The Power Spherical distribution

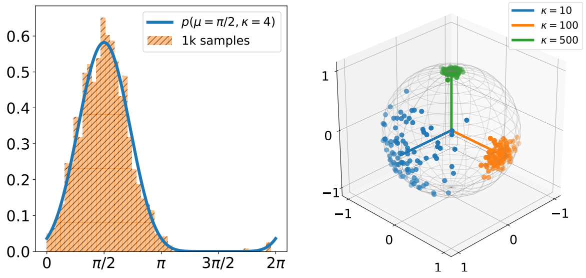
Nicola De Cao [1 2] **Wilker Aziz** [1]

Abstract

There is a growing interest in probabilistic models defined in hyper-spherical spaces, be it to accommodate observed data or latent structure. The von Mises-Fisher (vMF) distribution, often regarded as the Normal distribution on the hyper-sphere, is a standard modeling choice: it is an exponential family and thus enjoys important statistical results, for example, known Kullback-Leibler (KL) divergence from other vMF distributions. Sampling from a vMF distribution, however, requires a rejection sampling procedure which besides being slow poses difficulties in the context of stochastic backpropagation via the *reparameterization trick*. Moreover, this procedure is numerically unstable for certain vMFs, *e.g.*, those with high concentration and/or in high dimensions. We propose a novel distribution, the **Power Spherical** distribution, which retains some of the important aspects of the vMF (*e.g.*, support on the hyper-sphere, symmetry about its mean direction parameter, known KL from other vMF distributions) while addressing its main drawbacks (*i.e.*, scalability and numerical stability). We demonstrate the stability of Power Spherical distributions with a numerical experiment and further apply it to a variational auto-encoder trained on MNIST. Code at: github.com/nicola-decao/power_spherical

1. Introduction

Manifold learning and machine learning applications of directional statistics (Sra, 2018) have spurred interest in distributions defined in non-Euclidean spaces (*e.g.*, simplex, hyper-torus, hyper-sphere). Examples include learning rotations (*i.e.*, $\mathrm{SO}(n)$, Falorsi et al., 2018; 2019) and hierarchical structures on hyperbolic spaces (Mathieu et al., 2019; Nagano et al., 2019). Hyper-spherical distributions, in particular, find applications in clustering (Banerjee et al., 2005; Bijral et al., 2007), mixed-membership models (Reisinger et al., 2010), computer vision (Liu et al., 2017), and natural language processing (Kumar & Tsvetkov, 2019).



(a) On the circle $\mathbb{S}^1$.

(b) On the sphere $\mathbb{S}^2$.

Figure 1. Example of draws and density function of the Power Spherical distribution. For the circle (a) we plot both the density and histograms for 1k samples. For the sphere (b) we plot draws from 3 distributions of orthogonal directions (μ) and different concentration parameter κ.

The von Mises-Fisher distribution (vMF; Mardia & Jupp, 2009) is a natural and standard choice for densities in hyper-spheres. It is a two-parameter exponential family, one parameter being a mean direction and the other a scalar concentration, and it is symmetric about the former. Because of that, it is often regarded as the Normal distribution on spheres. Amongst other useful properties, it has closed-form Kullback–Leibler (KL) divergence with other vMF densities including the uniform distribution, one of its special cases.

Thanks to the *tangent-normal decomposition*, sampling from a vMF, no matter its dimensionality, requires only sampling from a univariate marginal distribution. Unfortunately, the inverse cumulative density function (*cdf*) of this marginal is not known analytically, which prevents straightforward generation of independent samples. Fortunately, a rejection sampling procedure is known (Ulrich, 1984), but unsurprisingly, rejection sampling is inefficient.

In deep learning, an appealing use of the vMF distribution is as a random generator in stochastic and differentiable computation graphs. For that, we need a *reparameterization trick* to enable unbiased and low variance estimates of gradients of samples with respect to the vMF parameters (Rezende et al., 2014; Kingma & Welling, 2014). With rejection sam-

[1] University of Amsterdam [2] The University of Edinburgh. Correspondence to: Nicola De Cao <nicola.decao@uva.nl>.

Second workshop on *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (ICML 2020), Virtual Conference

pling, reparameterization does not come naturally, requiring a correction term which has high variance (Naesseth et al., 2017). This plays against widespread use of vMFs. For example, Davidson et al. (2018) successfully used the vMF distribution to approximate the posterior distribution of a hyper-spherical variational auto-encoder (VAE; Kingma & Welling, 2014), but they had to omit the correction term, trading variance for bias. Additionally, due to its exponential form as well as a dependency on the modified Bessel function of the first kind (Weisstein, 2002), the vMF distribution is numerically unstable in high dimensions or with high concentration (Davidson et al., 2018).

To overcome all of the vMF drawbacks, we propose the novel **Power Spherical distribution**. We start from the *tangent-normal decomposition* of vectors in hyper-spheres and specifically design a univariate marginal distribution that admits an analytical inverse *cdf*. This marginal is used to derive a distribution on hyper-spheres of any dimension. The resulting distribution is not an exponential family and is defined via a power law. Crucially, it is numerically stable and dispenses with rejection sampling for independent sampling. We verify the stability of the Power Spherical distribution with a numerical experiment as well as reproducing some of the experiments of Davidson et al. (2018) while substituting the vMF with our proposed distribution.

Contributions We propose a new distribution defined on any d-dimensional hyper-sphere which

- has closed form marginal *cdf* and inverse *cdf*, thus it does not require rejection sampling;
- is fully reparameterizable without a correction term;
- is numerically stable in high dimensions and/or high concentrations.

