table 2 for an overview of those manifolds and their characteristics.

Manifold	**feature**	$\mathcal{U}$	$\mathbf{f}(\mathbf{u})$	$\det \mathbf{G}_\mathbf{f}(\mathbf{x})$	**Principal Curvature** $\lvert\kappa_1\rvert$
$\mathbb{S}^2$	closed	$[0, 2\pi] \times [0, \pi]$	$\begin{pmatrix}\cos(u_1)\sin(u_2)\\ \sin(u_1)\sin(u_2)\\ \cos(u_2)\end{pmatrix}$	$\sin(u_2)$	1
$\mathbb{T}^2$	closed	$[0, 2\pi] \times [0, 2\pi]$	$\begin{pmatrix}(1+0.6\cos(u_2))\cos(u_1)\\ (1+0.6\cos(u_2))\sin(u_1)\\ 0.6\sin(u_2)\end{pmatrix}$	$0.6(1+0.6\cos(u_2))$	$\max\left\{\frac{1}{0.6}, \left\lvert\frac{\cos(u_1)}{1+0.6\cos(u_1)}\right\rvert\right\}$
$\mathbb{H}^2$	diffeom. to $\mathbb{R}^2$	$[0, +\infty) \times [0, 2\pi]$	$\begin{pmatrix}\sinh(u_1)\cos(u_2)\\ \sinh(u_1)\sin(u_2)\\ \cosh(u_1)\end{pmatrix}$	$(\sinh^2(u_1)+\cosh^2(u_1))\sinh^2(u_1)$	$\frac{1}{(\sinh^2(u_1)+\cosh^2(u_1))^{\frac{1}{2}}}$
thin spiral	open	$(0, +\infty)$	$\pi\sqrt{z}\begin{pmatrix}-\cos(3\pi\sqrt{u})\\ \sin(3\pi\sqrt{u})\end{pmatrix}$	$\pi^2\frac{1+(3\pi\sqrt{u})^2}{4u^2}$	$3\frac{(3\pi)^2u+2}{((3\pi)^2u+1)^{\frac{3}{2}}}$
swiss roll	open	$(0, 1) \times (0, 1)$	$\begin{pmatrix}(\alpha+3\pi u_2)\cos(\alpha+3\pi u_2)\\ 21u_1\\ (\alpha+3\pi u_2)\sin(\alpha+3\pi u_2)\end{pmatrix}$	$(63\pi)^2(1+(0.5+2u_2)^2)$	$\frac{{u_2}^2+2}{({u_2}^2+1)^{\frac{3}{2}}}$
$(\mathbb{HS})^2$	chart 1	$(-\infty, 0] \times [0, 2\pi)$	$\begin{pmatrix}-\cosh(\lvert u_1\rvert)\cos(u_2)\\ -\cosh(\lvert u_1\rvert)\sin(u_2)\\ \sinh(\lvert u_1\rvert)\end{pmatrix}$	$(\sinh^2(\lvert u_1\rvert)+\cosh^2(\lvert u_1\rvert))\cosh^2(\lvert u_1\rvert)$	$\frac{1}{(\sinh^2(\lvert u_1\rvert)+\cosh^2(\lvert u_1\rvert))^{\frac{1}{2}}}$
	chart 2	$[0, \frac{\pi}{2}] \times [0, 2\pi)$	$\begin{pmatrix}\cos(u_2)\cos(u_1+\pi)\\ \sin(u_2)\cos(u_1+\pi)\\ \sin(u_1+\pi)\end{pmatrix}$	$\cos(u_1+\pi)$	1

Table 2: Characteristics of various manifolds.

In figure 5, we show different target densities in data and latent space (columns A and B), along with the learned latent distributions using our method (as described in point 4. of the *standard procedure*) with the normal inflation-deflation method NID (column C). We take the model with σ^2 corresponding to the best KS value. In the last column D, we show how the KS-statistics depends on σ^2 using IID, NID, and the FOM baseline. We refer to appendix B.2, B.2.1, and B.3.1 for the training details, exact latent densities, and additional figures showing the learned latent densities using IID and FOM.

Remarkable, our method performs well on a wide range of manifolds and different target distributions. Whether the manifold is closed (A 1-2, 3-4), open (A 5-6, 7-9), or consists of multiple charts (A 10-11), whether the latent variables are idependent (B 1,3,6,8,10) or dependent (B 2,4,5,9,11), whether the distribution is supported on points for which the Gram determinant is 0 (A 1-2) or on points for which the Gram determinant is arbitrarily large (A 7), the induced latent density (and therefore the data-density $p^*(x)$) is approximated well. This is not only reflected in the visual similarity to the target distribution (columns B vs. C) but also in the KS statistics (column D). Surprisingly, the best KS values for the IID and NID methods are of the same order as the FOM baseline (tables in D). This is striking as the IID and NID methods are trained in data space, in contrast to the FOM which is trained in latent space directly (see point 4. in the *standard procedure*).

In some cases, the NID even outperforms the FOM significantly (see tables in D 2-3). The optimal KS value for IID is only slightly worse than the one for NID showing that indeed our method can even be used without any explicit knowledge of the manifold (except its dimensionality for the right scaling factor).
Note that the NID method always allows for a greater range of σ^2 compared to IID, except for the thin spiral for which both curves have almost the same course (D 7). As an extreme case, the geometry of the hyperboloid $\mathbb{H}^2$ even allows for very large values of σ^2 when using NID (D 5-6). Notably, almost all the KS curves are U-shaped. However, for the torus (D 3-4) and swiss roll (D 8-9) the KS value for IID decreases approaching $\sigma^2 = 10^1$ before increasing again. For increasing σ^2, the induced latent distribution $\hat{\pi}_u$ is increasingly flat. Then, certain values of σ^2 lead to the right scaling such that $\int_{\mathcal{U}} \hat{\pi}_u(u)du \approx 1$ which decreases the KS value.

The lower bound based on the condition number nicely predicts the magnitude of noise required to approximate $p^*(x)$ well using IID. Also the upper bound behaves as predicted and matches almost the onset for which Gaussian noise leads to a bad KS value (except for the hyperboloid D 6). It is necessary (though not sufficient, see D 6) for σ^2 to be lower than this upper bound such that IID approximates NID well and thus can be used to approximate $p^*(x)$.

