

A APPENDIX

A.1 CONVERT POLAR COORDINATES TO CARTESIAN COORDINATES

We relate the definition of polar coordinates given in Section 3.1 to the definition of Cartesian coordinates (Blumenson, 1960) and the information expressed in embedding is not missing in the conversion. For example, it is easier to use orthogonal coordinates to calculate cosine similarity. Let $\mathbf{w} = (r, \theta, \varphi^1, \varphi^2, \dots, \varphi^{n-2})$ be a n -dimension vector in polar coordinates. Suppose $\bar{\mathbf{w}} = \{x_1, x_2, \dots, x_n\}$ is the corresponding vector of \mathbf{w} in Cartesian coordinates. Here, the angles θ and φ^k in n -dimensional polar coordinates are represented as follows:

$$\theta = 2 \operatorname{arccot} \frac{x_{n-1} + \sqrt{x_n^2 + x_{n-1}^2}}{x_n}, \quad \varphi^k = \arccos \frac{x_k}{\sqrt{x_n^2 + x_{n-1}^2 + \dots + x_k^2}}.$$

A.2 ANGLE DISTRIBUTIONS IN POLAR COORDINATES

As mentioned in Section 3.3.3, achieving a uniform distributions on a sphere with polar coordinates is complicated. Figure 6 shows distributions of θ and φ^k where samples uniformly distribute on a sphere at $r = 1$. Whereas the distribution of the θ dimension is intuitive, samples do not equally distribute in the φ^k dimensions; fewer samples around the poles and more samples around the center. Also, the distributions differ depending on the dimensions. This is stem from the fact that volume around the center and the poles is different on a sphere. Strictly saying, the differential cube is smaller when φ is close to the poles as in Figure 7. Therefore, to distribute samples uniformly on a sphere, the number of points should be smaller approaching to the poles $\varphi = 0, 2\pi$. More specifically, if we put the same number of samples at every positions as in the θ dimension, the number of samples per a differential cube of a sphere becomes larger around the poles than other positions.

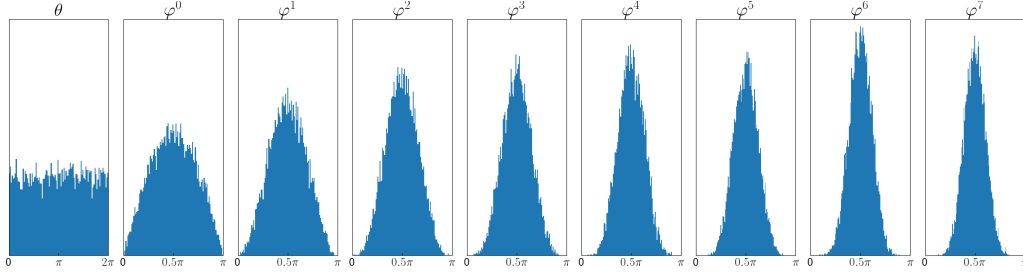


Figure 6: Angle distributions in ten-dimensional sphere.

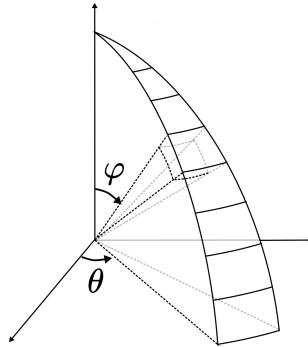


Figure 7: Relationships between angles and volume on a sphere in polar coordinates.

A.3 ANGLE DISTRIBUTIONS IN TRAINED MODELS

In addition to Figure 4 in the main content, Figure 8 shows the angle distributions of θ and φ^1 , respectively. The models are ones used in the ablation study and in Figure 4. Squared loss function (the left figures) produced biased distribution in both θ and φ dimensions. Welsch loss function (the center figures) soften the bias for the θ dimension (the upper figure) but not for the φ dimension much (the lower figure). SVGD significantly contributed to keep the distributions uniform in θ and φ .

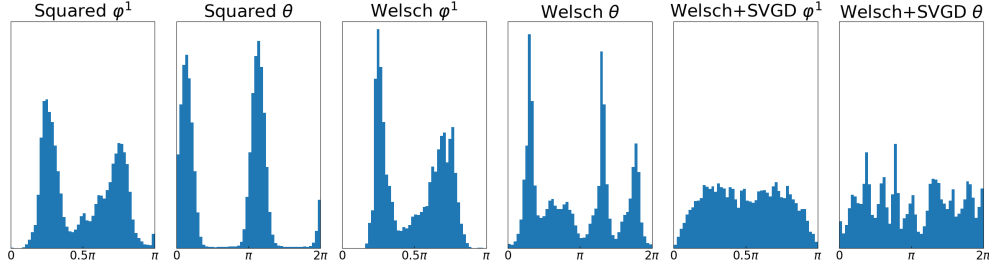


Figure 8: Angle distributions of polar embedding.

A.4 HYPERPARAMETERS

We tried hyperparameters listed in Table 3 and Table 4. In the experiment, we reported the results with the best one on the validation set for each model and each dataset.

Table 3: Hyperparameters for the noun hierarchy models

Module	Hyperparameter		Searched values
General	Dimension	-	5, 10
	Batch size	-	128, 256, 512, 1024
	Iteration	N	500000
	Learning rate	α	0.1, 0.3, 0.5, 0.7
	Learning rate decay	-	0.95, 0.99
	Negative sampling weight	β	0.1, 0.3, 0.5
Welsch loss	Parameter	c	0.4
SVGD	Batch size	-	128
	Interval	S	1000, 2000, 5000, 10000, 20000
	Iteration	M	2, 5, 20
	Learning rate	η	1
	Early stopping criterion	γ	0.99

Table 4: Hyperparameters for the mammal hierarchy model

Module	Hyperparameter		Searched values
General	Dimension	-	2
	Batch size	-	128
	Iteration	N	10000
	Learning rate	α	0.1, 0.3, 0.5
	Learning rate decay	-	0.95, 0.99
	Negative sampling weight	β	0.1, 0.3, 0.5
Welsch loss	Parameter	c	0.4
SVGD	Batch size	-	128
	Interval	S	500
	Iteration	M	5
	Learning rate	η	1
	Early stopping criterion	γ	0.99