

## Supplementary Material: Spherical Text Embedding

### A Proof of Theorem 1

**Definition 1** (Modified Bessel Function of the First Kind). The modified Bessel function of the first kind of order  $r$  can be defined as [26]:

$$I_r(\kappa) = \frac{(\kappa/2)^r}{\Gamma(r + \frac{1}{2}) \Gamma(\frac{1}{2})} \int_0^\pi \exp(\kappa \cos \theta) (\sin \theta)^{2r} d\theta,$$

where  $\Gamma(x) = \int_0^\infty \exp(-t) t^{x-1} dt$  is the gamma function.

**Lemma 1.** The definite integral of power of sin on the interval  $[0, \pi]$  is given by

$$J_p = \int_0^\pi (\sin x)^p dx = \frac{\sqrt{\pi} \Gamma(\frac{1+p}{2})}{\Gamma(1 + \frac{p}{2})} \quad (p \in \mathbb{Z}^+),$$

where  $\Gamma(x) = \int_0^\infty \exp(-t) t^{x-1} dt$  is the gamma function.

*Proof.*

$$\begin{aligned} J_p &= \int_0^\pi (\sin x)^p dx \\ &= \left( -(\sin x)^{p-1} \cos x \right) \Big|_0^\pi + (p-1) \int_0^\pi (\sin x)^{p-2} (\cos x)^2 dx \\ &= 0 + (p-1) \int_0^\pi (\sin x)^{p-2} dx - (p-1) \int_0^\pi (\sin x)^p dx \\ &= (p-1) J_{p-2} - (p-1) J_p \end{aligned}$$

Therefore,  $J_p = \frac{p-1}{p} J_{p-2}$ .

Using the above iteration relationship and the property of gamma function  $\Gamma(x+1) = x\Gamma(x)$ , we write  $J_p$  using gamma function:

- When  $p$  is an even integer:

$$\begin{aligned} J_p &= \frac{p-1}{p} \frac{p-3}{p-2} \cdots \frac{1}{2} J_0 \\ &= \frac{(p-1)/2}{p/2} \frac{(p-3)/2}{(p-2)/2} \cdots \frac{1/2}{2/2} J_0 \\ &= \frac{\Gamma(\frac{1+p}{2}) / \Gamma(\frac{1}{2})}{\Gamma(\frac{p}{2} + 1) / \Gamma(1)} J_0 \end{aligned}$$

Plugging in the base case  $J_0 = \pi$  and  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ ,  $\Gamma(1) = 1$ , we prove that

$$J_p = \frac{\sqrt{\pi} \Gamma(\frac{1+p}{2})}{\Gamma(1 + \frac{p}{2})} \quad (p \in \mathbb{Z}^+, p \text{ is even})$$

- When  $p$  is an odd integer:

$$\begin{aligned} J_p &= \frac{p-1}{p} \frac{p-3}{p-2} \cdots \frac{2}{3} J_1 \\ &= \frac{(p-1)/2}{p/2} \frac{(p-3)/2}{(p-2)/2} \cdots \frac{2/2}{3/2} J_1 \\ &= \frac{\Gamma(\frac{1+p}{2}) / \Gamma(1)}{\Gamma(\frac{p}{2} + 1) / \Gamma(\frac{3}{2})} J_1 \end{aligned}$$

Plugging in the base case  $J_1 = 2$  and  $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$ ,  $\Gamma(1) = 1$ , we prove that

$$J_p = \frac{\sqrt{\pi} \Gamma(\frac{1+p}{2})}{\Gamma(1 + \frac{p}{2})} \quad (p \in \mathbb{Z}^+, p \text{ is odd})$$