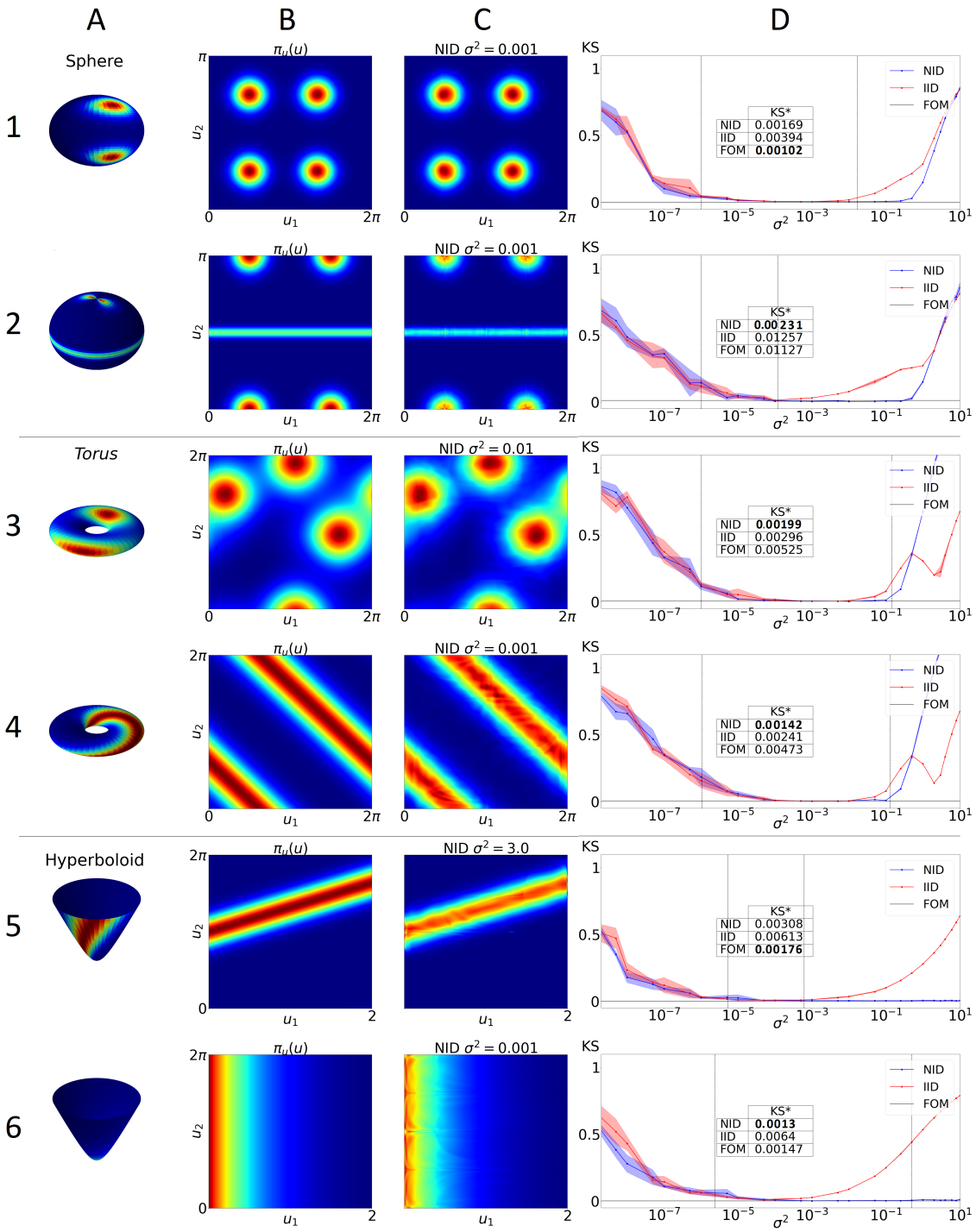

Figure 5: **Columns A and B:** Target density in data (A) and latent space (B) for various manifolds and different latent distributions. **Column C:** Best learned density using our method with the method NID. **Column D:** KS vs. σ^2 plot for the IID and NID methods (we used 3 seeds for the error bars) with the KS value of FOM as horizontal line. **Table in D:** Optimal KS values for the different models (best, that is lowest, in bold). **Vertical lines in D** Lower and upper bound (see point 5. of the standard procedure).

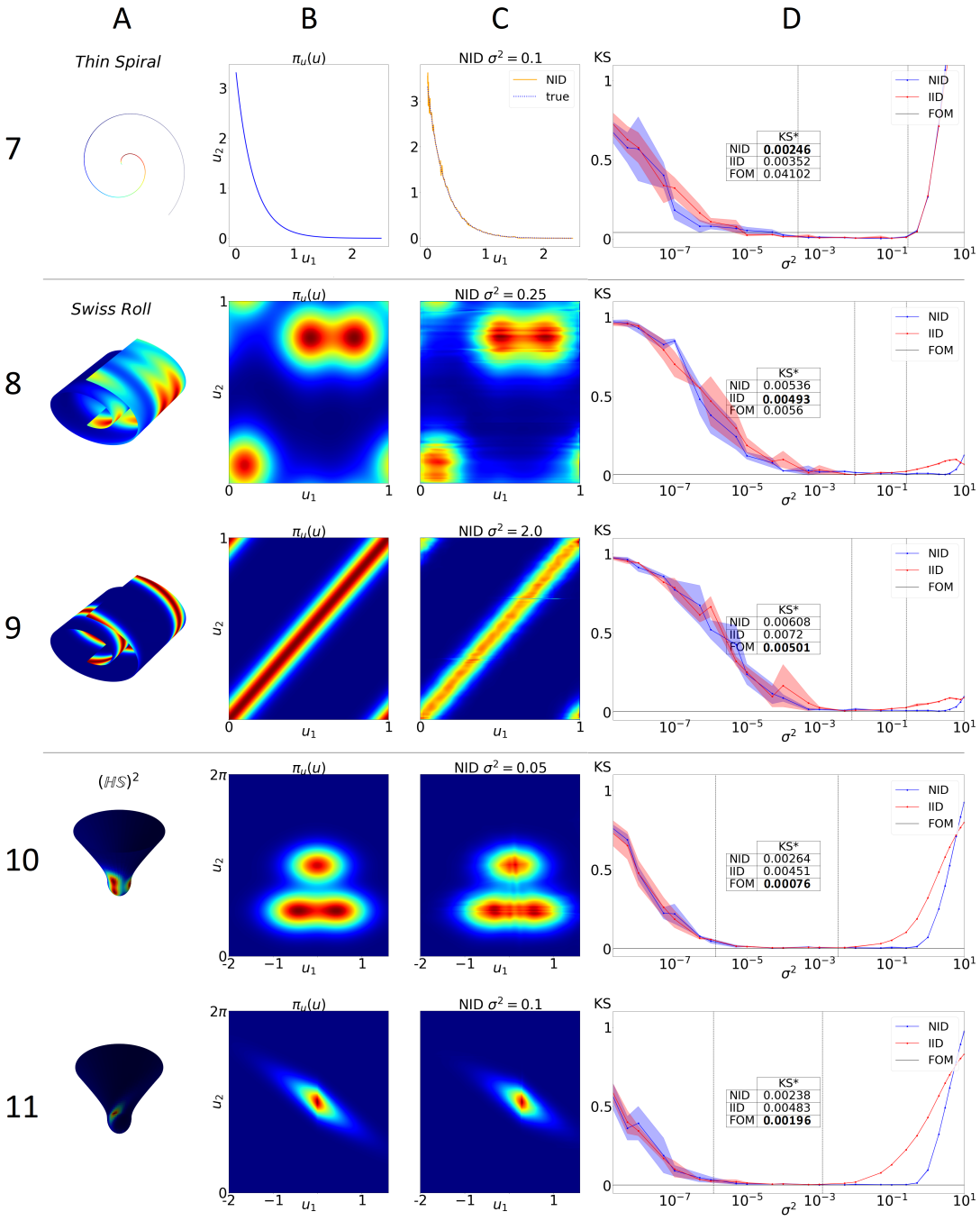

figure 5 (continuation): **Columns A and B:** Target density in data (A) and latent space (B) for various manifolds and different latent distributions. **Column C:** Best learned density using our method with the method NID. **Column D:** KS vs. σ^2 plot for the IID and NID methods (we used 3 seeds for the error bars) with the KS value of FOM as horizontal line. **Table in D:** Optimal KS values for the different models (best, that is lowest, in bold). **Vertical lines in C** Lower and upper bound (see point 5. of the standard procedure).

5.3 Density estimation on MNIST

Finally, we end this sectionwith an application on the handwritten digit dataset MNIST, Lecun et al. (1998). The manifold hypothesis states that real-world data, such as images, can be described by a few key features only, thus populating a low-dimensional manifold in the high-dimensional embedding space.

To estimate the density of digit 1 images, both the inflation-deflation method and the $\mathcal{M}-$flow need to know the manifold dimensionality d.[8] Estimating this intrinsic dimensionality d is an active research area, see Hein and Audibert (2005) and Facco et al. (2017). For instance, Hein and Audibert (2005) estimate the intrinsic dimensionality of MNIST digit 1 to be roughly 8.

We test the utility of learned digit 1 likelihoods for out-of-distribution detection (OOD) using IID (isotropic inflation-deflation) and the $\mathcal{M}-$flow. In figure 7, we show the log-likelihood densities (estimated using kernel density estimation) on the MNIST test set after training on digit 1 images from the training set only. For the IID, we preprocess the training set by adding Gaussian noise with $\sigma^2 = 0.1$ to the 8-bit images.[9] For the $\mathcal{M}-$flow, we leave the training set unaltered. Though, we did not find this preprocessing (or the absence of it) to have a significant impact on the log-likelihoods for both methods. We refer to appendix B.3 for more training details and additional plots for different preprocessing protocols.

In figure 7, we want to highlight two interesting observations. First, the log-likelihoods of digits other than 1 are not significantly different using the IID or $\mathcal{M}-$flow method for OOD. One can see this by comparing the area of intersectionof the digit 1 density (orange) with the other digits (other colors). The greater this area, the more out of distribution examples (in this case MNIST digits other than 1) would be classified as digit 1 when using a naive classifier based on an ad hoc log-likelihood threshold. This area is ≈ 0.07 for both methods. Our second observation is that the absolute log-likelihood values differ substantially. As both methods try to estimate the density $p^*(x)$ supported on a low-dimensional manifold, we would have expected similar log-likelihood values. The fact that these values are several magnitudes apart, together with the observation that an inflation is not strictly necessary using an NF to learn the data-density (see figure 10 in sectionB.3.1), indicates that the MNIST digit 1 images do not strictly lie on a low-dimensional manifold embedded in $\mathbb{R}^D, D = 784$. In such a case, the $\mathcal{M}-$flow would still try to fit the training set onto a manifold which may lead to overfitting and unforeseeable log-likelihood values when evaluating on a test set. In contrast, the IID method is by construction robust to overfitting as the addition of isotropic noise leads to similar log-likelihood values in the vicinity of the data-manifold.

