Poincaré Wasserstein Autoencoder

Ivan Ovinnikov[*]
Department of Computer Science
ETH Zürich
Zürich, Switzerland
ivan.ovinnikov@inf.ethz.ch

Abstract

This work presents a reformulation of the recently proposed Wasserstein autoencoder framework on a non-Euclidean manifold, the Poincaré ball model of the hyperbolic space $\mathbb{H}^n$. By assuming the latent space to be hyperbolic, we can use its intrinsic hierarchy to impose structure on the learned latent space representations. We demonstrate the model in the visual domain to analyze some of its properties and show competitive results on a graph link prediction task.

1 Introduction

Variational Autoencoders (VAE) [11, 17] are an established class of unsupervised machine learning models, which make use of amortized approximate inference to parametrize the otherwise intractable posterior distribution. They provide an elegant, theoretically sound generative model used in various data domains. One of the common issues in generative modeling of images is the sample quality of the generated samples both in terms of sharpness and structural coherence, for instance in samples of visual scenes. A hypothesized cause of this problem is the fact that the VAE latent space is unstructured. There has recently been a number of works which explicitly make use of properties of non-Euclidean geometry in order to perform machine learning tasks. The use of hyperbolic spaces in particular has been shown to yield improved results on datasets which either present a hierarchical tree-like structure such as word ontologies [15] or feature some form of partial ordering [1]. In this work, we propose a Wasserstein autoencoder [21] model which parametrizes a Gaussian distribution in the Poincaré ball model of the hyperbolic space. By treating the latent space as a Riemannian manifold with constant negative curvature, we can use the tree-like hierarchical properties of hyperbolic spaces to impose a structure on the latent space representations.

2 Related Work

There has been a number of extensions to the original VAE framework [11]. These extensions address various problematic aspects of the original model. The first type aims at improving the approximation of the posterior by selecting a richer family of distributions. Some prominent examples include the Normalizing Flow model [16] as well as its derivates [14], [10], [4]. A second direction aims at imposing structure on the latent space by selecting structured priors such as the mixture prior [3], learned autoregressive priors [23] or imposing informational constraints on the objective [8], [24]. The approach most similar to ours but with a hyperspherical latent space and a von-Mises variational distribution has been presented in [2]. The idea of graph generation in hyperbolic space and analysis of complex network properties has been studied in [13].

[*] Affiliated with Computer Graphics Lab

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

3 Model

In order to perform variational inference on the Poincaré Ball, we first need to define an appropriate distribution. Similarly to the VAE, we select a Gaussian distribution with zero mean and unit diagonal covariance as our prior. The Gaussian distribution has a closed form expression on the Poincaré ball [19].

3.1 Gaussian distribution in $\mathbb{H}^n$

Explicit formulation The Gaussian probability distribution function (p.d.f.) in hyperbolic space is defined analogously to the p.d.f. in the Euclidean space. The main difference is the use of the hyperbolic distance in the exponent and a different variance dependent normalization constant which accounts for the underlying geometry (first derived in [19]).

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{Z(\boldsymbol{\sigma})} e^{-\frac{d^2(\mathbf{x}, \boldsymbol{\mu})}{2\boldsymbol{\sigma}^2}} \qquad Z(\boldsymbol{\sigma}) = 2\pi \sqrt{\frac{\pi}{2}} \boldsymbol{\sigma} e^{\frac{\boldsymbol{\sigma}^2}{2}} \operatorname{erf}\left(\frac{\boldsymbol{\sigma}}{\sqrt{2}}\right)$$

on the Poincaré disk model of the hyperbolic space where the expression $d(\mathbf{x}, \boldsymbol{\mu})$ is the geodesic distance function between two points $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{H}^n$. The geodesic distance on the Poincaré ball is defined as

$$d(\mathbf{x}, \boldsymbol{\mu}) = \operatorname{arccosh}\left(1 + 2\frac{||\mathbf{x} - \boldsymbol{\mu}||^2}{(1 - ||\mathbf{x}||^2)(1 - ||\boldsymbol{\mu}||^2)}\right)$$

3.2 Hyperbolic reparametrization trick

The reparametrization trick is a common method to make the sampling operation differentiable by using a differentiable function $g(\epsilon, \theta)$ to obtain a reparametrization gradient for backpropagation through the stochastic layer of the network. For the location-scale family of distributions, the reparametrization function $g(\epsilon, \theta)$ can be written as $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \epsilon$ in the Euclidean space where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We adapt the reparametrization trick to the hyperbolic space by using the framework of gyrovector operators.