□

**Theorem 1.** When the corpus has infinite vocabulary, *i.e.*,  $|V| \rightarrow \infty$ , the analytic forms of Equations (1) and (2) are given by the von Mises-Fisher (vMF) distribution with the prior embedding as the mean direction and constant 1 as the concentration parameter, *i.e.*,

$$\lim_{|V| \rightarrow \infty} p(v | u) = \text{vMF}_p(\mathbf{v}; \mathbf{u}, 1), \quad \lim_{|V| \rightarrow \infty} p(u | d) = \text{vMF}_p(\mathbf{u}; \mathbf{d}, 1).$$

*Proof.* We give the proof for the first equality, and the second equality can be derived similarly. We generalize the relationship proportionality  $p(\mathbf{v} | \mathbf{u}) \propto \exp(\exp(\cos(\mathbf{v}, \mathbf{u})))$  in Equation (2) to the continuous case and obtain the following probability dense distribution:

$$\lim_{|V| \rightarrow \infty} p(v | u) = \frac{\exp(\cos(\mathbf{v}, \mathbf{u}))}{\int_{\mathbb{S}^{p-1}} \exp(\cos(\mathbf{v}', \mathbf{u})) d\mathbf{v}'} \triangleq \frac{\exp(\cos(\mathbf{v}, \mathbf{u}))}{Z}, \quad (8)$$

where we denote the integral in the denominator as  $Z$ .

To evaluate the integral  $Z$ , we make the transformation to polar coordinates. Let  $\mathbf{t} = Q\mathbf{v}'$ , where  $Q \in \mathbb{R}^{p \times p}$  is an orthogonal transformation so that  $d\mathbf{t} = d\mathbf{v}'$ . Moreover, let the first row of  $Q$  be  $\mathbf{u}$  so that  $t_1 = \mathbf{u}^\top \mathbf{v}'$ . Then we use  $(r, \theta_1, \dots, \theta_{p-1})$  to represent the polar coordinates of  $\mathbf{t}$  where  $r = 1$  and  $\cos \theta_1 = \mathbf{u}^\top \mathbf{v}$ . The transformation from Euclidean coordinates to polar coordinates is given by [35] via computing the determinant of the Jacobian matrix for the coordinate transformation:

$$d\mathbf{t} = r^{p-1} \prod_{j=2}^p (\sin \theta_{j-1})^{p-j} d\theta_{j-1}.$$

Then

$$Z = \int_0^\pi \exp(\cos \theta_1) (\sin \theta_1)^{p-2} d\theta_1 \prod_{j=3}^{p-1} \int_0^\pi (\sin \theta_{j-1})^{p-j} d\theta_{j-1} \int_0^{2\pi} d\theta_{j-1}.$$

By Lemma 1, we have

$$\prod_{j=3}^{p-1} \int_0^\pi (\sin \theta_{j-1})^{p-j} d\theta_{j-1} = \pi^{\frac{p-3}{2}} \frac{\Gamma(\frac{p-2}{2}) \Gamma(\frac{p-3}{2}) \cdots \Gamma(1)}{\Gamma(\frac{p-1}{2}) \Gamma(\frac{p-2}{2}) \cdots \Gamma(\frac{3}{2})} = \frac{\pi^{\frac{p-3}{2}}}{\Gamma(\frac{p-1}{2})}.$$

Then

$$\begin{aligned} Z &= \int_0^\pi \exp(\cos \theta_1) (\sin \theta_1)^{p-2} d\theta_1 \cdot \frac{\pi^{\frac{p-3}{2}}}{\Gamma(\frac{p-1}{2})} \cdot 2\pi \\ &= \frac{2\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})} \int_0^\pi \exp(\cos \theta_1) (\sin \theta_1)^{p-2} d\theta_1. \end{aligned}$$

According to Definition 1, the integral term of  $Z$  above can be expressed with  $I_{p/2-1}(1)$  as:

$$\int_0^\pi \exp(\cos \theta_1) (\sin \theta_1)^{p-2} d\theta_1 = \frac{\Gamma(\frac{p-1}{2}) \Gamma(\frac{1}{2})}{2^{1-p/2}} I_{p/2-1}(1).$$