2. Method

We start with an overview of the vMF distribution, also introducing results that help formulate the Power Spherical distribution. We then define the Power Spherical and present some of its proprieties such as mean, mode, variance, and entropy as well as its Kullback–Leibler divergence from a vMF and from a uniform distribution.

2.1. Preliminaries

Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ be the hyper-spherical set. A key idea in directional distribution theory is the *tangent-normal decomposition* around another vector $\mu \in \mathbb{S}^{d-1}$ in the same space.

Theorem 1 (9.1.2 in Mardia & Jupp (2009))**.** *Any unit vector $x \in \mathbb{S}^{d-1}$ can be decomposed as*

$$x = \mu t + v(1-t^2)^{\frac{1}{2}} \ , \tag{1}$$

with $t \in [-1, 1]$ and $v \in \mathbb{S}^{d-2}$ a tangent to $\mathbb{S}^{d-1}$ at μ.

Corollary 1 (9.3.1 in Mardia & Jupp (2009))**.** *The intersection of $\mathbb{S}^{d-1}$ with the plane through $t\mu$ and normal to μ is a $(d-2)$-sphere of radius $\sqrt{1-t^2}$. Moreover, t has density*

$$p_T(t; d) \propto \left(1-t^2\right)^{\frac{d-3}{2}} \quad \textit{with} \quad t \in [-1, 1] \ . \tag{2}$$

Importantly, it follows from Theorem 1 and Corollary 1 that every distribution that depends on x only through $t = \mu^\top x$ can be expressed in terms of a **marginal distribution** p_T and a **uniform distribution** p_V on the subspace $\mathbb{S}^{d-1}$. Density evaluation as well as sampling *just* requires dealing with the marginal p_T as p_V has constant density and it is trivial to sample from. An example from this class is the von Mises-Fisher distribution.

Definition 1. *Let's define a unnormalized density as*

$$p_X(x; \mu, \kappa) \propto \exp\left(\kappa \mu^\top x\right) \quad \textit{with} \quad x \in \mathbb{S}^{d-1} \ , \tag{3}$$

with direction $\mu \in \mathbb{S}^{d-1}$ and concentration $\kappa \in \mathbb{R}_{\geq 0}$. When normalized, this is the von Mises-Fisher distribution.

Theorem 2 (9.3.12 in Mardia & Jupp (2009))**.** *The marginal $p_T(t; d, \kappa)$ of a von Mises-Fisher distribution is*

$$p_T(t; d, \kappa) = C_T(\kappa, d) \cdot e^{\kappa t} \left(1-t^2\right)^{\frac{d-3}{2}} \ , \tag{4}$$

with normalizer $C_T(\kappa, d) =$

$$\left(\frac{\kappa}{2}\right)^{\frac{d}{2}-1} \left\{\Gamma\left(\frac{d-1}{2}\right)\Gamma\left(\frac{1}{2}\right) I_{\frac{d-1}{2}}(\kappa)\right\}^{-1} \ , \tag{5}$$

where $I_v(z)$ is the modified Bessel function of the first kind.

Although evaluation of p_T is tractable, the Bessel function (Weisstein, 2002) (see Appendix A) is slow to compute and unstable for large arguments. Besides, p_T does not admit a known closed-form *cdf* (nor its inverse). Thus, rejection sampling is normally used to draw samples from it (Ulrich, 1984).

2.2. The Power Spherical distribution

As all the issues of the vMF distribution stem from a problematic marginal, we address them by defining a new distribution that shares some basic proprieties of a vMF but none of its drawbacks. Namely *i)* it is rotationally symmetric about μ, *ii)* it can be expressed in terms of a marginal distribution p_T, and *iii)* this marginal has closed-form and stable *cdf* (and inverse *cdf*).

Definition 2. *Let's define an unnormalized density as*

$$p_X(x; \mu, \kappa) \propto \left(1 + \mu^\top x\right)^\kappa \quad \textit{with} \quad x \in \mathbb{S}^{d-1} \ , \tag{6}$$

*with direction $\mu \in \mathbb{S}^{d-1}$ and concentration $\kappa \in \mathbb{R}_{\geq 0}$. When normalized this is the **Power Spherical** distribution.*

Algorithm 1 Power Spherical sampling

```
**Input:** dimension $d$, direction $\mu$, concentration $\kappa$
sample $z \sim \mathrm{Beta}\left(Z; (d-1)/2+\kappa, (d-1)/2\right)$
sample $v \sim \mathcal{U}(\mathcal{S}^{d-2})$
$t \leftarrow 2z - 1$
$y \leftarrow [t; (\sqrt{1-t^2})v^\top]^\top$ {concatenation}
$\hat{u} \leftarrow e_1 - \mu$ {$e_1$ is the base vector $[1, 0, \cdots, 0]^\top$}
$u = \frac{\hat{u}}{\|\hat{u}\|_2}$
$x \leftarrow (I_d - 2uu^\top)y$ {$I_d$ is the identity matrix $d \times d$}
**Return:** $x$
```

