Supplementary for CO-SNE: Dimensionality Reduction and Visualization for Hyperbolic Data

Yunhui Guo Haoran Guo Stella X. Yu
UC Berkeley / ICSI

1. Riemannian Geometry

We give a brief overview of Riemannian geometry, for more details please refer to [1]. A Riemannian manifold $(\mathcal{M}, \mathfrak{g})$ is a real smooth manifold $\mathcal{M}$ with a Riemannian metric $\mathfrak{g}$. The Riemannian metric $\mathfrak{g}$ is a smoothly varying inner product which is defined on the tangent space $T_{\mathbf{x}}\mathcal{M}$ of $\mathcal{M}$. Given $\mathbf{x} \in \mathcal{M}$ and two vectors $\mathbf{v}, \mathbf{w} \in T_{\mathbf{x}}\mathcal{M}$, we can use the Riemannian metric $\mathfrak{g}$ to compute the inner product $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathbf{x}}$ as $\mathfrak{g}(\mathbf{v}, \mathbf{w})$. The norm of $\mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$ is defined as $\|\mathbf{v}\|_{\mathbf{x}} = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle_{\mathbf{x}}}$. A geodesic generalizes the notion of straight line in the manifold which is defined as a curve $\gamma : [0, 1] \to \mathcal{M}$ of constant speed that is everywhere locally a distance minimizer. The exponential map and the inverse exponential map are defined as follows: given $\mathbf{x}, \mathbf{y} \in \mathcal{M}, \mathbf{v} \in T_{\mathbf{x}}\mathcal{M}$, and a geodesic γ of length $\|\mathbf{v}\|$ such that $\gamma(0) = \mathbf{x}, \gamma(1) = \mathbf{y}, \gamma'(0) = \mathbf{v}/\|\mathbf{v}\|$, the exponential map $\text{Exp}_{\mathbf{x}} : T_{\mathbf{x}}\mathcal{M} \to \mathcal{M}$ satisfies $\text{Exp}_{\mathbf{x}}(\mathbf{v}) = \mathbf{y}$ and the inverse exponential map $\text{Exp}_{\mathbf{x}}^{-1} : \mathcal{M} \to T_{\mathbf{x}}\mathcal{M}$ satisfies $\text{Exp}_{\mathbf{x}}^{-1}(\mathbf{y}) = \mathbf{v}$.

2. Hyperbolic Student's t-Distribution

Recall the way to define the the Student's t-distribution which expresses the random variable t as,

$$t = \frac{u}{\sqrt{v/n}} \tag{1}$$

where u is a random variable sampled from a standard normal distribution and v is a random variable sampled from a χ^2-distribution of n degrees of freedom. The χ^2-distribution can also be derived from normal distribution. Let u_1, u_2, ..., u_n be independent standard normal random variables, then the sum of the squares,

$$v = \sum_{i=1}^{n} u_i^2 \tag{2}$$

is a χ^2-distribution with n degrees of freedom. Thus, the probability density function of the χ^2-distribution can be derived from the probability density function of the normal distribution which is,

$$T_n(v) = \frac{1}{2^{n/2}\Gamma(n/2)} v^{(n-2)/2} e^{-v/2} \tag{3}$$

Using Equation 1, we can further derive the probability density function of the Student's t-distribution,

$$f_n(t) = \frac{1}{\sqrt{n}B(1/2, n/2)} (1 + \frac{t^2}{n})^{-(n+1)/2} \tag{4}$$

where B is the Beta function.

The probability density function of hyperbolic Cauchy distribution can be derived in a similar way using hyperbolic normal distribution.

3. Hyperbolic Cauchy Distribution

Similar to the Student's t-Distribution, the probability density function of Cauchy distribution can derived from the probability density function of the normal distribution. In particular, let X and Y be independent standard normal random variables, then $Z = \frac{X}{X+Y}$ is a Cauchy random variable. The probability density function of Cauchy distribution can be written as,

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma[1 + (\frac{x - x_0}{\gamma})^2]} \tag{5}$$

The probability density function of hyperbolic Cauchy distribution can be derived in a similar way using hyperbolic normal distribution.

The repulsion and attraction forces in CO-SNE depend on the term $p_{ij} - q_{ij}$ in Equation 12 of the main text. p_{ij} is fixed during training which depends on the distribution of the high-dimensional datapoints. To create more repulsion forces between two close low-dimensional embeddings y_i and y_j, we aim at increasing the probability that the point y_i would select the point y_j as its neighbor (i.e., q_{ij}). By using a small γ in hyperbolic Cauchy distribution, the distance between y_i and y_j is scaled up. When the point y_i is fixed, the probability of selecting y_j as a neighbor (i.e., q_{ij}) is scaled up relatively to some point y_k which is far away