Finally, we want to revisit our remark on the computational complexity of the $\mathcal{M}-$flow, see section2 . To evaluate the density using the $\mathcal{M}-$flow, one needs to calculate the Gram determinant which has a computational complexity of $\mathcal{O}(d^2D) + \mathcal{O}(d^3)$.[10] Indeed, to evaluate 1000 digits using a batch size of 1, the $\mathcal{M}-$flow needs about 17.5 hours. For the same amount, the inflation-deflation method needs less than 10 seconds.

8. Note that the inflation-deflation method only needs to know the dimensionality d for the right scaling factor during testing. The $\mathcal{M}-$flow, however, needs to know d for training. In exchange, the $\mathcal{M}-$flow learns a low-dimensional representation which the inflation-deflation method does not.
9. Note that this is different from the usual uniform dequantization performed on images.
10. The necessary Jacobian is computed using differentiation.

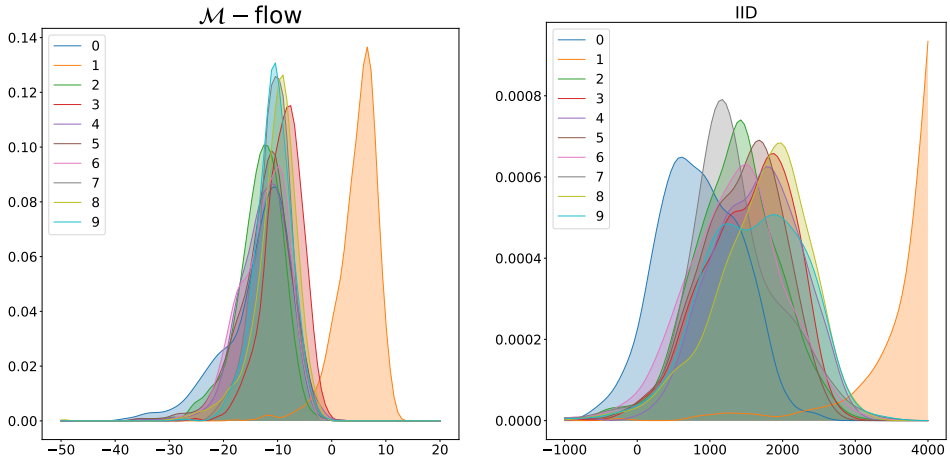

Figure 7: Log likelihoods on various MNIST test digits using $\mathcal{M}-$flow (**left**) and IID (**right**) trained on digit 1 only.

6. Discussion

To overcome the limitations of NFs to learn a density $p^*(x)$ defined on a low-dimensional manifold, we proposed to embed the manifold into the ambient space such that it becomes diffeomorphic to $\mathbb{R}^D$, learn this inflated density using an NF, and, finally, deflate the inflated density according to theorem 4. There, we provided sufficient conditions on the choice of inflation such that we can compute $p^*(x)$ exactly. Our method depends on some critical points that we addressed in section3.3. So far, the magnitude of noise σ^2 when using NFs on real-world data is somewhat chosen arbitrarily. As a step to overcome this arbitrariness, we derived an upper bound for σ^2 in proposition 7 and established an interesting connection to the manifold learning literature in theorem 8. However, proposition 7 may not be very useful as such in real-world application and numerical methods need to be considered which potentially suffer from the curse of dimensionality. On a more positive note, our various experiments on different manifolds suggest that a great range for σ^2 leads to good results, even when using full Gaussian noise. Thus, including σ^2 into the standard hyperparameter search will likely suffice.

Our theoretical results open new research avenues. Using full Gaussian noise to learn the inflated distribution smears information on $p^*(x)$, in particular, if $p^*(x)$ has many local extrema. This loss of information may be especially impactful in out-of-distribution (OOD) detection or when it comes to adversarial robustness. Therefore, developing methods that allow generating noise in the manifold's normal space could improve the performance of NFs on such tasks.

Another interesting direction is to exploit the product form of equation (2) and learn low-dimensional representations by forcing the NF to be noise insensitive in the first d-components and noise sensitive in the remaining ones. Inverting the corresponding flow allows sampling directly on the manifold.

Acknowledgments

We would like to thank Johann Brehmer for clarifying discussions on the manifold flow, and Simone C. Surace for useful discussions on manifolds.
This study has been supported by the Swiss National Science Foundation grant 31003A_175644.

Appendix A. Appendix

A.1 Proof of theorem 4

Let $x \in \mathcal{X}$. Since $\mathcal{X}$ is a $d-$dimensional C^2 manifold, there exists an open neighborhood $\mathcal{B}_x$ of x in $\mathcal{X}$, an open set $\mathcal{U}_x$ in $\mathbb{R}^d$, and an invertible map $f : \mathcal{U}_x \mapsto \mathcal{B}_x$, $\mathcal{U}_x \subset \mathbb{R}^d$, such that f and f^{-1} are twice continuously differentiable. It follows that the Gram determinant of J_f is non-zero for all $x \in \mathcal{B}_x$, that is $\det G_f(x) \neq 0 \; \forall x \in \mathcal{B}_x$. We exploit this by constructing a local diffeomorphism $\tilde{f}$ on the inflated space $\widetilde{\mathcal{X}} = \bigcup_{x \in \mathcal{X}} N_{q_{\mathrm{n}}(\cdot|x)}$ in the following.
For that, we denote by A_u the matrix with columns consisting of normal vectors spanning the normal space in $x = f(u), u \in \mathcal{U}_x$. Without loss of generality, we can set $\det A_u^T A_u = 1$. With $\mathcal{V}_x \subset \mathbb{R}^{D-d}$, we define $\tilde{f} : \mathcal{U}_x \times \mathcal{V}_x \subset \mathbb{R}^d \times \mathbb{R}^{D-d} \to \widetilde{\mathcal{B}}_x$ for some $\widetilde{\mathcal{B}}_x \subset \widetilde{\mathcal{X}}$ as follows:

$$\tilde{f}(u, v) = f(u) + A_u v.$$

Note that, by assumption, $0 \in \mathcal{V}_x$. Thus, for v sufficiently small, that is $||v|| < \varepsilon$ for some $\varepsilon > 0$, $\tilde{f}$ is indeed a diffeomorphism which follows from the inverse function theorem.[11] Our key observation is that

$$\det G_{\tilde{f}}(x) = \det G_f(x) \tag{5}$$

which allows us to relate the density on $\widetilde{\mathcal{X}}$ to the density on $\mathcal{X}$. For the sake of clarity, we prove equation (5) in lemma 10 below.

Now let $\tilde{x} = x + \varepsilon_{\mathrm{n}} \in \widetilde{\mathcal{X}}$ such that $\tilde{x} \in \widetilde{\mathcal{B}}_x$. Since $\mathcal{X}$ is $Q-$normally reachable, $\mathcal{P}_{\tilde{X}}-$almost all $\tilde{x}$ are uniquely determined by some $(u, v)^T = \tilde{f}^{-1}(\tilde{x}) \in \mathcal{U}_x \times \mathcal{V}_x$, and since u and v are sampled independently by assumption, it must hold that

$$q_{\mathrm{n}}(\tilde{x}) = (\det G_{\tilde{f}}(\tilde{x}))^{-\frac{1}{2}} \pi_u(u) \pi_v(v)$$

where $\pi_v(v)$ is the noise generating latent distribution. Note that $q_{\mathrm{n}}(\tilde{x})$ is the density of $d\mathbb{P}_{\tilde{X}}$ with respect to the volume form $dV_{\tilde{f}}$. For $\tilde{x} = x$, we have that $v = 0$ and thus

$$q_{\mathrm{n}}(x) = (\det G_{\tilde{f}}(x))^{-\frac{1}{2}} \pi_u(u) \pi_v(0).$$

Now since $\det G_{\tilde{f}}(x) = \det G_f(x)$, we have that

$$\begin{aligned} q_{\mathrm{n}}(x) =& (\det G_{\tilde{f}}(x))^{-\frac{1}{2}} \pi_u(u) \pi_v(0) \\ =& (\det G_f(x))^{-\frac{1}{2}} \pi_u(u) \pi_v(0) \\ =& p^*(x) \pi_v(0) \\ =& p^*(x) q_{\mathrm{n}}(x|x) \end{aligned}$$

where in the last step we have used that the Gram determinant of the normal space generating mapping is 1 such that $\pi_v(0) = q_{\mathrm{n}}(x|x)$. As x was chosen arbitrarily on the manifold, this ends the proof.