As we show in Theorem 12 (Appendix C.1), the marginal of the Power Spherical distribution has the valuable property of being defined in terms of an affine transformation of a Beta-distributed variable, *i.e.*,

$$T = 2Z - 1 \quad \text{with} \quad Z \sim \mathrm{Beta}\left(\alpha, \beta\right) \ , \tag{7}$$

where $\alpha = \frac{d-1}{2} + \kappa$ and $\beta = \frac{d-1}{2}$. Therefore, its density can be easily assessed via the change of variable theorem (Theorem 8 in Appendix B). Sampling and evaluating a Beta distribution is numerically stable and, crucially, it permits backpropagation though sampling via implicit reparameterization gradients (Figurnov et al., 2018). The properly normalized density of the Power Spherical distribution is derived in Theorem 13 (Appendix C.2) and is $p_X(x; \mu, \kappa) =$

$$\underbrace{\left\{2^{\alpha+\beta}\pi^{\beta}\frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha+\beta\right)}\right\}^{-1}}_{=N_X(\kappa,d)\ \text{(normalizer)}}\left(1+\mu^\top x\right)^{\kappa} \ . \tag{8}$$

Sampling As for the vMF,[1] draws are obtained sampling

$$t \sim p_T(t; \kappa, d) \quad \text{and} \quad v \sim \mathcal{U}(\mathbb{S}^{d-2}) \ , \tag{9}$$

and constructing $y = [t; v^\top\sqrt{1-t^2}]^\top$ using Theorem 1. Finally, we apply a Householder reflection about μ to y to obtain a sample x (see Algorithm 1). All these operations are differentiable which means we can use the reparameterization trick to have low variance and unbiased estimation of gradients of Monte Carlo samples with respect to the parameters of the density (Rezende et al., 2014; Kingma & Welling, 2014). Importantly, and differently from a vMF, sampling from a Power Spherical does not require rejection sampling. This leads to two main advantages: *i)* fast sampling (as we demonstrate in Section 3), and *ii)* no need for a high variance gradient correction term that compensates for sampling from a proposal distribution rather than the true one (Naesseth et al., 2017; Davidson et al., 2018).

Property	Value
$\mathbb{E}[X]$	$\mu(\alpha-\beta)/(\alpha+\beta)$
$\mathrm{var}(X)$	$\frac{2\alpha}{(\alpha+\beta)^2(\alpha+\beta+1)}\big((\beta-\alpha)\mu\mu^\top + (\alpha+\beta)I_d\big)$
Mode	μ (for $\kappa > 0$)
$\mathrm{H}(T)$	$\mathrm{H}(\mathrm{Beta}(\alpha,\beta)) + \log 2$
$\mathrm{H}(X)$	$\log N_X(\kappa, d) - \kappa\big(\log 2 + \psi\left(\alpha\right) - \psi\left(\alpha+\beta\right)\big)$

Table 1. Properties of $X \sim \mathrm{PowerSpherical}(\mu, \kappa)$. Recall that $\alpha = (d-1)/2 + \kappa$ and $\beta = (d-1)/2$.

2.3. Proprieties

Table 1 summarizes some basic properties of the Power Spherical distribution. See Appendix C.3 for derivations. In particular, note that having a closed-form differential entropy allows using the Power Spherical in applications such as variational inference (VI; Jordan et al., 1999) and mutual information minimization.

Kullback–Leibler divergence The KL divergence between a Power Spherical P and a uniform $Q = \mathcal{U}(\mathbb{S}^{d-1})$ is

$$\mathrm{D}_{\mathrm{KL}}[P\|Q] = -\,\mathrm{H}(P) + \mathrm{H}(Q) \ . \tag{10}$$

See Theorem 17 (Appendix C.5) for the full derivation. Being able to compute the KL divergence from a uniform distribution in closed-form is useful in the context of variational inference as Q can be used as a prior. Another important result we present here is a closed-form KL divergence of a Power Spherical distribution P with parameters μ_p, κ_p from a vMF Q with parameters μ_q, κ_q, which evaluates to

$$-\,\mathrm{H}(P) + \log C_X(\kappa_q, d) - \kappa_q\mu_q^\top\mu_p\left(\frac{\alpha-\beta}{\alpha+\beta}\right) \tag{11}$$

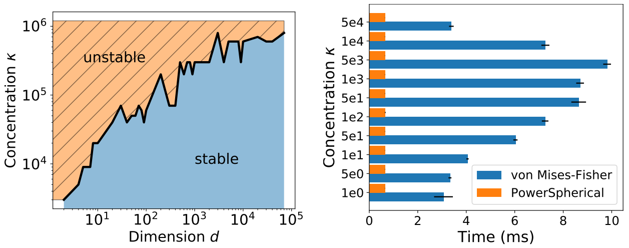
with $C_X(\kappa_q, d)$ the vMF normalizer (see Theorem 16 in Appendix C.5). This is valuable when using the Power Spherical to approximate a vMF, as we can assess the quality of the approximation. For example, if one has a vMF prior, it is then straightforward to use a Power Spherical approximate posterior in VI. If a vMF is necessary for a particular application, we can train with Power Spherical (enabling fast sampling and stable optimization) and then return the vMF that is closest to it in terms of KL.

3. Experiments

In this section we aim to show that Power Spherical distributions are more stable than vMFs, they are also faster to sample from, and lead to comparable performance when used in the context of variational auto-encoders.

Stability We tested numerical stability of both distributions for dimensions $d \in \{a \cdot 10^b\}$ and concentrations $\kappa \in \{a \cdot 10^b\}$ for all $a \in \{1, \ldots, 9\}$, $b \in \{0, \ldots, 5\}$. For

[1] This is equal to the method of Davidson et al. (2018) (Algorithms 1 and 3) for sampling from a vMF where we use the marginal of the Power Spherical instead.



