A. Density Transformations on Manifolds

In this section, we explain how to update the density of a distribution transformed from one Riemannian manifold to another by a smooth map. We only consider the case where both manifolds are sub-manifolds of Euclidean spaces.

Let $\mathcal{M}$ and $\mathcal{N}$ be D-dimensional manifolds embedded into Euclidean spaces $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. For example, $\mathcal{M}$ and $\mathcal{N}$ could be $\mathbb{S}^D$ embedded in $\mathbb{R}^{D+1}$ as in Section 2.3. Both manifolds inherit a Riemannian metric from their embedding spaces. Let T be a smooth injective map $T : \mathcal{M} \to \mathcal{N}$. We will assume that T can be extended to a smooth map between open neighbourhoods of the embedding spaces that contain $\mathcal{M}$ and $\mathcal{N}$, and that we have chosen such an extension. For example, the exponential-map flow in Equation (19) can be written using the coordinates of the embedding space $\mathbb{R}^{D+1}$, and can thus be extended to open neighbourhoods of the embedding spaces as desired.

In what follows, we will use the fact that if $u_1, \ldots, u_D$ are vectors in $\mathbb{R}^n$, then the volume of the parallelepiped with sides $u_1, \ldots, u_D$ is $\sqrt{\det(U^\top U)}$, where U is the $n \times D$ matrix with column vectors $u_1, \ldots, u_D$. If $u_1, \ldots, u_D$ form an orthonormal system, this volume is 1.

Let $\pi : \mathcal{M} \to \mathbb{R}^+$ be a density on $\mathcal{M}$. This defines a distribution on $\mathcal{M}$ and we can use T to transform it into a distribution on $\mathcal{N}$. Let $p : \mathcal{N} \to \mathbb{R}^+$ be the density of the transformed distribution. We are interested in computing p assuming we know π. Let x be a point on $\mathcal{M}$, and $e_1, \ldots, e_D$ be an orthonormal basis of the tangent space $\mathrm{T}_x\mathcal{M}$. Define E to be the $m \times D$ matrix with i-th column vector e_i. Let J be the $n \times m$ Jacobian of T, where T is seen as a map between open sets in $\mathbb{R}^m$ and $\mathbb{R}^n$. The tangent map of T at x transforms each e_i to Je_i, and the matrix that collects all transformed vectors in its columns is JE. Hence, the volume of the parallelepiped with sides the transformed vectors is $\sqrt{\det((JE)^\top JE)} = \sqrt{\det(E^\top J^\top JE)}$. Therefore, the density p is given by

$$p(T(x)) = \frac{\pi(x)}{\sqrt{\det(E^\top J^\top JE)}}. \tag{24}$$

In the special case where $\mathcal{M} = \mathcal{N} = \mathbb{R}^D$ and $m = n = D$, the matrix E is an orthogonal matrix, and the above reduces to the familiar density update in Equation (2).

A.1. The case of $T_{c\to s} : \mathbb{S}^{D-1} \times [-1, 1] \to \mathbb{S}^D$

In this section, we specialize to $\mathcal{M} = \mathbb{S}^{D-1} \times (-1, 1)$ and $\mathcal{N} = \mathbb{S}^D$ with $D \geq 2$. In particular, we will prove:

Proposition 1. *Let π be a density $\pi : \mathbb{S}^{D-1} \times (-1, 1) \to \mathbb{R}^+$. Let $p : \mathbb{S}^D \to \mathbb{R}^+$ be the density of the transformed distribution under $T_{c\to s}$. Then:*

$$p(T_{c\to s}(z, r)) = \frac{\pi(z, r)}{(1 - r^2)^{\frac{D}{2}-1}}. \tag{25}$$

Proof. The sphere $\mathbb{S}^{D-1}$ is embedded in $\mathbb{R}^D$. This gives us an embedding of $\mathcal{M}$ in $\mathbb{R}^{D+1}$. The map $T_{c\to s}$, introduced in Equation (14), is easily extended to a map $\mathbb{R}^D \times (-1, 1) \to \mathbb{R}^{D+1}$ using the same formula $T_{c\to s}(z, r) = (\sqrt{1 - r^2}z, r)$. Its Jacobian is an upper triangular $D + 1$ by $D + 1$ matrix

$$J = \begin{bmatrix} \sqrt{1-r^2} & 0 & \cdots & \frac{-x_1 r}{\sqrt{1-r^2}} \\ 0 & \sqrt{1-r^2} & & \frac{-x_2 r}{\sqrt{1-r^2}} \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 1 \end{bmatrix}.$$