11. In fact, we need to show that $\det J_{\tilde{f}}(u, 0) \neq 0$ for all $(u, 0)$. Because this implies the existence of a local neighborhood such that $\tilde{f}$ is diffeomorphic to the image of this local neighborhood. That $\det J_{\tilde{f}}(u, 0) \neq 0$ follows immediately from lemma 10 .

Lemma 10 *For $\tilde{f}$, f and x as defined above, we have that* $\det G_{\tilde{f}}(x) = \det G_f(x)$.

Proof The Jacobian of $\tilde{f}$ is given by

$$J_{\tilde{f}}(u,v) = \left[\begin{array}{c:c} J_f(u) + \frac{\partial}{\partial u} A_u v & A_u \end{array}\right]$$

where $\frac{\partial}{\partial u}$ denotes the Jacobian of a function depending on u, and the dashed line separates two block matrices. Here we need that $f \in C^2$ to ensure the Jacobian is real. For points on the manifold is $v = 0$, and thus the Gram determinant reduces to

$$\begin{aligned}
\det G_{\tilde{f}}(x) &= \det \left(J_{\tilde{f}}(u,0)^T J_{\tilde{f}}(u,0) \right) \\
&= \det \left[\begin{array}{c:c} J_f(u)^T J_f(u) & J_f(u)^T \cdot A_u \\ \hdashline A_u^T \cdot J_f(u)^T & A_u^T A_u \end{array}\right] \\
&= \det \left[\begin{array}{c:c} J_f(u)^T J_f(u) & 0_{d \times D-d} \\ \hdashline 0_{D-d \times d} & A_u^T A_u \end{array}\right] \\
&= \det J_f(u)^T J_f(u) \cdot \det A_u^T A_u \\
&= \det J_f(u)^T J_f(u) \\
&= \det G_f(x)
\end{aligned}$$

where for the third equality we have exploited the fact that the column vectors of J_f and A_u are orthogonal. This was to be shown. ■

A.2 Lower bound

With the same conditions on the manifold as in theorem 4, inflating the manifold in normal direction can be locally described by the chart $\tilde{f} : \mathcal{U}_x \times \mathcal{V}_x \to \tilde{\mathcal{B}}_x$ with

$$\tilde{f}(u,v) = f(u) + \sigma A_u v$$

where σ is sufficiently small and A_u denotes the matrix with columns consisting of normal vectors spanning the normal space in $x = f(u), u \in \mathcal{U}_x$, see sectionA.1 for more details. The condition number of the Jacobian of $\tilde{f}$ is given by the ratio of the greatest and lowest singular value. The singular values are given by the square roots of the eigenvalues of the Gram matrix of $J_{\tilde{f}}$. Since

$$J_{\tilde{f}}(u,v) = \left(\begin{array}{c:c} J_f(u) + \frac{\partial}{\partial u} A_u v & A_u \end{array}\right)$$

we get

$$\begin{aligned}
G_{\tilde{f}}(x) &= \left(J_{\tilde{f}}(u,v)^T J_{\tilde{f}}(u,v) \right) \\
&= \left(\begin{array}{c:c} J_f(u)^T J_f(u) + \mathcal{O}(\sigma^2) & 0 \\ \hdashline 0 & \sigma^2 A_u^T A_u \end{array}\right) \\
&= \left(\begin{array}{c:c} J_f(u)^T J_f(u) + \mathcal{O}(\sigma^2) & 0 \\ \hdashline 0 & \sigma^2 I \end{array}\right)
\end{aligned}$$

where in the first step we have exploited that A_u is normal to the tangent space spanned by the columns of $J_f(u)$, and in the second step we set without loss of generality $A_u^T A_u = I$. If σ^2 is very small, then the singular values of $G_{\tilde{f}}(x)$ are given by the eigenvalues of $J_f(u)^T J_f(u)$ and σ^2. We denote the maximum singular value of $J_f(u)^T J_f(u)$ by λ_{max}. Thus, the ratio of the greatest and lowest singular value, the condition number κ_c, is given by

$$\kappa_c = \frac{\sqrt{\lambda_{\text{max}}}}{\sigma}.$$

We choose σ such that $\kappa_c \leq 10^3$ which leads to

$$\sigma \geq \frac{\sqrt{\lambda_{\text{max}}}}{10^3}. \tag{6}$$

Therefore, as a lower bound, we calculate the mean value (with respect to the target distribution) of the right hand side of equation (6).

A.3 Proof of proposition 5

This follows immediately from the universality of standard NFs, see section2, and theorem 4.

A.4 Proof of statement in remark 6

We denote the probability measure of the random variable X as $\mathbb{P}_X$ and it is defined on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ where $\mathcal{B}(\mathcal{X})$ is the set of Borel sets in $\mathbb{R}^D$ intersected with $\mathcal{X}$. For a realization of X, say x, we denote the probability measure of the shifted random variable $x + \mathcal{E}_{\text{n}}$ as $\mathbb{P}_{\tilde{X}|X=x}$ and it is defined on $(\mathcal{N}_x, \mathcal{B}(N_x))$. We extend both measures to $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$ by setting the probabilities to 0 whenever a set $A \in \mathcal{B}(\mathbb{R}^D)$ has no intersectionwith $\mathcal{X}$ or N_x, respectively. For instance, that means for $\tilde{x} \in N_x$ that

$$\mathbb{P}[x + \mathcal{E}_{\text{n}} \in (\tilde{x}, \tilde{x} + d\tilde{x})] = \mathbb{P}[x + \mathcal{E}_{\text{n}} \in (\tilde{x}, \tilde{x} + d\tilde{x}) \cap N_x] = \mathbb{P}_{\tilde{X}|X=x}[(\tilde{x}, \tilde{x} + d\tilde{x}) \cap N_x]$$

where $(\tilde{x}, \tilde{x} + d\tilde{x})$ denotes an infinitesimal volume element around $\tilde{x}$.

The mapping $(x, \varepsilon_{\text{n}}) \mapsto x + \varepsilon_{\text{n}}$ is $\mathcal{B}(\mathbb{R}^D) \times \mathcal{B}(\mathbb{R}^D)-$measurable, and thus $\tilde{X} = X + \mathcal{E}_{\text{n}}$ is a random variable on $(\mathbb{R}^D, \mathcal{B}(\mathbb{R}^D))$ and has the pushforward of $\mathbb{P}_{(X,\mathcal{E}_{\text{n}})}$ with regard to the mapping $(x, \varepsilon_{\text{n}}) \to x + \varepsilon_{\text{n}}$ as probability measure where $\mathbb{P}_{(X,\mathcal{E}_{\text{n}})}$ is the joint measure of X and $\mathcal{E}_{\text{n}}$. Thus, for $A \in \mathcal{B}(\widetilde{\mathcal{X}})$, we have that

$$\mathbb{P}_{\tilde{X}}(A) = \mathbb{P}_{(X,\mathcal{E}_{\text{n}})}\left(\{(x, \varepsilon_{\text{n}}) \in \mathbb{R}^D \times \mathbb{R}^D | x + \varepsilon_{\text{n}} \in A\}\right). \tag{7}$$

