Justification for the used gradient in the curved von Mises Fisher distribution experiments

General smooth curves on the sphere

We define the log density of the curved von Mises Fisher distribution as

$$\log p(x) = \kappa \max_{t\in[0,1]} \{x^\top \mu(t)\}$$

where $\mu : [0,1] \to \mathbb{S}^{d-1}$ is a smooth curve on the sphere. We have

$$\log p(x) = \kappa\, x^\top \mu(t^*(x))$$

where $t^*(x) := \arg\max_{t\in[0,1]} \{x^\top \mu(t)\} = \arg\min_{t\in[0,1]} \|x - \mu(t)\|^2$ determines the orthogonal projection of x onto μ. Then, the gradient of $\log p$ is

$$\nabla_x \log p(x) = \kappa\, \Big(\mu(t^*(x)) + x^\top \mu'(t^*(x))\, \nabla_x t^*(x)\Big)\,.$$

Because $t^*(x)$ is the maximizer of $x^\top\mu(t)$, it follows that $x^\top\mu'(t^*(x)) = 0$. This is intuitively clear, since $t^*(x)$ determines the orthogonal projection of x onto the curve. The gradient simplifies to

$$\nabla_x \log p(x) = \kappa\, \mu(t^*(x))\,.$$

A problem arises at locations $x \in \mathbb{S}^{d-1}$ where the orthogonal projection is not uniquely defined. At these locations, the log density and its gradient remain undefined. However, for slerps this occurs only on a nullset.

Spherical linear interpolation

Let us now look specifically at a curve μ given as spherical linear interpolation (slerp) between two points $a, b \in \mathbb{S}^{d-1}$:

$$\mu(t) = \mathrm{slerp}_{(a,b)}(t) := \frac{\sin(\theta(1-t))}{\sin(\theta)}\, a + \frac{\sin(\theta t)}{\sin(\theta)}\, b \tag{1}$$

where $t \in [0,1]$ and $\theta = \arccos(a^\top b)$. We have $\mathrm{slerp}_{(a,b)}(0) = a$ and $\mathrm{slerp}_{(a,b)}(1) = b$. The (continuated) slerp can be rewritten by using the projection of b onto $\mathbb{S}^{d-2}_a$:

$$\mu(t) = \cos(\theta t)\, a + \sin(\theta t)\, \frac{b - a^\top b\, a}{\sin(\theta)} = \cos(\theta t)\, a + \sin(\theta t)\, v, \qquad t \in (-\pi/\theta, \pi/\theta)$$

where $v = (I - aa^\top)b / \|(I - aa^\top)b\| \in \mathbb{S}^{d-2}_a$ with $a^\top v = 0$. From $x^\top \mu'(t^*) \overset{!}{=} 0$, we obtain

$$t^*(x) = \theta^{-1} \arctan(x^\top v / x^\top a) \qquad \text{or} \qquad t^*(x) = \theta^{-1}\big(\arctan(x^\top v/x^\top a) + \pi\big),$$

where $x^\top a = 0$ happens only on a null set. We obtain

$$\begin{aligned}
\mathrm{slerp}_{(a,b)}(t^*(x)) &= \frac{aa^\top + vv^\top}{\sqrt{(x^\top a)^2 + (x^\top v)^2}}\, x \\
&= \frac{aa^\top x + v(b^\top x - a^\top b b^\top x)/(\sin\theta)^2}{\sqrt{(x^\top a)^2 + (x^\top b - a^\top b x^\top a)^2/(\sin\theta)^2}} \\
&= \frac{(x^\top b)\,(I - aa^\top)b + (x^\top a)\,(I - bb^\top)a}{\sqrt{(x^\top a)^2 + (x^\top b)^2 - 2(a^\top b)(x^\top a)(x^\top b)}}
\end{aligned}$$

To ensure that $\mu(t^*(x))$ lies on the spherical segment from a to b, we restrict $t^*(x)$ to the interval $[0,1]$ and make the approximation

$$\begin{aligned}\mu(t^*(x)) &\approx \mathrm{slerp}_{(a,b)}(\min\{1, \max\{0, t^*(x)\}\}) \\ &= \begin{cases} a & \text{if } t^*(x) \le 0, \\ \frac{\sin(\theta(1-t^*(x)))\, a + \sin(\theta t^*(x))\, b}{\sin(\theta)} & \text{if } 0 < t^*(x) < 1, \\ b & \text{if } t^*(x) \ge 1, \end{cases}\end{aligned}$$

where $t^*(x) = \theta^{-1} \arctan(x^\top v / x^\top a)$. For $x \in \mathbb{S}^{d-1}$ with $t^*(x) = 0$ or $t^*(x) = 1$, the log density defined by the slerp is continuous, but not differentiable due to the restriction to $[0,1]$. For all other x we use

$$\nabla_x \log p(x) = \kappa\, \mu(t^*(x))\,.$$

Piecewise linear curve

In our model, we concatenate k slerps defined by $k+1$ points $a_i \in \mathbb{S}^{d-1}, i = 0, 1, \ldots, k$ associated with $t_i \in [0,1]$, $0 = t_0 < t_1 < \cdots < t_{k-1} < t_k = 1$:

$$\mu(t) = \sum_{i=1}^{k} \mathbb{1}_{[t_{i-1}, t_i)}(t)\, \mathrm{slerp}_{(a_{i-1}, a_i)}((t - t_{i-1})/(t_i - t_{i-1}))$$

where $t_i = \sum_{j=1}^{i} \theta_j / \sum_{j'=1}^{k} \theta_{j'}$ and $\theta_i := \arccos({a_{i-1}}^\top a_i)$ for $i = 1, \ldots, k$. In our choice of the “knots” $a_i \in \mathbb{S}^{d-1}$, we ensure that ${a_i}^\top a_{i-1} \neq 1$, i.e. two successive knots do not coincide such that $\theta_i > 0$.

To evaluate $\log p(x)$, we first compute $t_i^*(x)$:

$$t_i^*(x) = t_{i-1} + \frac{t_i - t_{i-1}}{\theta_i} \arctan(({a_i}^\top x - {a_{i-1}}^\top x \cos\theta_i)/({a_{i-1}}^\top x \sin\theta_i))$$

for all k slerps. We then assign x to the nearest slerp indexed by i^*:

$$i^* = \underset{i \in \{1,\ldots,k\}}{\arg\max} \left\{ x^\top \mathrm{slerp}_{(a_{i-1}, a_i)}(\min\{1, \max\{0, (t_i^*(x) - t_{i-1})/(t_i - t_{i-1})\}\}) \right\}\,.$$

We then use as gradient at x

$$\nabla_x \log p(x) = \kappa\, \mathrm{slerp}_{(a_{i^*-1}, a_{i^*})}(\min\{1, \max\{0, (t_{i^*}^*(x) - t_{i^*-1})/(t_{i^*} - t_{i^*-1})\}\}).$$