Therefore, with the fact that  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ ,

$$Z = \frac{2\pi^{\frac{p-1}{2}}}{\Gamma(\frac{p-1}{2})} \frac{\Gamma(\frac{p-1}{2}) \Gamma(\frac{1}{2})}{2^{1-p/2}} I_{p/2-1}(1) = (2\pi)^{p/2} I_{p/2-1}(1).$$

Plugging  $Z$  back to Equation (8), we finally arrive that

$$\lim_{|V| \rightarrow \infty} p(v | u) = \frac{1}{(2\pi)^{p/2} I_{p/2-1}(1)} \exp(\cos(\mathbf{v}, \mathbf{u})) = \text{vMF}_p(\mathbf{v}; \mathbf{u}, 1).$$

□

## B Proof of Theorem 2

**Lemma 2.** Let  $(X_n)_{n \in \mathbb{N}}$  be a non-negative stochastic process with bounded positive variations, *i.e.*,  $\sum_{n=0}^{\infty} \mathbb{E} [\max (\mathbb{E} [X_{n+1} - X_n \mid \mathcal{F}_n], 0)] < \infty$ . Then this process is a quasi-martingale, *i.e.*,

$$\Pr \left( \sum_{n=0}^{\infty} |\mathbb{E} [X_{n+1} - X_n \mid \mathcal{F}_n]| < \infty \right) = 1, \quad \Pr \left( \lim_{n \rightarrow \infty} X_n = X^* \right) = 1,$$

where  $\mathcal{F}_n$  is the increasing sequence of  $\sigma$ -algebras generated by variables before time  $n$ ,  $\mathcal{F}_n = \{s_0, \dots, s_{n-1}\}$ , such that  $X_n$  computed from  $s_0, \dots, s_{n-1}$  is  $\mathcal{F}_n$  measurable.

*Proof.* See [14]. □

Before proving Theorem 2, we first prove the update rule without approximation, *i.e.*, replacing the retraction  $R_x$  in Equation (7) with the exponential map  $\exp_x$  defined by Equation (4), leads to almost surely convergence.

**Lemma 3.** When the update rule given by

$$\mathbf{x}_{t+1} = R_{\mathbf{x}_t} \left( -\eta_t \left( 1 + \frac{\mathbf{x}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|} \right) (I - \mathbf{x}_t \mathbf{x}_t^\top) \nabla f(\mathbf{x}_t) \right).$$

is applied to  $\mathcal{L}(\theta)$ , and the learning rate satisfies the usual condition in stochastic approximation, *i.e.*,  $\sum_t \eta_t^2 < \infty$  and  $\sum_t \eta_t = \infty$ ,  $\theta$  converges almost surely to a critical point  $\theta^*$  and  $\text{grad } \mathcal{L}(\theta)$  converges almost surely to 0, *i.e.*,

$$\Pr \left( \lim_{t \rightarrow \infty} \mathcal{L}(\theta_t) = \mathcal{L}(\theta^*) \right) = 1, \quad \Pr \left( \lim_{t \rightarrow \infty} \text{grad } \mathcal{L}(\theta_t) = 0 \right) = 1.$$

*Proof.* We use  $l(\mathbf{x}_t, z_t)$  to denote the approximated loss function  $\mathcal{L}(\mathbf{x}_t)$  evaluated at a training instance  $z_t$ , *i.e.*,  $\mathcal{L}(\mathbf{x}_t) = \mathbb{E}_z[l(\mathbf{x}_t, z)]$ .

Let

$$\xi(\mathbf{x}_t, z_t) = \eta_t \left( 1 + \frac{\mathbf{x}_t^\top \nabla l(\mathbf{x}_t, z_t)}{\|\nabla l(\mathbf{x}_t, z_t)\|} \right), \quad \text{grad } l(\mathbf{x}_t, z_t) = (I - \mathbf{x}_t \mathbf{x}_t^\top) \nabla l(\mathbf{x}_t, z_t),$$

and we omit the arguments of  $\xi$  and  $\text{grad } l$  in the following derivation because we only care about the upper bound of both.