Now let $\tilde{x} \in N_x$ for an $x \in \mathcal{X}$. Since $\mathcal{X}$ is $Q-$normally reachable, $\mathbb{P}_{\tilde{X}}-$almost all $\tilde{x}$ are uniquely determined by $(x, \varepsilon_{\text{n}})$ such that $\tilde{x} = x + \varepsilon_{\text{n}}$. Therefore, we have for $\mathbb{P}_{\tilde{X}}-$almost all $\tilde{x} = x + \varepsilon_{\text{n}}$ that

$$\begin{aligned}
\mathbb{P}_{\tilde{X}}((\tilde{x}, \tilde{x} + d\tilde{x}) \cap \widetilde{\mathcal{X}}) &= \mathbb{P}_{(X,\mathcal{E}_{\text{n}})}\left(\{(x, \varepsilon_{\text{n}}) \in \mathbb{R}^D \times \mathbb{R}^D | x + \varepsilon_{\text{n}} \in (\tilde{x}, \tilde{x} + d\tilde{x}) \cap \widetilde{\mathcal{X}}\}\right) \\
&= \mathbb{P}\left(X + \mathcal{E}_{\text{n}} \in (\tilde{x}, \tilde{x} + d\tilde{x}) \cap \widetilde{\mathcal{X}}\right) \\
&= \mathbb{P}\left(X \in (x, x + dx) \cap \mathcal{X}\right) \cdot \mathbb{P}\left(x + \mathcal{E}_{\text{n}} \in (\tilde{x}, \tilde{x} + d\tilde{x}) \cap N_x\right) \\
&= \mathbb{P}_X\left((x, x + dx) \cap \mathcal{X}\right) \cdot \mathbb{P}_{\tilde{X}|X=x}\left((\tilde{x}, \tilde{x} + d\tilde{x}) \cap N_x\right)
\end{aligned}$$

where for the first equality we used equation (7) and for the third the fact that $(x, \varepsilon_{\mathrm{n}})$ is almost surely uniquely determined by $\tilde{x}$.

Both probability measures on the right-hand side have a density. For $\mathbb{P}_X$ with respect to dV_f, see section2, this density is $p^*(x)$. Similarly, since N_x is a linear subspace of $\mathbb{R}^D$, $q_{\mathrm{n}}(\tilde{x}|x)$ is the density of $\mathbb{P}_{\tilde{X}|X=x}$ with respect to a volume form dV_h where h is the mapping from $\mathbb{R}^{D-d}$ to N_x. Then, the corresponding density of $\mathbb{P}_{\tilde{X}}$ with respect to the product measure $V_\otimes := V_f \otimes V_h$ is given by

$$q_{\mathrm{n}}^{\otimes}(\tilde{x}) = p^*(x) q_{\mathrm{n}}(\tilde{x}|x)$$

and it holds that

$$\begin{aligned}\int_{\widetilde{\mathcal{X}}} q_{\mathrm{n}}^{\otimes}(\tilde{x}) dV_\otimes(\tilde{x}) &= \int_{\mathcal{X}} \int_{N_x} p^*(x) q_{\mathrm{n}}(\tilde{x}|x) dV_h(\tilde{x}) dV_f(x) \\ &= \int_{\mathcal{X}} p^*(x) dV_f(x) \\ &= 1,\end{aligned}$$

as needed for a density on $\widetilde{\mathcal{X}}$. This ends the proof.

A.5 Proof of proposition 7

The generating function f is an embedding for $\mathcal{X}$ and $X = f(u)$ has the density $p^*(x)$ for $x \in \mathcal{X}$. We may extend the domain of $p^*(x)$ to include all points $x \in \mathbb{R}^D$ using the Dirac-delta function. We denote this density with $\bar{p}(x)$ at it is given by

$$\bar{p}(x) = \int_{\mathcal{U}} \delta(x - f(u)) \pi_u(u) du,$$

see Au and Tam (1999). After inflating X, we have that

$$p_\Sigma(\tilde{x}) = \int_{\mathcal{U}} \mathcal{N}(\tilde{x}; f(u), \Sigma) \pi_u(u) du$$

with covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$ where for $\Sigma = \sigma^2 I$ we have that $\lim_{\sigma \to 0} p_\Sigma(\tilde{x}) = \bar{p}(\tilde{x})$. Assume $\tilde{x} = x$ for some $x \in \mathcal{X}$. We Taylor expand $f(u)$ around $u_0 = f^{-1}(x)$ up to first order,

$$f(u) \approx f(u_0) + J_f(u_0)(u - u_0),$$

and $\pi_u(u)$ up to second order,

$$\pi_u(u) \approx \pi_u(u_0) + \pi_u(u_0)'(u - u_0) + \frac{1}{2}(u - u_0)^T \pi_u''(u_0)(u - u_0).$$

where $\pi_u(u_0)'$ denotes the gradient and $\pi_u''(u_0)$ the Hessian of π evaluated at u_0, thus $\pi_u(u_0)' \in \mathbb{R}^d$ and $\pi_u''(u_0) \in \mathbb{R}^{d \times d}$. Then, we can approximate $p_\Sigma(x)$ as follows:

$$\begin{aligned}p_\Sigma(x) \approx & \frac{1}{\sqrt{(2\pi)^D \det(\Sigma)}} \int_{\mathcal{U}} \exp\left(-\frac{1}{2}(u - u_0)^T J_f^T \Sigma^{-1} J_f (u - u_0)\right) \cdot \\ & \cdot \left(\pi_u(u_0) + \pi_u'(u_0)^T (u - u_0) + \frac{1}{2}(u - u_0)^T \pi_u''(u_0)(u - u_0)\right) du.\end{aligned}$$

Now define $\hat{\Sigma}^{-1} = J_f^T \Sigma^{-1} J_f$. Then,

$$p_\Sigma(x) \approx \frac{\sqrt{\det(\hat{\Sigma})}}{\sqrt{(2\pi)^{D-d}\det(\Sigma)}} \int_{\mathcal{U}} \frac{1}{\sqrt{(2\pi)^d \det(\hat{\Sigma})}} \exp\left(-\frac{1}{2}(u-u_0)^T \hat{\Sigma}^{-1}(u-u_0)\right) \cdot$$
$$\cdot (\pi_u(u_0) + \pi_u'(u_0)^T(u-u_0) + \frac{1}{2}(u-u_0)^T \pi_u''(u_0)(u-u_0))du.$$

Thus, we can exploit the Gaussian in $\mathcal{U}$-space and get

$$\begin{aligned} p_\Sigma(x) \approx & \frac{\sqrt{\det(\hat{\Sigma})}}{\sqrt{(2\pi)^{D-d}\det(\Sigma)}} (\pi_u(u_0) + \frac{1}{2}\mathbb{E}\left[(u-u_0)^T \pi_u''(u_0)(u-u_0)\right]) \\ = & \frac{\sqrt{\det(\hat{\Sigma})}}{\sqrt{(2\pi)^{D-d}\det(\Sigma)}} (\pi_u(u_0) + \frac{1}{2}||\pi_u''(u_0) \odot \hat{\Sigma}||_+), \end{aligned}$$

where $\odot$ stands for the elementwise multiplication and $||A||_+ = \sum_{i,j=1}^d A_{ij}$ for a $\mathbb{R}^d \times \mathbb{R}^d$ matrix A.

For the special case where $\Sigma = \sigma^2 I_D$, we can simplify this expression by exploiting that

$$\begin{aligned} \frac{\sqrt{\det(\hat{\Sigma})}}{\sqrt{(2\pi)^{D-d}\det(\Sigma)}} &= \frac{1}{(2\pi)^{\frac{D-d}{2}}} \frac{\sigma^{-D}}{\sigma^{-d}\sqrt{\det G_f}} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{D-d}{2}}\sqrt{\det G_f}}. \end{aligned}$$

Thus, in total, we get for this special choice of Σ

$$\begin{aligned} p_\sigma(x) \approx & \frac{1}{(2\pi\sigma^2)^{\frac{D-d}{2}}\sqrt{\det G_f}} (\pi_u(u_0) + \frac{\sigma^2}{2}||\pi_u''(u_0) \odot (J_f^T J_f)^{-1}||_+) \\ = & \frac{1}{(2\pi\sigma^2)^{\frac{D-d}{2}}\sqrt{\det G_f}} \pi_u(u_0)(1 + \frac{\sigma^2}{2\pi_u(u_0)}||\pi_u''(u_0) \odot (J_f^T J_f)^{-1}||_+) \end{aligned} \tag{8}$$

Now, if

$$\left|\frac{\sigma^2}{2\pi_u(u_0)}||\pi_u''(u_0) \odot (J_f^T J_f)^{-1}||_+\right| \ll 1,$$

then $q_{\mathrm{n}}(x|x) \approx 1/(2\pi\sigma^2)^{\frac{D-d}{2}}$ as $1/(2\pi\sigma^2)^{\frac{D-d}{2}}$ from equation (8) is exactly the normalization constant obtained when inflating the manifold with Gaussian noise in the normal space, $q_{\mathrm{n}}(x|x) = 1/(2\pi\sigma^2)^{\frac{D-d}{2}}$. Therefore, multiplying equation (8) on both sides with $q_{\mathrm{n}}(x|x)$ and letting σ tend to zero, we have that

$$\lim_{\sigma^2 \to 0} \frac{q_\sigma(x)}{q_{\mathrm{n}}(x|x)} = p^*(x)$$

This ends the proof.