We will use a symmetry argument to simplify the computation of the determinant in Equation (24). Let G be a rotation of $\mathbb{R}^{D+1}$ that leaves the last coordinate invariant.[1] Note that for any point $x \in \mathbb{R}^{D+1}$, we have $T_{c\to s}(Gx) = GT_{c\to s}(x)$. This means that the Jacobian J transforms as a function of x as $J(Gx) = GJ(x)G^\top$. Note also that if x is in $\mathcal{M}$, and $E(x)$ is a matrix where the column vectors form a basis of the tangent space at x, then Gx is also in $\mathcal{M}$, and $GE(x)$ is a matrix where the column vectors form a basis of the tangent space at Gx. So we can choose $E(Gx) = GE(x)$. With that choice

$$\begin{aligned} &\det\big(E(Gx)^\top J(Gx)^\top J(Gx)E(Gx)\big) \\ &= \det\big(E(x)^\top G^\top GJ(x)^\top G^\top GJ(x)G^\top GE(x)\big) \\ &= \det\big(E(x)J(x)^\top J(x)E(x)\big). \end{aligned}$$

Since for any $x \in \mathcal{M}$, we can always choose G such that Gx is of the form $(\sqrt{1 - r^2}, 0, \ldots, 0, r)$, we can restrict ourselves to this case. For such a choice, the Jacobian simplifies to

$$J = \begin{bmatrix} \sqrt{1-r^2} & 0 & \cdots & \frac{-r}{\sqrt{1-r^2}} \\ 0 & \sqrt{1-r^2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 1 \end{bmatrix}.$$

For E, we can simply choose the $D + 1$ by D matrix made by removing the first column from the identity matrix. Then JE is equal to J with the first column removed:

$$JE = \begin{bmatrix} 0 & 0 & \cdots & \frac{-r}{\sqrt{1-r^2}} \\ \sqrt{1-r^2} & 0 & & 0 \\ 0 & \sqrt{1-r^2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 1 \end{bmatrix}.$$

The product $(JE)^\top JE$ is simply a diagonal matrix of size D with diagonal $(1 - r^2, \ldots, 1 - r^2, \frac{1}{1-r^2})$. Taking the determinant and applying Equation (24) concludes the proof. □

[1] The set of such rotations is the group of rotations of $\mathbb{R}^D$ embedded in $\mathbb{R}^{D+1}$.

At first, Proposition 1 might seem worrying since the density ratio in that proposition vanishes when r is -1 or 1. So, as r approaches the boundary of the interval $[-1,1]$, it seems that the correction term to the density will tend to infinity and lead to numerical instability.

What saves us is that we do not use $T_{c\to s}$ on its own, and instead combine it with a particular flow transformation on $\mathbb{S}^{D-1}\times[-1,1]$ and the inverse $T_{s\to c}$, as shown in Equations (12) to (14). In these formulas, the map g is a spline on the interval $[-1,1]$ which maps -1 to -1, 1 to 1, and has strictly positive slopes $g'(-1)$ and $g'(1)$. Looking only at -1 (the case 1 can be similarly dealt with), this means

$$g(-1+\epsilon)\approx -1+g'(-1)\epsilon.$$

As ϵ goes to 0, the density corrections coming from $T_{c\to s}$ and $T_{s\to c}$ combine to

$$\left(\frac{D}{2}-1\right)\log\frac{1-g(-1+\epsilon)^2}{1-(-1+\epsilon)^2},$$

which is equivalent to

$$\left(\frac{D}{2}-1\right)\log g'(-1)$$

as ϵ goes to 0. In particular, the terms that tend to infinity cancel each other, and the flow is well-behaved. When implementing the flow, numerical stability is achieved by not adding the terms that cancel each other. Finally, we note a subtle point about what we proved: the sequence of transformations $T_{c\to s}\circ T_{c\to c}\circ T_{s\to c}$ will transform a distribution with finite density into another distribution with finite density, but we do not guarantee that the resulting density will be continuous.