(a) Stability of the vMF distribution. Ours does not have numerical issues in these intervals.

(b) Sampling time (on GPU) with $d = 64$ of a batch of 100 vectors of varius concentrations.

Figure 2. Comparing stability (a) and running time (b) of the von Mises-Fisher and the Power Spherical distribution.

every combination of $\langle d, \kappa \rangle$, we sample 10 vectors $x^{(i)}$ and compute the gradient $g^{(i)} = \nabla_\kappa \mu^\top x^{(i)}$. If at least one of the samples $x^{(i)}$ or one of the gradients $g^{(i)}$ returns *Not a Number* (NaN), we mark $\langle d, \kappa \rangle$ as unstable for that distribution. In Figure 2(a) we show the regions of instability. As intended, the Power Spherical does not present numerical issues in these intervals, while the vMF does. This makes our distribution more suitable where high dimensional vectors are needed such as for language modelling Kumar & Tsvetkov (2019).

Efficiency We also compare sampling efficiency between the Power Spherical and the vMF to highlight that rejection sampling is an undesirable bottleneck. We measured sampling time with $d = 64$ of a batch of 100 vectors of various concentrations $\kappa \in \{a \cdot 10^b\}$ with $a \in \{1, \dots, 5\}$, $b \in \{0, \dots, 4\}$.[2] For every concentration κ we computed mean and variance of 7 trials that consisted of computing the mean execution time (in milliseconds) sampling 100 times. Figure 2(b) shows the results. Sampling from a Power Spherical is at least 6× faster than sampling from a vMF. For some concentrations where the rejection ratio is worse, it is almost 20× faster. Noticeably, and differently from a vMF, sampling time is constant regardless of the concentration κ.

Variational inference Finally, we employed our Power Spherical distribution in the context of variational auto-encoders (Kingma & Welling, 2014). In particular, we replicated some of Davidson et al.'s (2018) experiments comparing the Power Spherical and the vMF on the the MNIST dataset (LeCun et al., 1998). We implement a simple feed-forward encoder $[784, 256, \tanh, 128, \tanh, d]$ and decoder $[d, 128, \tanh, 256, \tanh 784]$ and optimize the evidence lower bound for 100 epochs for $d \in \{5, 10, 20, 40\}$. We used Adam (Kingma & Ba, 2014) with learning rate 10^{-3} and batch size 64. Table 2 shows importance sampling estimates (with 5k Monte Carlo samples) of log-likelihood and the evidence lower bound on the test set. We observe no substantial difference in performance between the two distributions. This shows, across dimensions, that the Power Spherical is sufficiently expressive to replace the vMF in this application. However, training with the Power Spherical was $>2\times$ faster than with the vMF. One can notice that both log-likelihood and ELBO do not improve after $d = 40$. This is in line with Davidson et al.'s (2018) findings and a consequence of a shallow architecture. It is not the purpose of this experiment to show improvements on this task.

Method	**vMF**		**Power Spherical**	
	LL	ELBO	LL	ELBO
$d = 5$	-114.51	-117.68	-114.49	-118.01
$d = 10$	-97.37	-101.78	-97.46	-101.86
$d = 20$	-93.80	-99.38	-93.70	-99.27
$d = 40$	-98.64	-108.44	-98.63	-108.32

Table 2. Comparison between the vMf and Power Spherical distributions in a VAE on MNIST with different dimensional latent spaces $\mathbb{S}^{d-1}$. We show estimated (with 5k Monte Carlo samples) log-likelihood (LL) and evidence lower bond (ELBO) on test set.

4. Conclusion and future work

We presented a novel distribution on the d-dimensional sphere that, unlike the typically used von Mises-Fisher, *i)* is numerically stable in high dimensions and concentration, *ii)* has gradients of its samples with respect to its parameters, and *iii)* does not require rejection sampling allowing faster computation and exact reparameterization gradient without a high variance correction term. We empirically show that our distribution is more numerically stable and faster to sample from compared to a vMF while preforming equally well in a variational auto-encoder setting. As shown in our experiments, high dimensional hyperspaces suffer from surface area collapse at the expense of the expressivity of latent space embeddings. Davidson et al. (2019) addressed some of these issues, future work might explore this direction further. Another future direction is to use the Power Spherical in combination with normalizing flows (Rezende & Mohamed, 2015). Though neural autoregressive flows (Huang et al., 2018; De Cao et al., 2019) have been shown to be remarkably flexible, they are still subject to topological constraints (Dinh et al., 2019; Cornish et al., 2019), which motivates extensions for complex manifolds (Brehmer & Cranmer, 2020; Wu et al., 2020) including spheres and tori (Rezende et al., 2020). In addition, a recent work by Falorsi & Forré (2020) extend Neural Ordinary Differential Equations (Chen et al., 2018) to arbitrary Riemannian manifolds allowing the use of continuous normalizing flow to such spaces.

[2] On a NVIDIA Titian X 12GB GPU.

Acknowledgements

Authors want to thank Luca Falorsi and Sergio De Cao for helpful discussions and technical support as well as Srinivas Vasudevan and Robin Scheibler for spotting some typos. This project is supported by SAP Innovation Center Network, ERC Starting Grant BroadSem (678254).

References

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep): 1345–1382, 2005.

Bijral, A. S., Breitenbach, M., and Grudic, G. Mixture of watson distributions: a generative model for hyperspherical embeddings. In *Artificial Intelligence and Statistics*, pp. 35–42, 2007.