A.6 Proof of theorem 8

The result follows directly from the definition of the reach number $\tau_{\mathcal{X}}$ of $\mathcal{X}$. It is defined as the supremum of all $r \geq 0$ such that the orthogonal projection $\mathrm{pr}_{\mathcal{X}}$ on $\mathcal{X}$ is well-defined on the $r-$neighbourhood $\mathcal{X}^r$ of $\mathcal{X}$,

$$\mathcal{X}^r := \{\tilde{x} \in \mathbb{R}^D | \ \mathrm{dist}(\tilde{x}, \mathcal{X}) \leq r\}$$

where $\mathrm{dist}(\tilde{x}, \mathcal{X})$ denotes the distance of $\tilde{x}$ to $\mathcal{X}$. Thus,

$$\tau_{\mathcal{X}} = \sup \left\{ r \geq 0 \mid \forall \tilde{x} \in \mathbb{R}^D, \ \mathrm{dist}(\tilde{x}, \mathcal{X}) \leq r \implies \exists! x \in \mathcal{X} \text{ s.t. } \mathrm{dist}(\tilde{x}, \mathcal{X}) = ||\tilde{x} - x|| \right\},$$

see Definition 2.1. in Berenfeld and Hoffmann (2019). By assumption $\tau_{\mathcal{X}} > 0$. Thus for all $\tilde{x} \in \mathcal{X}^{\tau_{\mathcal{X}}}$ we have that $x := \mathrm{pr}_{\mathcal{X}}(\tilde{x})$ is unique. Since $\mathcal{X}$ is a closed manifold, it must hold that $\tilde{x} \in N_x$ where N_x denotes the normal space in x. Let the noise generating distributions be a uniform distribution on the ball with radius $\tau_{\mathcal{X}}$, thus

$$q_{\mathrm{n}}(\tilde{x}|x) = \mathrm{Uniform}(\tilde{x}; B(x, \tau_{\mathcal{X}}) \cap N_x),$$

where $B(x, \tau_{\mathcal{X}})$ denotes a $D-$dimensional ball with radius $\tau_{\mathcal{X}}$ and center x. Then, we have for $\widetilde{\mathcal{X}} = \bigcup_{x \in \mathcal{X}} N_{q_{\mathrm{n}}(\cdot|x)}$ that

$$\widetilde{\mathcal{X}} = \mathcal{X}^{\tau_{\mathcal{X}}}.$$

Thus, $\mathcal{X}$ is $Q-$normally reachable where $Q := \{q_{\mathrm{n}}(\cdot|x)\}_{x \in \mathcal{X}}$.

Appendix B. Experiments

For all expriments, we use Adam optimizer with an initial learning rate 0.1, a learning rate decay of 0.5 after 2000 optimization steps without improvement (learning rate patience). We use a batch size of 200. No hyperparameter fine-tuning was done.

B.1 Technical details for circle experiments

We use a BNAF (block neural autoregressive flow) to learn the inflated density, see table 3 for the details. There, we report the number of hidden layers, hidden dimensions (which scale with the dimensionality of the embedding space), total parameters of the model, and, finally, the number of gradient steps (iterations).

For the FOM and χ^2-noise models, we use the same architecture as for the $D = 2$ case.

Data dimension	hidden layers	hidden dimension	total parameters	iterations
2	3	100	31,204	70000
5	3	250	192,010	70000
10	3	500	764,000	70000
15	3	750	1,716,030	100000
20	3	1000	3,048,040	100000

Table 3: BNAF details for circle experiments.

B.2 Technical details for density estimation tasks

We use a BNAF (block neural autoregressive flow) to learn the inflated density, see table 4 for the details. There, we report the number of hidden layers, hidden dimensions, total parameters of the model, and, finally, the number of gradient steps (iterations).

Data dimension	hidden layers	hidden dimension	total parameters	iterations
1	6	210	31,204	50000
2	6	210	268,384	50000
3	6	210	268,806	50000
4	6	200	244,408	50000

Table 4: BNAF details for density estimation experiments.

B.2.1 LATENT DENSITIES

In table 5 we show the latent densities used in the experiments in order of appearance.

Manifold	$\pi_u(u) \propto$	Parameters
$\mathbb{S}^2$	$\sum_{i=1}^{4} \exp(6\cos(u_1-\mu_i)) \cdot \exp(6\cos(2(u_2-m_i)))$	table 6
	$\sum_{i=1}^{2} \exp(6\cos(u_1-\mu_i)) \cdot \exp(6\cos(2(u_2-m_i))) + \frac{1}{2\pi} \cdot 2\exp(50\cos(2(u_2-m_3)))$	table 7
$\mathbb{T}^2$	$\sum_{i=1}^{3} \exp(2\cos(u_1-\mu_i)) \cdot \exp(2\cos(u_2-m_i))$	table 8
	$\frac{1}{2\pi}\exp(2\cos(u_1+u_2-1.94))$	
$\mathbb{H}^2$	$2\exp\left(-\frac{u_1}{2}\right)\frac{1}{2\pi}$	
	$\frac{1}{2}\exp(6\cos(u_2-u_1-\pi))$	
thin spiral	$\frac{1}{0.3}\exp(0.3z)$	
swiss roll	$\sum_{i=1}^{3} \exp(\kappa\cos(2\pi u_1-\mu_i)) \cdot \exp(\kappa\cos(2\pi u_2-\mu_i))$	table 9
	$\frac{1}{2\pi}\cdot 2\pi\exp(\kappa\cos(2\pi(u_2-u_1)))$	
$(\mathbb{HS})^2$	$\sum_{i=1}^{3} \exp(\kappa\cos(u_1-\mu_i)) \cdot \exp(\kappa\cos(u_2-\mu_i))$	
	$\frac{1}{2}\exp(0.3\lvert u_1\rvert)\exp(\kappa\cos(u_2-u_1-\pi))$	

Table 5: Latent densities.

i	μ_i	m_i
1	$\frac{\pi}{2}$	$\frac{\pi}{4}$
2	$\frac{\pi}{2}$	$\frac{3\pi}{4}$
3	$\frac{3\pi}{2}$	$\frac{\pi}{4}$
4	$\frac{3\pi}{2}$	$\frac{3\pi}{3}$

Table 6: Mixture parameters of von Mises on $\mathbb{S}^2$.

i	μ_i	m_i
1	0	$\frac{\pi}{2}$
2	π	$\frac{3\pi}{2}$

Table 7: Mixture parameters of von Mises on $\mathbb{S}^2$.

i	μ_i	m_i
1	0.21	2.85
2	1.89	6.18
3	3.77	1.56

Table 8: Mixture parameters of von Mises on $\mathbb{T}^2$.

i	μ_i	m_i
1	0.1	0.1
2	0.5	0.8
3	0.8	0.8

Table 9: Mixture parameters of von Mises on swiss roll.

B.3 Density estimation on MNIST digit 1

For fair comparison, we tried to use the same architectures for the IID and $\mathcal{M}-$flow. As the latter requires to invert the flow during training, we have used rational-quadratic splines[12] for the flows which can be efficiently inverted, see Durkan et al. (2019) and table 10. Note that the $\mathcal{M}-$flow learns the latent density using an additional flow f_ϕ after learning to reconstruct the data using f_ψ, see Brehmer and Cranmer (2020).

	flow f_ψ		flow h_ϕ		
model	# couplings	coupling type	# couplings	coupling type	#paramters
IID	10	spline with $B = 11$,$K = 10$	-	-	14M
$\mathcal{M}$-flow	10	spline with $B = 11$,$K = 10$	10	spline with $B = 11$,$K = 10$	14.3M

Table 10: Architectures for standard IID and $\mathcal{M}-$flow on MNIST digit 1.

We train on 100 epochs with a batch size of 100, and take the model yielding the best result on the validation set (10% of the training set). We use AdamW optimizer (Loshchilov and Hutter (2017)) and anneal the learning rate to 0 after 100 epochs using a cosine schedule (Loshchilov and Hutter (2016)).