We consider two consecutive gradient update steps of the parameter,  $\mathbf{x}_t$  and  $\mathbf{x}_{t+1}$ . There exists a geodesic segment  $\gamma(s) = \exp_{\mathbf{x}_t}(-s \cdot \text{grad } l)$  linking  $\mathbf{x}_t$  and  $\mathbf{x}_{t+1}$ , and its length is obtained by

$$\ell(\gamma) = \int_0^\xi \|\gamma'(s)\| ds.$$

Then we apply the Taylor formula on Riemannian manifold [6] and have

$$\begin{aligned} \mathcal{L}(\mathbf{x}_{t+1}) &= \mathcal{L}(\exp_{\mathbf{x}_t}(-\xi \cdot \text{grad } l)) \\ &= \mathcal{L}(\mathbf{x}_t) - \xi \cdot \text{grad } l^\top \text{grad } \mathcal{L}(\mathbf{x}_t) + \int_0^\xi (\xi - s) \gamma'(s)^\top \text{Hess } \mathcal{L}(\gamma(s)) \gamma'(s) ds, \end{aligned} \quad (9)$$

where  $\text{grad } \mathcal{L}$  is the Riemannian gradient satisfying

$$\left. \frac{d}{dt} \mathcal{L}(\exp_{\mathbf{x}}(t\alpha)) \right|_{t=0} = \alpha^\top \text{grad } \mathcal{L}(\mathbf{x}),$$

and  $\text{Hess } \mathcal{L}(\mathbf{x})$  is the Riemannian Hessian satisfying

$$\left. \frac{d}{dt} (\text{grad } \mathcal{L}(\exp_{\mathbf{x}}(t\alpha))^\top \text{grad } \mathcal{L}(\exp_{\mathbf{x}}(t\alpha))) \right|_{t=0} = 2 \text{grad } \mathcal{L}(\mathbf{x})^\top \text{Hess } \mathcal{L}(\mathbf{x}) \alpha.$$

For spherical space, the Riemannian Hessian is given by [11, 12]

$$\text{Hess } f(\mathbf{x}) := (I - \mathbf{x} \mathbf{x}^\top) (\nabla^2 f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \mathbf{x} I),$$

where  $\nabla^2 f(\mathbf{x})$  is the Euclidean Hessian of  $f(\mathbf{x})$ .

Therefore, at each training instance  $z_t$ , the Riemannian Hessian of the approximated loss is bounded:

$$\begin{aligned}\|\text{Hess } l(\mathbf{x}_t, z_t)\| &\leq \|I - \mathbf{x}_t \mathbf{x}_t^\top\| \|\nabla^2 l(\mathbf{x}_t, z_t) - \nabla l(\mathbf{x}_t, z_t)^\top \mathbf{x}_t I\| \\ &\leq (1 + \|\mathbf{x}_t\|^2) \|\nabla l(\mathbf{x}_t, z_t)\| \|\mathbf{x}_t\| \\ &\leq 4,\end{aligned}$$

where we use the fact that  $\nabla^2 l(\mathbf{x}_t, z_t) = 0$  and  $\|\nabla l(\mathbf{x}_t, z_t)\| \leq 2$  by computing the Euclidean Hessian and gradient of Equation (3).