Brehmer, J. and Cranmer, K. Flows for simultaneous manifold learning and density estimation. *arXiv preprint arXiv:2003.13913*, 2020.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6571–6583. 2018.

Cornish, R., Caterini, A. L., Deligiannidis, G., and Doucet, A. Relaxing bijectivity constraints with continuously indexed normalising flows. *arXiv preprint arXiv:1909.13833*, 2019.

Davidson, T. R., Falorsi, L., De Cao, N., Kipf, T., and Tomczak, J. M. Hyperspherical variational auto-encoders. *34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

Davidson, T. R., Tomczak, J. M., and Gavves, E. Increasing Expressivity of a Hyperspherical VAE. *NeurIPS 2019, in Workshop on Bayesian Deep Learning*, 2019.

De Cao, N., Titov, I., and Aziz, W. Block Neural Autoregressive Flow. *35th Conference on Uncertainty in Artificial Intelligence (UAI19)*, 2019.

Dinh, L., Sohl-Dickstein, J., Pascanu, R., and Larochelle, H. A rad approach to deep mixture models. In *ICLR, Workshop track*, 2019.

Falorsi, L. and Forré, P. Neural ordinary differential equations on manifolds. *arXiv preprint arXiv:2006.06663*, 2020.

Falorsi, L., de Haan, P., Davidson, T. R., De Cao, N., Weiler, M., Forré, P., and Cohen, T. S. Explorations in homeomorphic variational auto-encoding. *ICML workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

Falorsi, L., de Haan, P., Davidson, T. R., and Forré, P. Reparameterizing distributions on lie groups. *AISTATS*, 2019.

Figurnov, M., Mohamed, S., and Mnih, A. Implicit reparameterization gradients. In *Advances in Neural Information Processing Systems*, pp. 441–452, 2018.

Huang, C.-W., Krueger, D., Lacoste, A., and Courville, A. Neural autoregressive flows. *International Conference on Machine Learning*, 2018.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Kumar, S. and Tsvetkov, Y. Von mises-fisher loss for training sequence to sequence models with continuous outputs. *Seventh International Conference on Learning Representations (ICLR 2019)*, 2019.

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

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., and Song, L. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

Mardia, K. V. and Jupp, P. E. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

Mathieu, E., Le Lan, C., Maddison, C. J., Tomioka, R., and Teh, Y. W. Continuous hierarchical representations with poincaré variational auto-encoders. In *Advances in neural information processing systems*, pp. 12544–12555, 2019.

Naesseth, C., Ruiz, F., Linderman, S., and Blei, D. Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms. *AISTATS*, pp. 489–498, 2017.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *International Conference on Machine Learning*, pp. 4693–4702, 2019.

Reisinger, J., Waters, A., Silverthorn, B., and Mooney, R. J. Spherical topic models. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 903–910, 2010.

Rezende, D. and Mohamed, S. Variational Inference with Normalizing Flows. *ICML*, 37:1530–1538, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *ICML*, pp. 1278–1286, 2014.

Rezende, D. J., Papamakarios, G., Racanière, S., Albergo, M. S., Kanwar, G., Shanahan, P. E., and Cranmer, K. Normalizing flows on tori and spheres. *arXiv preprint arXiv:2002.02428*, 2020.

Sra, S. Directional statistics in machine learning: a brief review. *Applied Directional Statistics: Modern Methods and Case Studies*, pp. 225, 2018.

Ulrich, G. Computer Generation of Distributions on the m-Sphere. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 33(2):158–163, 1984.

Weisstein, E. W. Bessel function of the first kind. *Wolfram*, 2002.

Wu, H., Köhler, J., and Noé, F. Stochastic normalizing flows. *arXiv preprint arXiv:2002.06707*, 2020.

A. Definitions

Definition 3. *The Gamma function is:*

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) \, \mathrm{d}u \; . \tag{12}$$

Definition 4. *The Beta function:*

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \; . \tag{13}$$

Definition 5. *The incomplete Beta function is:*

$$B_x(a, b) = \int_0^x u^{a-1}(1-u)^{b-1} \, \mathrm{d}u \; . \tag{14}$$

Definition 6. *The regularized incomplete Beta function is:*

$$I_x(a, b) = \frac{B_x(a, b)}{B(a, b)} \; . \tag{15}$$

Definition 7. *The surface area of the hyper-sphere $\mathbb{S}^{d-1}$ is:*

$$A_{d-1} = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \; . \tag{16}$$

Definition 8. *The uniform distribution on $\mathbb{S}^{d-1}$ has constant density equal to the reciprocal of the surface area (Definition 7):*

$$p_X(x) = \frac{1}{A_{d-1}} \; . \tag{17}$$

Definition 9. *The modified Bessel function of the first kind is defined as*

$$I_v(z) = \left(\frac{1}{2}z\right)^v \sum_{k=0}^{\infty} \frac{\left(\frac{1}{4}z^2\right)^k}{k!\Gamma(v+k+1)} \; . \tag{18}$$

An useful integral function

$$\begin{aligned} &= \int (1+x)^a (1-x)^b \, \mathrm{d}x \\ &= 2^{a+b+1} B_{\frac{x+1}{2}} \left(a+1, b+1\right) + C \; . \end{aligned} \tag{19}$$

Definition 10. *For a distribution P of a continuous random variable the differential entropy is defined to be the integral*

$$\mathrm{H}(P) = -\mathbb{E}_{p(x)} \left[\log p(x)\right] \; , \tag{20}$$

where p denotes the probability density function of P.