12. An interval $[-B, B]$ is split into K equidistant bins, and on each subinterval, a rational-quadratic spline is defined such that the derivatives are continuous at the boundary points. The parameters of the splines are again outcomes of neural networks. We refer to B as the spline range and K as the bin size in the following. Outside of the interval $[-B, B]$, the transformation is set to the identity.

We apply weight decay with a prefactor of 10^{-6} without dropout. Furthermore, a L_2-regularization on the latent variable with a prefactor of 0.01 was used to stabilize the training.

B.3.1 ADDITIONAL FIGURES

We show the performance of the IID, FOM together with the NID method on the density estimation tasks from section5.2 in figure 8 and 9.

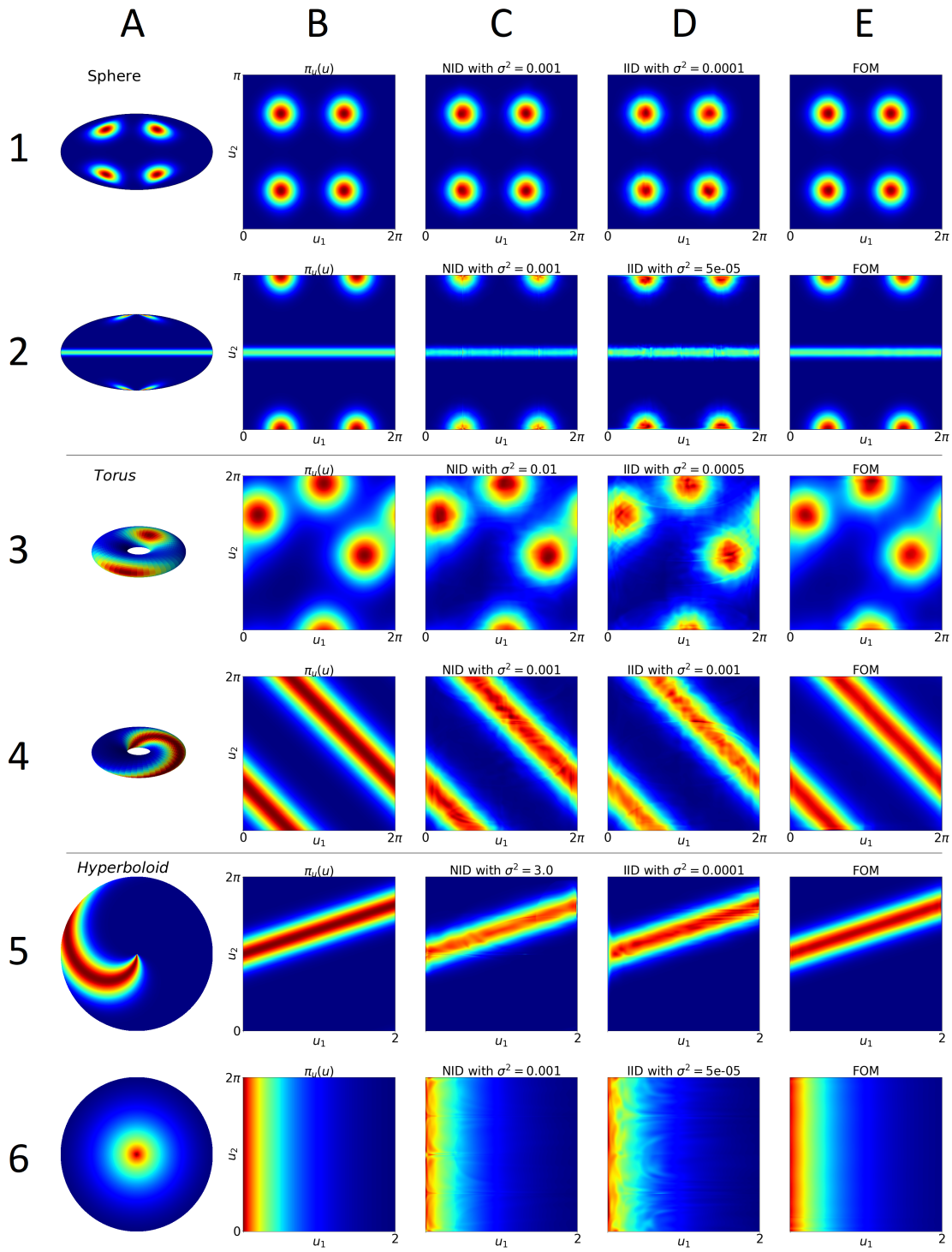

Figure 8: **Columns A and B:** Target density in data (A) and latent space (B) for various manifolds and different latent distributions. **Column C:** Best learned density using our method with NID. **Column D:** Best learned density using our method with IID. **Column D:** Learned density using FOM.

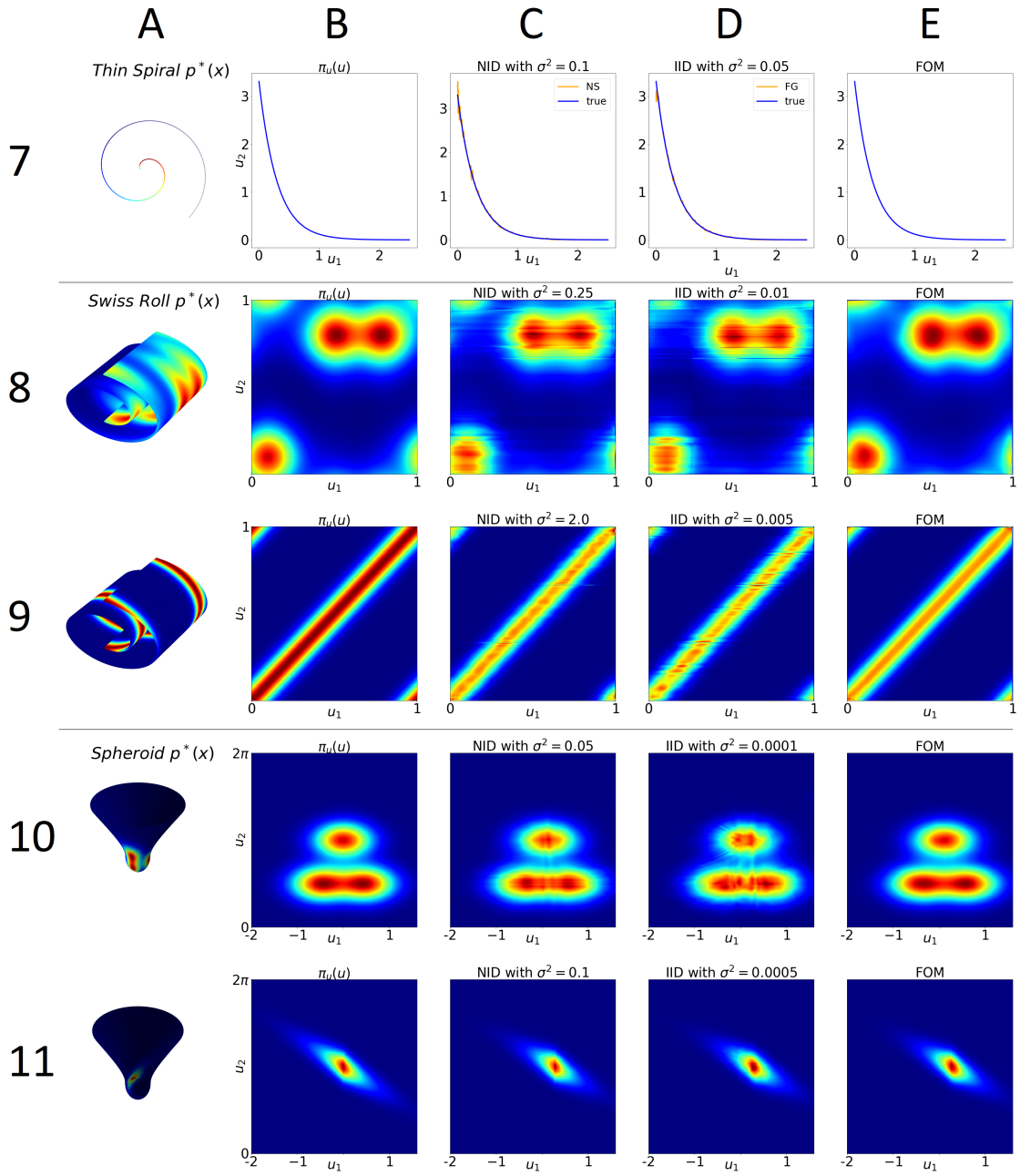

Figure 9: **Columns A and B:** Target density in data (A) and latent space (B) for various manifolds and different latent distributions. **Column C:** Best learned density using our method with NID. **Column D:** Best learned density using our method with IID. **Column D:** Learned density using FOM.