Consequently, the Riemannian Hessian of the original loss is bounded:

$$\|\text{Hess } \mathcal{L}(\mathbf{x}_t)\| \leq \mathbb{E}_z [\|\text{Hess } l(\mathbf{x}_t, z)\|] \leq 4.$$

Similarly, the Riemannian gradient is also bounded:

$$\|\text{grad } l(\mathbf{x}_t, z_t)\| \leq \|I - \mathbf{x}_t \mathbf{x}_t^\top\| \|\nabla l(\mathbf{x}_t, z_t)\| \leq 4.$$

Also,  $\xi$  is bounded by

$$\xi(\mathbf{x}_t, z_t) = \eta_t \left( 1 + \frac{\mathbf{x}_t^\top \nabla l(\mathbf{x}_t, z_t)}{\|\nabla l(\mathbf{x}_t, z_t)\|} \right) \leq 2\eta_t.$$

Then the integral in Equation (9) is bounded by

$$\int_0^\xi (\xi - s) \gamma'(s)^\top \text{Hess } \mathcal{L}(\gamma(s)) \gamma'(s) ds \leq \xi^2 \|\text{grad } l\|^2 \|\text{Hess } \mathcal{L}(\mathbf{x}_t)\| \leq 256\eta_t^2.$$

Therefore, Equation (9) shows that

$$\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_t) \leq -\xi \cdot \text{grad } l^\top \text{grad } \mathcal{L}(\mathbf{x}_t) + 256\eta_t^2. \quad (10)$$

Now let  $\mathcal{F}_t$  be the increasing sequence of  $\sigma$ -algebras generated by variables before time  $t$ ,  $\mathcal{F}_t = \{z_0, \dots, z_{t-1}\}$ , such that  $\mathbf{x}_t$  computed from  $z_0, \dots, z_{t-1}$  is  $\mathcal{F}_t$  measurable.

Then we take the expectation over  $z$  under  $\mathcal{F}_t$  of both sides of Equation (10)

$$\begin{aligned}\mathbb{E}_z[\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_t) \mid \mathcal{F}_t] &\leq -\xi \mathbb{E}_z[\text{grad } l(\mathbf{x}_t, z_t)^\top \text{grad } \mathcal{L}(\mathbf{x}_t) \mid \mathcal{F}_t] + 256\eta_t^2 \\ &= -\xi \mathbb{E}_z[\text{grad } l(\mathbf{x}_t, z_t)^\top \text{grad } \mathcal{L}(\mathbf{x}_t)] + 256\eta_t^2 \\ &= -\xi \|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2 + 256\eta_t^2,\end{aligned} \quad (11)$$

since  $z_t$  is independent of  $\mathcal{F}_t$ .

As  $\mathcal{L}(\mathbf{x}_t) \geq 0$  and  $\sum_t \eta_t^2 < \infty$ , this shows that  $\mathcal{L}(\mathbf{x}_t) + \sum_t^\infty 256\eta_t^2$  is a non-negative supermartingale, thus it converges almost surely. Therefore,  $\mathcal{L}(\mathbf{x}_t)$  converges almost surely.

Next, to prove  $\text{grad } \mathcal{L}(\mathbf{x}_t)$  converges, we repeat the above proof and replace  $\mathcal{L}(\mathbf{x}_t)$  with  $\|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2$ . Specifically, we can bound the second derivative of  $\|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2$  and arrive at a very similar form as Equation (11). We then prove  $\text{grad } \mathcal{L}(\mathbf{x}_t)$  almost surely converge.

Finally, we prove  $\text{grad } \mathcal{L}(\mathbf{x}_t)$  must converge to 0. Summing over  $t$  of Equation (11), we have

$$\sum_{t \geq 0} \mathbb{E}_z[\mathcal{L}(\mathbf{x}_{t+1}) \mathcal{L}(\mathbf{x}_t) \mid \mathcal{F}_t] \leq \sum_{t \geq 0} 256\eta_t^2 - \sum_{t \geq 0} \xi \|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2.$$

From Equation (11), we know that  $\mathcal{L}(\mathbf{x}_t)$  satisfies the assumption in Lemma 2. Hence,  $\sum_{t \geq 0} \mathbb{E}_z[\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_t) \mid \mathcal{F}_t]$  converges almost surely, implying  $\sum_{t \geq 0} \xi \|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2$  also converges almost surely. Combining with the fact that  $\text{grad } \mathcal{L}(\mathbf{x}_t)$  converges almost surely, which we have proved, we show that  $\text{grad } \mathcal{L}(\mathbf{x}_t)$  must converge to 0.  $\square$