B. Detailed Diagram of Recursive Flow on $\mathbb{S}^D$

In Figure 6, we provide an illustration of the recursive construction in Equations (12) to (14), showing the specific wiring order of the conditional maps inside the flow. This order is the one implied by the recursion. In general, any other order can be used, or a composition of autoregressive flows with multiple orders.

C. Examples of Möbius Transformations

To illustrate the kind of densities we get on $\mathbb{S}^1$ using the Möbius flows, we show a few random examples in Figure 7.

D. Fourier Transformations on $\mathbb{S}^1$

Another family of circle transformations that we considered are *Fourier transformations*, defined by

$$f_{\alpha,\phi,w}(\theta)=\theta+\sum_i\frac{\alpha_i}{w_i}\sin(w_i\theta-\phi_i)+\mu, \qquad (26)$$

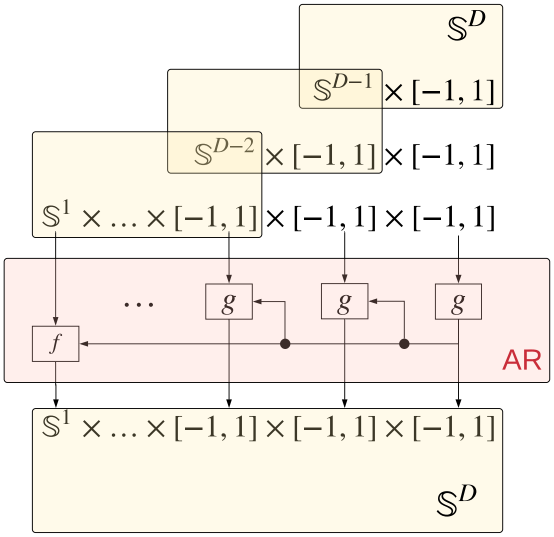
$\mathbb{S}^D$
$\mathbb{S}^{D-1}$ × [−1, 1]
$\mathbb{S}^{D-2}$ × [−1, 1] × [−1, 1]
$\mathbb{S}^1$ × … × [−1, 1] × [−1, 1] × [−1, 1]
…
g
g
g
f
AR
$\mathbb{S}^1$ × … × [−1, 1] × [−1, 1] × [−1, 1]
$\mathbb{S}^D$

Figure 6. Detailed illustration of the recursive flow on the sphere $\mathbb{S}^D$ showing the explicit wiring of the conditional flows. The sphere $\mathbb{S}^D$ is recursively transformed to the cylinder $\mathbb{S}^1\times[-1,1]^{D-1}$, then an autoregressive flow is applied to the cylinder, and finally the cylinder is transformed back to the sphere.

where $\mu=\sum_i\alpha_i\sin(\phi_i)$, $w_i\in\mathbb{Z}$, $\phi_i\in[0,2\pi]$ and $\sum_i|\alpha_i|\le 1$. The integers w_i are fixed frequencies in the Fourier basis.

We found empirically that this family of transformations is not competitive with the other transformations considered in this paper, especially for highly concentrated densities as shown in Figure 8.

E. Polynomial Exponential Map

The polynomial exponential map of Sei (2013) is the exponential-map flow built using the scalar field

$$\phi(x)=\mu^\top x+x^\top Ax, \qquad (27)$$

where $x\in\mathbb{S}^D$ in the coordinates of the embedding space $\mathbb{R}^{D+1}$. The parameters μ and A must satisfy the constraint $\|\mu\|_1+\|A\|_1\le 1$ where $\|\cdot\|_1$ is the elementwise ℓ_1 norm.

F. Target Densities Used in Experiments

For the experiments on the torus $\mathbb{T}^2$, we used targets built from densities in the von Mises family as shown in Table 2.

On the sphere $\mathbb{S}^2$, the target was a mixture of the form

$$p(x)\propto\sum_{k=1}^{4}e^{10x^\top T_{s\to e}(\mu_k)}, \qquad (28)$$

where $\mu_1=(0.7,1.5)$, $\mu_2=(-1,1)$, $\mu_3=(0.6,0.5)$, $\mu_4=(-0.7,4)$, $T_{s\to e}$ maps from spherical to Euclidean