Definition 11. *For distributions P and Q of a continuous random variable the Kullback–Leibler divergence is defined to be the integral*

$$\mathrm{D}_{\mathrm{KL}}(P \| Q) = -\mathrm{H}(P) - \mathbb{E}_{p(x)} \left[\log q(x)\right] \; , \tag{21}$$

where p and q denote the probability density function of P and Q respectively.

Definition 12. *A random variable X is distributed according to the Beta distribution if its probability density function is*

$$B(\alpha, \beta)^{-1} x^{\alpha-1} (1-x)^{\beta-1} \; , \tag{22}$$

for $x \in [0, 1]$ and zero elsewhere where $\alpha \in \mathbb{R}^{>0}$ and $\beta \in \mathbb{R}^{>0}$ are shape parameters.

B. Theorems

Theorem 3. *The entropy of a random variable X Beta distributed is*

$$\begin{aligned} \mathrm{H}(X) = \log B(\alpha, \beta) &+ (\alpha+\beta-2)\psi(\alpha+\beta) \\ &- (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) \end{aligned} \tag{23}$$

where $\psi(x) = \frac{\partial}{\partial x} \log(\Gamma(x))$.

Theorem 4. *The expectation of X of a random variable X Beta distributed is*

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta} \; . \tag{24}$$

Theorem 5. *The expectation of $\log X$ of a random variable X Beta distributed is*

$$\mathbb{E}[\log X] = \psi(\alpha) - \psi(\alpha+\beta) \; . \tag{25}$$

Theorem 6. *The expectation of X^2 of a random variable X Beta distributed is*

$$\mathbb{E}[X^2] = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)} \; . \tag{26}$$

Theorem 7. *The variance of X of a random variable X Beta distributed is*

$$\mathrm{var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \; . \tag{27}$$

Theorem 8. *Given a bijective function $f : \mathcal{X} \to \mathcal{Y}$ between two continuous random variables $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$, a relation between the probability density functions $p_Y(y)$ and $p_X(x)$ is*

$$p_Y(y) = p_X(x) \left|\det \mathbf{J}_{f(x)}\right|^{-1} \; , \tag{28}$$

where $y = f(x)$, and $\left|\det \mathbf{J}_{f(x)}\right|$ is the absolute value of the determinant of the Jacobian of f evaluated at x.

Theorem 9. *Given a bijective function $f : \mathcal{X} \to \mathcal{Y}$ between two continuous random variables $X \in \mathcal{X} \subseteq \mathbb{R}^d$ and $Y \in \mathcal{Y} \subseteq \mathbb{R}^d$, a relation between the two respective entropies $\mathrm{H}(Y)$ and $\mathrm{H}(X)$ is*

$$\mathrm{H}(Y) = \mathrm{H}(X) + \int_{\mathcal{X}} p_X(x) \log \left|\det \mathbf{J}_{f(x)}\right| \mathrm{d}x \; , \tag{29}$$

where $y = f(x)$, and $\left|\det \mathbf{J}_{f(x)}\right|$ is the absolute value of the determinant of the Jacobian of f evaluated at x.

Theorem 10 (9.3.33 in Mardia & Jupp (2009))**.** *Continuous distributions with rotational symmetry about a direction μ and with probability density functions of the form $p_X \propto g(\mu^\top x)$ have mean*

$$\mathbb{E}[X] = \mathbb{E}[T]\mu \ . \tag{30}$$

Theorem 11 (9.3.34 in Mardia & Jupp (2009))**.** *Continuous distributions with rotational symmetry about a direction μ and with probability density functions of the form $p_X \propto g(\mu^\top x)$ have variance*

$$\operatorname{var}[X] = \operatorname{var}[T]\mu\mu^\top + \frac{1-\mathbb{E}[T^2]}{d-1}(I_d - \mu\mu^\top) \ , \tag{31}$$

where I_d is the $d \times d$ identity matrix.

C. Derivations

C.1. The Power Spherical marginal

Theorem 12. *Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ be the hyper-spherical set. Let an unnormalized density be*

$$p_X(x;\mu,\kappa) \propto \left(1+\mu^\top x\right)^\kappa \quad \textit{with} \quad x \in \mathbb{S}^{d-1} \ , \tag{32}$$

with $x \in \mathbb{S}^{d-1}$, direction $\mu \in \mathbb{S}^{d-1}$, and concentration parameter $\kappa \in \mathbb{R}_{\geq 0}$. Let T bet a random variable that denotes the dot-product $t = \mu^\top x$, then

$$T = 2Z - 1 \quad \textit{with} \quad Z \sim \operatorname{Beta}(\alpha,\beta) \ , \tag{33}$$

where $\alpha = \frac{d-1}{2} + \kappa$ and $\beta = \frac{d-1}{2}$.