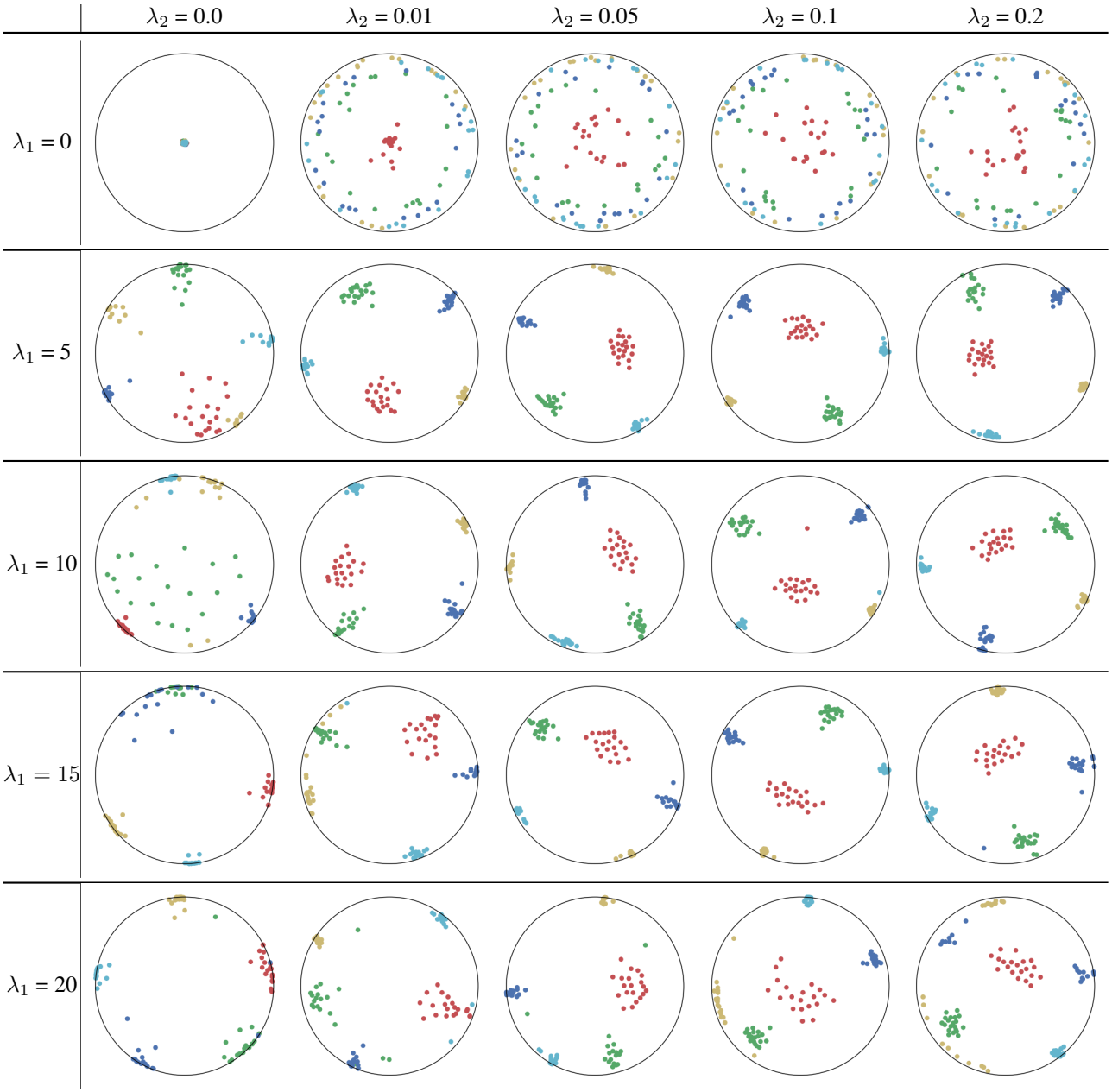

Figure 1. The effect of λ_1 and λ_2 on the quality of the visualization. We show the visualization results of CO-SNE on a mixture of five hyperbolic normal distributions in a five-dimensional hyperbolic space. We can observe that λ_1 is responsible for preserving the local similarity structure and λ_2 is responsible for preserving the global hierarchical structure. *This shows that both the KL-divergence and the distance loss are important for producing good visualization*. CO-SNE is robust to wide choices of λ_1 and λ_2. Generally, λ_1 should be larger than λ_2 since the magnitude of the gradients of the KL-divergence is smaller.

from y_i. As a result, $p_{ij} - q_{ij}$ can potentially become negative. This creates additional repulsion forces to push the low-dimensional points apart.

4. The Effect of λ_1 and λ_2

Recall the objective function of CO-SNE,

$$\mathcal{L} = \lambda_1 \mathcal{C} + \lambda_2 \mathcal{H} \tag{6}$$

λ_1 and λ_2 are used to balance the KL-divergence $\mathcal{C}$ and the distance loss $\mathcal{H}$ and can be regarded as the learning rates.

For ablation studies on the effect of λ_1 and λ_2, we reuse the settings in Section 4.1 of the main text.

We consider different settings of λ_1 and λ_2 to investigate the effect of the KL-divergence and the distance loss. The results are shown in Figure 1. We have several observations from the results.

1. In the first row, $\lambda_1 = 0.0$. This means that only the distance loss is presented. We can observe that the low-dimensional embeddings can only approximate the magnitude of the high-dimensional datapoints but not the similarity structure.

2. In the first column, $\lambda_2 = 0.0$. This means that only the KL-divergence is presented. We can observe that the low-dimensional embeddings can only preserve the similarity structure in the high-dimensional datapoints but not the hierarchical information.

3. In other cases, we can observe that a larger λ_2 can better preserve the hierarchical structure in the high-dimensional datapoints but may distort the similarity structure. A larger λ_1 may also lead to a bad visualization of the similarity structure since the KL-divergence might diverge. Nevertheless, CO-SNE is robust to wide choices of λ_1 and λ_2. *Both the KL-divergence and the distance loss are important for producing good visualization*.

References

[1] Manfredo Perdigao do Carmo. *Riemannian geometry*. Birkhäuser, 1992. 1