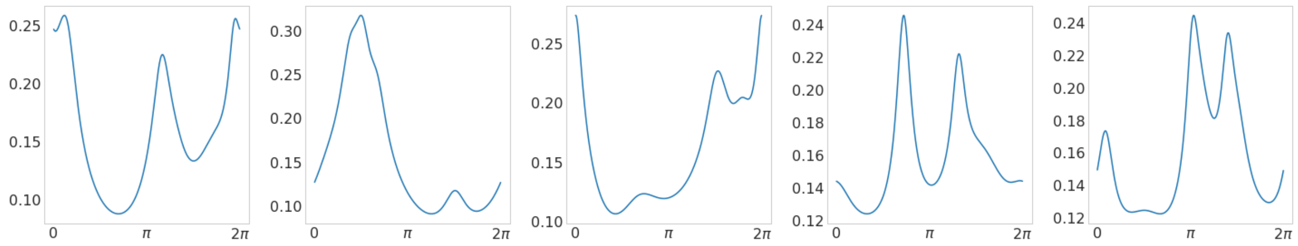

Figure 7. Probability density functions of convex combinations of 15 Möbius transformations applied to a uniform base distribution on the circle $\mathbb{S}^1$. Each of these distributions required $30 = 15 \times 2$ parameters.

Target	**Expression**	**Parameters**
Unimodal	$p_A(\theta_1, \theta_2) \propto \exp[\cos(\theta_1 - \phi_1) + \cos(\theta_2 - \phi_2)]$	$\phi = (4.18, 5.96)$
Multi-modal	$p_B(\theta_1, \theta_2) \propto \frac{1}{3}\sum_{i=1}^{3} p_A(\theta_1, \theta_2; \phi_i)$	$\phi = \{(0.21, 2.85), (1.89, 6.18), (3.77, 1.56)\}$
Correlated	$p_C(\theta_1, \theta_2) \propto \exp[\cos(\theta_1 + \theta_2 - \phi)]$	$\phi = 1.94$

Table 2. Target densities used for experiments on $\mathbb{T}^2$.

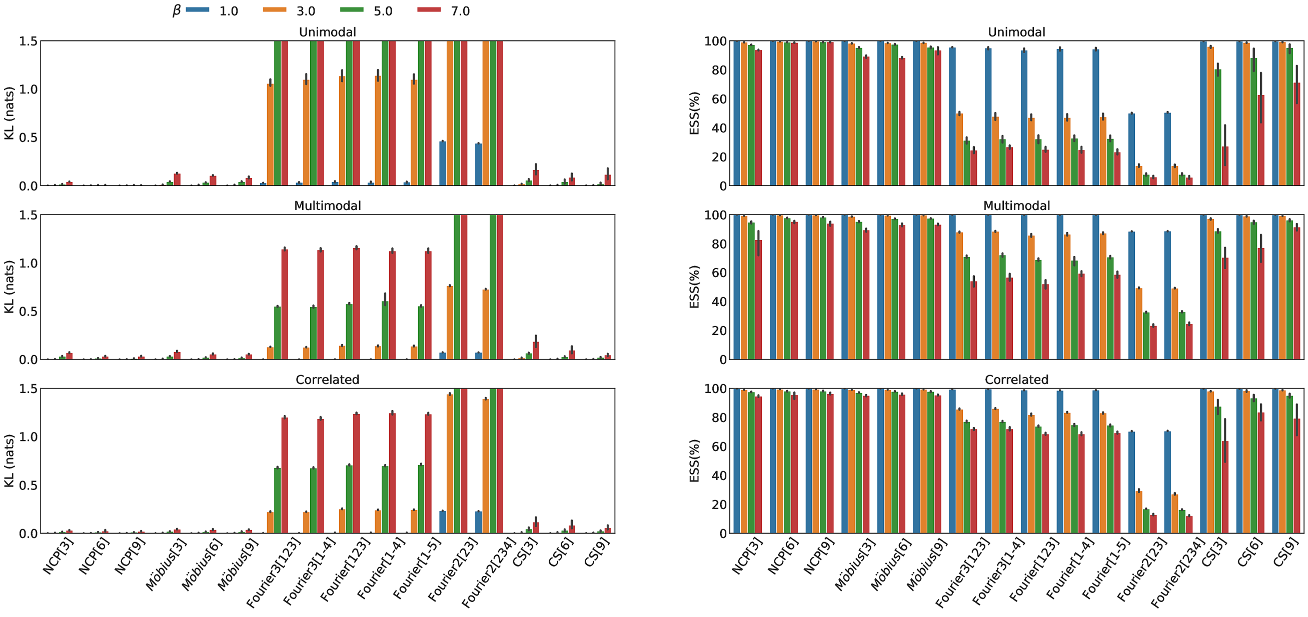

Figure 8. Same as Figure 2, with KL and values for the Fourier transforms added. For the Fourier models, the numbers between brackets represent used frequencies, and a number before the bracket means each frequency was repeated. For example, Fourier3$[1 - 4]$ is a Fourier model with 12 frequencies: 3 frequencies of k for each $k = 1, \ldots, 4$.