Proof. Given Corollary 1, the marginal distribution of the dot-product t is $\propto (1+t)^\kappa(1-t^2)^{\frac{d-3}{2}}$ so its normalizer is

$$N_T(\kappa,d) = \int_{-1}^{1} (1+t)^\kappa \left(1-t^2\right)^{\frac{d-3}{2}} \mathrm{d}t \tag{34}$$

$$= \int_{-1}^{1} (1+t)^{\frac{d-3}{2}+\kappa} (1-t)^{\frac{d-3}{2}} \mathrm{d}t \tag{35}$$

$$\overset{(19)}{=} 2^{d+\kappa-2}\left(B_1\left(\frac{d-1}{2}+\kappa, \frac{d-1}{2}\right) - \underbrace{B_0\left(\frac{d-1}{2}+\kappa, \frac{d-1}{2}\right)}_{=0}\right) \tag{36}$$

$$= 2^{d+\kappa-2} B\left(\frac{d-1}{2}+\kappa, \frac{d-1}{2}\right) \ . \tag{37}$$

It follows that the **probability density function** of the dot-product marginal distribution is $p_T(t;\kappa,d) =$

$$= N_T(\kappa,d)^{-1} (1+t)^\kappa \left(1-t^2\right)^{\frac{d-3}{2}} \tag{38}$$

$$= N_T(\kappa,d)^{-1} (1+t)^{\frac{d-3}{2}+\kappa} (1-t)^{\frac{d-3}{2}} \tag{39}$$

$$\overset{*}{=} N_T(\kappa,d)^{-1} (2z)^{\frac{d-2}{2}+\kappa-1} (2-2z)^{\frac{d-1}{2}-1} \tag{40}$$

$$= B(\alpha,\beta)^{-1} z^{\alpha-1} (1-z)^{\beta-1} \ . \tag{41}$$

where $*$ indicates a substitution $t = 2z - 1$ and eventually $\alpha = \frac{d-1}{2} + \kappa$, $\beta = \frac{d-1}{2}$. Notice that Equation 41 is a Beta distribution. Therefore, if we define the random variable $Z \sim \operatorname{Beta}(\alpha,\beta)$ turns out that T is simply $T = 2Z - 1$. □

Corollary 2. *Following Theorem 12, the marginal of a Power Spherical distribution has **cumulative density function** $F(t;\kappa,d) =$*

$$= N_T(\kappa,d)^{-1} \int_{-1}^{t} (1+x)^\kappa \left(1-x^2\right)^{\frac{d-3}{2}} \mathrm{d}x \tag{42}$$

$$= B(\alpha,\beta)^{-1} B_{\frac{x+1}{2}}(\alpha,\beta) \tag{43}$$

$$= I_{\frac{t+1}{2}}(\alpha,\beta) \ , \tag{44}$$

*and **inverse cumulative density function***

$$F^{-1}(y;\kappa,d) = 2I_y^{-1}(\alpha,\beta) - 1 \ . \tag{45}$$

Corollary 3. *Using Theorem 3 and 9 we derive the differential entropy of the marginal Power Spherical p_T. Using $t = 2z - 1 = f(z)$, $\det \mathbf{J}_{f(z)} = 2$ for all z, and then*

$$\mathrm{H}(T) = \mathrm{H}(Z) + \log 2 \tag{46}$$

$$\overset{(23)}{=} \log B(\alpha,\beta) + (\alpha+\beta-2)\psi(\alpha+\beta) - (\alpha-1)\psi(\alpha) - (\beta-1)\psi(\beta) + \log 2 \ . \tag{47}$$

C.2. The normalized Power Spherical distribution

Theorem 13. *The normalized density of the Power Spherical distribution is $p_X(x;\mu,\kappa) =$*

$$\left\{2^{\alpha+\beta}\pi^\beta \frac{\Gamma(\alpha)}{\Gamma(\alpha+\beta)}\right\}^{-1} \left(1+\mu^\top x\right)^\kappa \ , \tag{48}$$

with $\alpha = \frac{d-1}{2} + \kappa$ and $\beta = \frac{d-1}{2}$.

Proof. The Power Spherical is expressed via the tangent normal decomposition (Theorem 1) as a joint distribution between $T \sim p_T(t;\kappa,d)$ (from Theorem 12) and $V \sim \mathcal{U}(\mathbb{S}^{d-2})$. Since $T \perp\!\!\!\perp V$, the Power Spherical normalizer $N_X(p,\kappa)$ is the product of the normalizer of $p_T(t;\kappa,d)$ and the uniform distribution on $\mathbb{S}^{d-2}$ (whose probability is constant on the $d-2$-sphere – see Definition 8), that is

$$N_X(\kappa, d) = N_T(\kappa, d) \cdot A_{d-2} \tag{49}$$

$$= 2^{\alpha+\beta-1} B\left(\alpha, \beta\right) \frac{2\pi^\beta}{\Gamma\left(\beta\right)} \tag{50}$$

$$= 2^{\alpha+\beta} \pi^\beta \frac{\Gamma\left(\alpha\right)}{\Gamma\left(\alpha+\beta\right)} \ . \tag{51}$$

Thus, $p_X(x; \mu, \kappa) = N_X(\kappa, d)^{-1}(1+\mu^\top x)^\kappa$. □

C.3. Power Spherical properties

Corollary 4. *Directly applying Theorem 10, the mean of a Power Spherical is*

$$\mathbb{E}[X] = \mathbb{E}[T]\mu = (2\mathbb{E}[Z]-1)\mu \overset{(24)}{=} \left(\frac{\alpha-\beta}{\alpha+\beta}\right)\mu \ , \tag{52}$$

with $\alpha = \frac{d-1}{2} + \kappa$, $\beta = \frac{d-1}{2}$.