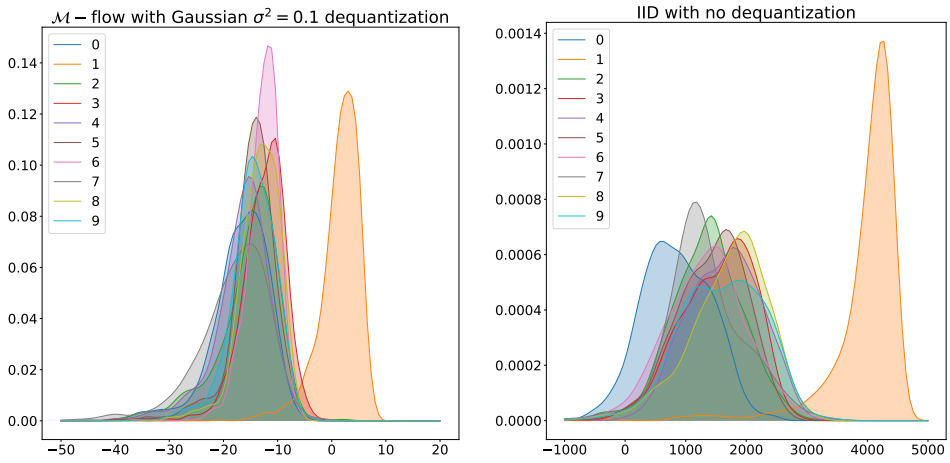

Figure 10: Log likelihoods on various MNIST test digits using $\mathcal{M}-$flow (**left**) and IID (**right**) trained on digit 1 only.

B.4 Manifold Flow for the mixture of von Mises distributions on S^2

In this subsection, we apply the manifold flow, see section4, on a mixture of von Mises distributions on a sphere. We do not attempt to find the optimal hyperparameters and training settings (such as batch- and training size, optimization method, or training scheduler) to maximize the performance.

The manifold flow (MF) proposed by Brehmer and Cranmer, 2020 uses two flows, one for encoding the data manifold to the latent space, and another for learning the latent density. To avoid calculating the Gram determinant of the encoding flow, they proposed different training procedures, an alternating, and a sequential (see section4 for more details). In figure 11, we show that both methods learn the density reasonable good (top left and right). However, if we add Gaussian noise with magnitude 0.01 to the dataset, the two training schemes lead to very different results (bottom left and right). This illustrates the drawback of not having a unified maximum likelihood objective. We used the same model and training settings as for a similar dataset (a two-dimensional manifold embedded in $\mathbb{R}^3$) studied in Brehmer and Cranmer, 2020.

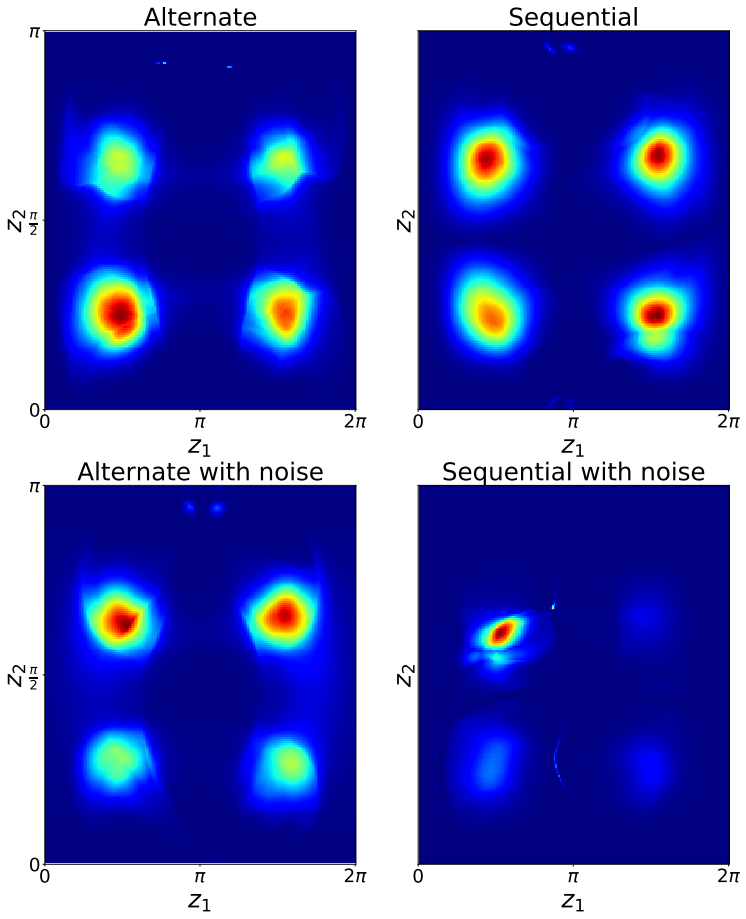

Figure 11: Performance of MF on the mixture of von Mises distributions on a sphere (top) and noisy sphere (bottom), using different training schemes (alternating left, and sequential right).

References

Chi Au and Judy Tam. Transforming variables using the dirac generalized function. *The American Statistician*, 53(3):270–272, 1999. doi: 10.1080/00031305.1999.10474472. URL `https://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474472`.

Jan Jetze Beitler, Ivan Sosnovik, and Arnold Smeulders. Pie: Pseudo-invertible encoder. 2018.

David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

Clément Berenfeld and Marc Hoffmann. Density estimation on an unknown submanifold. *arXiv preprint arXiv:1910.08477*, 2019.

Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

Edmond Cunningham, Renos Zabounidis, Abhinav Agrawal, Ina Fiterau, and Daniel Sheldon. Normalizing flows across dimensions. *arXiv preprint arXiv:2006.13070*, 2020.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Block neural autoregressive flow. In *Uncertainty in Artificial Intelligence*, pages 1263–1273. PMLR, 2020.

Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, pages 7511–7522, 2019.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Wendelin Feiten, Muriel Lang, and Sandra Hirche. Rigid motion estimation using mixtures of projected gaussians. In *Proceedings of the 16th International Conference on Information Fusion*, pages 1465–1472. IEEE, 2013.

Robert J Geller. Earthquake prediction: a critical review. *Geophysical Journal International*, 131 (3):425–450, 1997.

Mevlana C Gemici, Danilo Rezende, and Shakir Mohamed. Normalizing flows on riemannian manifolds. *arXiv preprint arXiv:1611.02304*, 2016.

Thomas Hamelryck, John T Kent, and Anders Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Comput Biol*, 2(9):e131, 2006.

Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 289–296, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102388. URL https://doi.org/10.1145/1102351.1102388.

Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. volume 80 of *Proceedings of Machine Learning Research*, pages 2078–2087, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Chin-Wei Huang, Laurent Dinh, and Aaron Courville. Augmented normalizing flows: Bridging the gap between generative flows and latent variable models. *arXiv preprint arXiv:2002.07101*, 2020.

Hyeongju Kim, Hyeonseung Lee, Woo Hyun Kang, Joun Yeop Lee, and Nam Soo Kim. Softflow: Probabilistic framework for normalizing flow on manifolds. *Advances in Neural Information Processing Systems*, 33, 2020.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

J.M. Lee. *Introduction to Riemannian Manifolds*. Graduate Texts in Mathematics. Springer International Publishing, 2019.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser-Nam Lim, and Christopher De Sa. Neural manifold ordinary differential equations. *arXiv preprint arXiv:2006.10254*, 2020.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *arXiv preprint arXiv:2006.10605*, 2020.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR, 2015.

Danilo Jimenez Rezende, George Papamakarios, Sébastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pages 8083–8092. PMLR, 2020.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL `http://jmlr.org/papers/v15/srivastava14a.html`.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL `https://doi.org/10.1145/1390156.1390294`.

Han Zhang, Xi Gao, Jacob Unterman, and Tom Arodz. Approximation capabilities of neural ODEs and invertible residual networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11086–11095. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/zhang20h.html`.