coordinates, and $x \in \mathbb{R}^3$ is a point on the embedded sphere in Euclidean coordinates.

On $\mathrm{SU}(2) \cong \mathbb{S}^3$, the target was a mixture of the same form where $\mu_1 = (1.7, -1.5, 2.3)$, $\mu_2 = (-3.0, 1.0, 3.0)$, $\mu_3 = (0.6, -2.6, 4.5)$, $\mu_4 = (-2.5, 3.0, 5.0)$, and $x \in \mathbb{R}^4$ is a point on the embedded sphere in Euclidean coordinates.

G. Misaligned Density on $\mathbb{S}^2$

The recursive formulas shown in Equations (12) to (14) require choosing a sequence of axes in order to construct the cylindrical coordinate system. This may introduce artifacts to the density related to this choice of axes. To test if this results in numerical problems, we compare the flow from Equations (12) to (14) on a target density that forms a non-axis-aligned ring against a composition of the same flow with a learned rotation.

The results of this experiment are shown in Figure 9. We compared both large ($K_s = 32$, $K_m = 12$) and small ($K_s = 3$, $K_m = 3$) versions of the auto-regressive Möbius-Spline flow and observed no significant differences between the two models on $\mathbb{S}^2$.

More experiments would be necessary to investigate this potential effect in higher dimensions.

H. NCP as a complex Möbius transformation

For a general Möbius transformation

$$f(z) = \frac{az+b}{cz+d}, \tag{29}$$

where $a, b, c, d, z \in \mathbb{C}$, to define a diffeomorphism on $\mathbb{S}^1$ it must be constrained to be of the form

$$h(z) = \frac{z-a}{1-\bar{a}z}. \tag{30}$$

This form ensures that $h(z)\bar{h}(z) = 1$ if $z\bar{z} = 1$ and has two real-valued free parameters $\Re(a)$ and $\Im(a)$.

In what follows we show that for the choice $\Im(a) = 0$ and $\Re(a) = -\frac{1-\alpha}{1+\alpha}$, the transformation $h(z)$ is equivalent to an NCP transform $w = 2\arctan\left(\alpha \tan\left(\frac{\theta}{2}\right) + \beta\right)$ with scale parameter α and offset parameter $\beta = 0$ (assuming $w, \theta \in (-\pi, \pi)$). If we define θ via $z = e^{i\theta}$, the goal is to show that w defined via

$$e^{iw} = \frac{e^{i\theta} + \frac{1-\alpha}{1+\alpha}}{1 + \frac{1-\alpha}{1+\alpha}e^{i\theta}}, \tag{31}$$

follows the NCP transformation rule

$$w = 2\arctan\left(\alpha \tan\left(\frac{\theta}{2}\right)\right) \mod 2\pi.$$

We begin by expanding Equation (31) in terms of more basic trigonometric quantities,