Corollary 5. *Directly applying Theorem 11, the variance of a Power Spherical is* $\mathrm{var}[X] =$

$$= \mathrm{var}[T]\mu\mu^\top + \frac{1-\mathbb{E}[T^2]}{d-1}(I_d - \mu\mu^\top) \tag{53}$$

$$= 4\,\mathrm{var}[Z]\mu\mu^\top + 4\frac{\mathbb{E}[Z]-\mathbb{E}[Z^2]}{2\beta}(I_d - \mu\mu^\top) \tag{54}$$

$$\overset{(26,27)}{=} 4\,\mathrm{var}[Z]\mu\mu^\top + 4\frac{\mathrm{var}[Z](\alpha+\beta)}{2\beta}(I_d - \mu\mu^\top) \tag{55}$$

$$= \frac{2\,\mathrm{var}[Z]}{\beta}\left((\beta-\alpha)\mu\mu^\top + (\alpha+\beta)I_d\right) \tag{56}$$

$$\overset{(27)}{=} \frac{2\alpha\left((\beta-\alpha)\mu\mu^\top + (\alpha+\beta)I_d\right)}{(\alpha+\beta)^2(\alpha+\beta+1)} \ , \tag{57}$$

with $\alpha = \frac{d-1}{2} + \kappa$, $\beta = \frac{d-1}{2}$ *and* I_d *is the* $d \times d$ *identity matrix.*

Theorem 14. *The mode of a Power Spherical with* $\kappa > 0$ *is*

$$\underset{x \in \mathbb{S}^{d-1}}{\arg\max}\, p_X(x; \mu, \kappa) = \mu \ . \tag{58}$$

Proof. We have $\underset{x \in \mathbb{S}^{d-1}}{\arg\max}\, p_X(x; \mu, \kappa) =$

$$= \underset{x \in \mathbb{S}^{d-1}}{\arg\max}\, N_X(\kappa, d)^{-1}(1+\mu^\top x)^\kappa \tag{59}$$

$$\overset{\kappa \geq 0}{=} \underset{x \in \mathbb{S}^{d-1}}{\arg\max} \log\left(1+\mu^\top x\right) \tag{60}$$

$$= \underset{x \in \mathbb{S}^{d-1}}{\arg\max}\, \mu^\top x = \mu \ . \tag{61}$$

□

C.4. Power Spherical differential entropy

Theorem 15. *The differential entropy of the Power Spherical* p_X *that is* $\mathrm{H}(X) =$

$$\log N_X(\kappa, d) - \kappa\left(\log 2 + \psi\left(\alpha\right) - \psi\left(\alpha+\beta\right)\right) \ , \tag{62}$$

with $\alpha = \frac{d-1}{2} + \kappa$, $\beta = \frac{d-1}{2}$.

Proof. Applying Definition 10, $\mathrm{H}(X) =$

$$= -\mathbb{E}_X[\log q(X)] \tag{63}$$

$$= \log N_X(\kappa, d) - \kappa\, \mathbb{E}_X[\log\left(1+\mu^\top X\right)] \tag{64}$$

$$= \log N_X(\kappa, d) - \kappa\left(\log 2 + \mathbb{E}_Z[\log Z]\right) \tag{65}$$

$$\overset{(25)}{=} \log N_X(\kappa, d) - \kappa\left(\log 2 + \psi(\alpha) - \psi(\alpha+\beta)\right) \ . \tag{66}$$

□

C.5. Kullback–Leibler divergence with the von Mises-Fisher distribution

Theorem 16. *The Kullback–Leibler divergence* $\mathrm{D_{KL}}$ *(Definition 11) between a Power Spherical distribution* P *with parameters* μ_p, κ_p *and von Mises-Fisher and* Q *with parameters* μ_q, κ_q *is* $\mathrm{D_{KL}}[P\|Q] =$

$$-\mathrm{H}(P) + \log C_X(\kappa_q, d) - \kappa_q \mu_q^\top \mu_p \left(\frac{\alpha-\beta}{\alpha+\beta}\right) \ , \tag{67}$$

with $\alpha = \frac{d-1}{2} + \kappa$, $\beta = \frac{d-1}{2}$.

Proof. Applying Definition 11, $\mathrm{D_{KL}}[P\|Q] =$

$$= -\mathrm{H}(P) - \mathbb{E}_{p_X}[\log q(X)] \tag{68}$$

$$= -\mathrm{H}(P) + \log C_X(\kappa_q, d) - \mathbb{E}_{p_X}[\kappa_q \mu_q^\top X] \tag{69}$$

$$= -\mathrm{H}(P) + \log C_X(\kappa_q, d) - \kappa_q \mu_q^\top \mathbb{E}_{p_T}[T]\mu_p \tag{70}$$

$$= -\mathrm{H}(P) + \log C_X(\kappa_q, d) - \kappa_q \mu_q^\top \mu_p \left(\frac{\alpha-\beta}{\alpha+\beta}\right) \ , \tag{71}$$

where $C_X(\kappa_q, d)$ is the von Mises-Fisher normalizer. □

C.6. Kullback–Leibler divergence with $\mathcal{U}(\mathbb{S}^{d-1})$

Theorem 17. *The Kullback–Leibler divergence* $\mathrm{D_{KL}}$ *(Definition 11) between a Power Spherical distribution* P *and a uniform distribution on the sphere* $Q = \mathcal{U}(\mathbb{S}^{d-1})$ *is*

$$\mathrm{D_{KL}}[P\|Q] = -\mathrm{H}(P) + \mathrm{H}(Q) \tag{72}$$

Proof. Applying Definition 11, $\mathrm{D_{KL}}[P\|Q] =$

$$= -\mathrm{H}(P) - \mathbb{E}_{p_X}[\log q(X)] \tag{73}$$

$$= -\mathrm{H}(P) + \mathrm{H}(Q) \ , \tag{74}$$

where $\mathrm{H}(Q) = \log A_{d-1}$ (from Definition 8). □