**Theorem 2.** When the update rule given by Equation (7) is applied to  $\mathcal{L}(\mathbf{x})$ , and the learning rate satisfies the usual condition in stochastic approximation, i.e.,  $\sum_t \eta_t^2 < \infty$  and  $\sum_t \eta_t = \infty$ ,  $\mathbf{x}$  converges almost surely to a critical point  $\mathbf{x}^*$  and  $\text{grad } \mathcal{L}(\mathbf{x})$  converges almost surely to 0, i.e.,

$$\Pr \left( \lim_{t \rightarrow \infty} \mathcal{L}(\mathbf{x}_t) = \mathcal{L}(\mathbf{x}^*) \right) = 1, \quad \Pr \left( \lim_{t \rightarrow \infty} \text{grad } \mathcal{L}(\mathbf{x}_t) = 0 \right) = 1.$$

*Proof.* Let

$$\mathbf{x}_{t+1}^{\text{exp}} = \exp_{\mathbf{x}_t} \left( -\eta_t \left( 1 + \frac{\mathbf{x}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|} \right) (I - \mathbf{x}_t \mathbf{x}_t^\top) \nabla f(\mathbf{x}_t) \right)$$

be the updated point mapped via exponential mapping.

Let

$$\mathbf{x}_{t+1} = R_{\mathbf{x}_t} \left( -\eta_t \left( 1 + \frac{\mathbf{x}_t^\top \nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|} \right) (I - \mathbf{x}_t \mathbf{x}_t^\top) \nabla f(\mathbf{x}_t) \right)$$

be the updated point mapped via retraction.

The retraction is a first-order approximation of the exponential mapping, *i.e.*,  $\exists M > 0$  such that  $d(R_{\mathbf{x}_t}(\epsilon \boldsymbol{\alpha}), \exp_{\mathbf{x}_t}(\epsilon \boldsymbol{\alpha})) < M\epsilon^2$  for  $\epsilon > 0$  sufficiently small, where  $\|\boldsymbol{\alpha}\| = 1$ .

Then,

$$\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_t) \leq |\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_{t+1}^{\text{exp}})| + \mathcal{L}(\mathbf{x}_{t+1}^{\text{exp}}) - \mathcal{L}(\mathbf{x}_t), \quad (12)$$

where  $\mathcal{L}(\mathbf{x}_{t+1}^{\text{exp}}) - \mathcal{L}(\mathbf{x}_t)$  is proved to be bounded in Equation (10) of Lemma 3, and the term  $|\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_{t+1}^{\text{exp}})|$  is also bounded

$$|\mathcal{L}(\mathbf{x}_{t+1}) - \mathcal{L}(\mathbf{x}_{t+1}^{\text{exp}})| \leq 256M\eta_t^2.$$

by applying a similar derivation as in Lemma 3 from Equation (9) to Equation (10).

Therefore,  $\mathcal{L}(\mathbf{x}_t)$  is a quasi-martingale and converges almost surely according to Lemma 2. Also,  $\sum_{t=1}^{\infty} \xi \|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2 < \infty$  almost surely, which means  $\|\text{grad } \mathcal{L}(\mathbf{x}_t)\|$  can only converge to 0 if it converges because  $\sum_t \eta_t = \infty$ .

Finally, to prove  $\text{grad } \mathcal{L}(\mathbf{x}_t)$  almost surely converges, we repeat the above proof by replacing  $\mathcal{L}(\mathbf{x}_t)$  with  $\|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2$  so that we can arrive at a similar form of Equation (12) and use the same procedure to show  $\|\text{grad } \mathcal{L}(\mathbf{x}_t)\|^2$  is a quasi-martingale and converges almost surely.  $\square$