$$\begin{aligned}
e^{iw} &= \frac{e^{i\theta} + \frac{1-\alpha}{1+\alpha}}{1 + \frac{1-\alpha}{1+\alpha}e^{i\theta}} \\
&= \frac{e^{i\theta} + \frac{1-\alpha}{1+\alpha}}{1 + \frac{1-\alpha}{1+\alpha}e^{i\theta}} \frac{1 + \frac{1-\alpha}{1+\alpha}e^{-i\theta}}{1 + \frac{1-\alpha}{1+\alpha}e^{-i\theta}} \\
&= \frac{e^{i\theta} + 2\frac{1-\alpha}{1+\alpha} + \left(\frac{1-\alpha}{1+\alpha}\right)^2 e^{-i\theta}}{\frac{2(\cos(\theta)+\alpha^2(-\cos(\theta))+\alpha^2+1)}{(\alpha+1)^2}} \\
&= \frac{\left(e^{i\frac{\theta}{2}} + \frac{1-\alpha}{1+\alpha}e^{-i\frac{\theta}{2}}\right)^2}{\frac{2(\cos(\theta)+\alpha^2(-\cos(\theta))+\alpha^2+1)}{(\alpha+1)^2}} \\
&= \frac{1}{2}\frac{\left((1+\alpha)e^{i\frac{\theta}{2}} + (1-\alpha)e^{-i\frac{\theta}{2}}\right)^2}{\cos(\theta) + \alpha^2(-\cos(\theta)) + \alpha^2 + 1} \\
&= \frac{2\left(\cos\left(\frac{\theta}{2}\right) + i\alpha\sin\left(\frac{\theta}{2}\right)\right)^2}{\cos(\theta) + \alpha^2(-\cos(\theta)) + \alpha^2 + 1}.
\end{aligned}$$

In order to isolate w, only the numerator $v = 2\left(\cos\left(\frac{\theta}{2}\right) + i\alpha\sin\left(\frac{\theta}{2}\right)\right)^2$ of the expression above matters as we are only interested in ratios of the imaginary and real parts of this expression, $\tan(w) = \frac{\Im(v)}{\Re(v)}$. The numerator can be expanded as

$$\begin{aligned}
\frac{2\left(\cos\left(\frac{\theta}{2}\right) + i\alpha\sin\left(\frac{\theta}{2}\right)\right)^2}{2\cos^2\left(\frac{\theta}{2}\right)} &= -\alpha^2\tan^2\left(\frac{\theta}{2}\right) \\
&\quad + 2\alpha\tan\left(\frac{\theta}{2}\right)\mathbf{i} + 1,
\end{aligned}$$

from which we conclude,

$$\tan(w) = \frac{\Im(v)}{\Re(v)} = \frac{2\alpha\tan\left(\frac{\theta}{2}\right)}{1 - \alpha^2\tan^2\left(\frac{\theta}{2}\right)}. \tag{32}$$

Using the trigonometric formula $\tan(2x) = \frac{2\tan(x)}{1-\tan(x)^2}$, we arrive at the final result

$$\begin{aligned}
&\tan\left(\frac{w}{2}\right) = \alpha\tan\left(\frac{\theta}{2}\right) \\
&\Leftrightarrow \\
&w = 2\arctan\left(\alpha\tan\left(\frac{\theta}{2}\right)\right) \mod 2\pi.
\end{aligned}$$

I. Application: Multi-Link Robot Arm

As a concrete application of flows on tori, we consider the problem of approximating the posterior density over joint angles $\theta_{1,\ldots,6}$ of a 6-link 2D robot arm, given (soft) constraints on the position of the tip of the arm. The possible configurations of this arm are points in $\mathbb{T}^6$. The position r_k of a joint $k = 1, \dots, 6$ of the robot arm is given by

$$r_k = r_{k-1} + \left(l_k \cos\left(\sum_{j \le k} \theta_j \right), l_k \sin\left(\sum_{j \le k} \theta_j \right) \right),$$

where $r_0 = (0, 0)$ is the position where the arm is affixed, $l_k = 0.2$ is the length of the k-th link, and θ_k is the angle of the k-th link in a local reference frame. The constraint on the position of the tip of the arm, r_6, is expressed in the form of a Gaussian-mixture likelihood $p(r_6 \,|\, \theta_{1,\dots,6})$ with two components. The prior $p(\theta_{1,\dots,6})$ is taken to be a uniform distribution on $\mathbb{T}^6$. The experimental results are illustrated in Figure 10.

J. Application: Learning from samples

In most of the experiments shown on this paper, we trained the models to fit a target density known up to a normalization constant (i.e. an inference problem). In this experiment we train our flow directly on data samples instead.

Training a flow-based model from data samples via maximum likelihood requires an explicit computation of the inverse map as shown in Equation (2). To demonstrate this is feasible with data coming from a non-trivial target density on the sphere $\mathbb{S}^2$ (i.e. that would require a large number of mixture components from simpler densities such as von Mises), we created a dataset of samples on the sphere coming from a density shaped as Earth's continental map as shown in Figure 11 (left).

We trained a flow built from stacking two autoregressive flows. Each flow in the stack used circular splines and standard splines on the interval. The model was trained to maximize the likelihood of the dataset for 100,000 training steps. Both splines used $K_s = 80$ segments. The neural networks producing the spline parameters are the same as for the other experiments. In Figure 11 (middle) we show samples from the learned model overlaid on Earth's map and in Figure 11 (right) we show a heat map of the learned density.

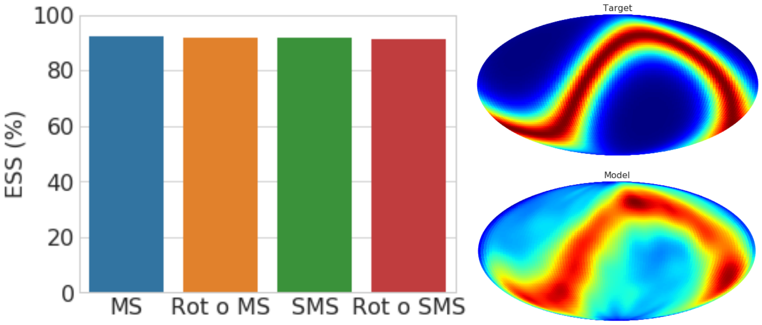

Figure 9. Learning a non-axis-aligned density on $\mathbb{S}^2$ using Equations (12) to (14) with and without composing with a learnable rotation. We compare Möbius-spline flow (MS) ($K_s = 32$, $K_m = 12$), learnable rotation composed with MS (Rot $\circ$ MS), small MS ($K_s = 3$, $K_m = 3$) (SMS) and learnable rotation composed with SMS (Rot $\circ$ SMS). We observed no substantial differences between these models, suggesting that the particular choice of axis inside the flow has no impact on performance.

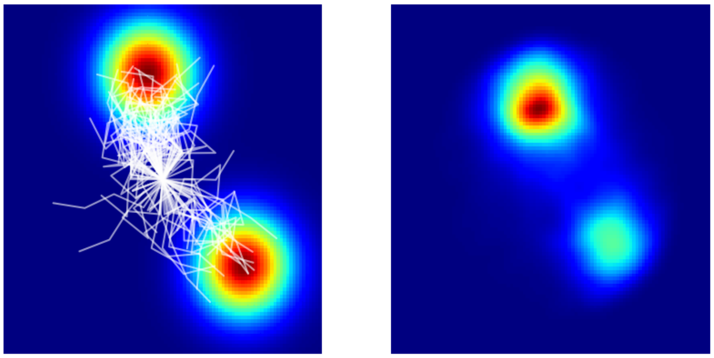

Figure 10. Learning the posterior density over joint angles of a 6-link 2D robot arm. We used an autoregressive Möbius flow on the torus $\mathbb{T}^6$ to approximate the posterior density of joint angles resulting in a Gaussian mixture density for the tip of the robot arm. **Left**: The heat map shows the target density for the tip of the robot arm, a Gaussian mixture with two modes with centres at $(-0.5, 0.5)$ and $(0.6, -0.1)$. White paths show arm configurations sampled from the learned model in angle space converted to position space. **Right**: Density of the tips of the robot arm using samples from the learned model.

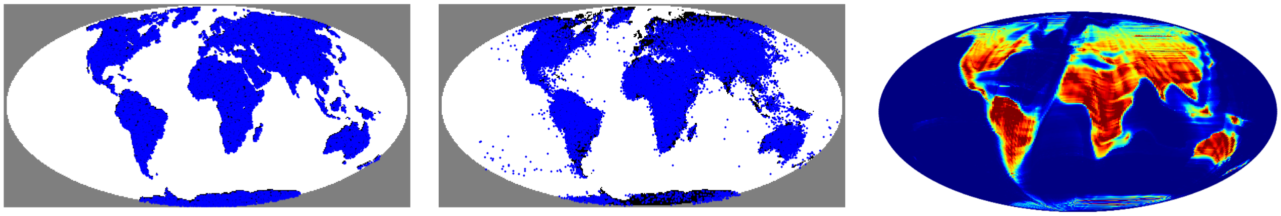

Figure 11. Learning a complex density from data samples using autoregressive spline flows. **Left**: Target density from which i.i.d. data samples were generated. **Middle**: Model samples overlaid on target density; **Right**: Heat map of the learned density.