Gyrovector spaces In order to perform arithmetic operations on the Poincaré ball model, we rely on the concept of gyrovector spaces, which is a generalization of Euclidean vector spaces to models of hyperbolic space based on Möbius transformations. First proposed by [22], they have been recently used to describe typical neural network operations in the hyperbolic space [6]. In order to perform the location-scale reparametrization in hyperbolic space, we need the gyrovector addition and Hadamard product defined as a diagonal matrix-gyrovector multiplication.

$$\mathbf{x} \oplus \mathbf{y} = \frac{(1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c||\mathbf{y}||^2)\mathbf{x} + (1 - c||\mathbf{x}||^2)\mathbf{y}}{1 + 2c\langle \mathbf{x}, \mathbf{y} \rangle + c^2||\mathbf{x}||^2||\mathbf{y}||^2]}$$

$$M^{\otimes}\mathbf{x} = \frac{1}{\sqrt{c}} \tanh\left(\frac{||M\mathbf{x}||}{||x||} \operatorname{arctanh}(\sqrt{c}||\mathbf{x}||)\right) \frac{M\mathbf{x}}{||M\mathbf{x}||}$$

We then obtain the posterior samples using the following relation:

$$\mathbf{z} = \mu_{\mathcal{H}}(\mathbf{x}) \oplus \operatorname{diag}(\sigma_{\mathcal{H}}(\mathbf{x}))^{\otimes} \mathbf{z}_{\text{prior}}$$

Mean and variance parametrization In order to obtain posterior samples in hyperbolic space, the parametrization of the mean and variance uses a hyperbolic feedforward layer as the last layer of the network (proposed in [6]). The outputs of the underlying Euclidean network $\mathbf{h}$ are projected using the exponential map at the origin: $\exp_{\mathbf{0}}(\mathbf{h}) = \tanh(\sqrt{(c)}||\mathbf{h}||)\frac{\mathbf{h}}{\sqrt{c}||\mathbf{h}||}$ and transformed using the hyperbolic feedforward layer map: $f_h(\mathbf{h}) = \varphi_h(W^{\otimes} \exp_{\mathbf{0}}(\mathbf{h}) \oplus \mathbf{b_h})$ where φ_h is the respective hyperbolic nonlinearity $\varphi_h(\mathbf{x}) = \exp_{\mathbf{0}}(\varphi(\log_{\mathbf{0}} \mathbf{x}))$ for the mean and variance parametrization.

Sampling from prior in hyperbolic space We choose the hyperbolic standard prior $\mathcal{N}_H(0, I)$. as prior $p(\mathbf{z})$. Sampling from a uniform distribution on the Poincaré disk is not trivial, because the uniform distribution is not described by a constant function as it is in Euclidean space. In order to generate samples from the standard prior, we use an approach based on the area ratio of disks in $\mathbb{H}^2$ to obtain the uniform samples on the Poincaré disk [13] and subsequently use a rejection sampling procedure to obtain the Gaussian standard prior. Since we use a mean field approximation of the posterior, we can simply stack low-dimensional samples according to the desired number of latent space dimensions and do not suffer the curse of dimensionality inherent to rejection sampling.

3.3 Optimization

Evidence Lower Bound The variational autoencoder relies on the evidence lower bound (ELBO) reformulation in order to perform tractable optimization of the Kullback-Leibler divergence (KLD) between the true and approximate posteriors. In the Euclidean VAE formulation, the KLD integral has a closed-form expression, which simplifies the optimization procedure considerably. Due to the nonlinearity of the geodesic distance in the exponent, we cannot derive a closed form solution of the expectation expression $\int q_\phi(\mathbf{z}) \log q_\phi(\mathbf{z})$. One possibility is to use a Taylor expansion of the first two moments of the expectation of the squared logarithm $\mathbb{E}_{q_\phi(\mathbf{z}')} \log^2(\mathbf{z}')$. This is however problematic from a numerical standpoint due to the small convergence radius of the Taylor expansion.

Wasserstein metric In the current version of the model in order to circumvent this obstacle, we use a recently proposed Wasserstein Maximum Mean Discrepancy (MMD) metric [7] with an appropriate positive definite RKHS [2] kernel as divergence measure. MMD is known to perform well when matching high-dimensional standard normal distributions [21]. This optimal transport distance measure has been proposed in the context of Wasserstein Autoencoders and aims to address the sample quality of regular VAEs. The WAE objective is derived from the optimal transport cost by relaxing the constraint on the posterior q:

$$\mathcal{L}_{\text{WAE}} = \inf_{q_\phi(\mathbf{z}|\mathbf{x}) \in \mathcal{Q}} \mathbb{E}_{p(\mathbf{x})} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} (\log p(\mathbf{x}|\mathbf{z})) + \beta D_{\text{MMD}} \tag{1}$$

MMD is a metric on the space of probability distributions under the condition that the selected RKHS kernel is characteristic. Geodesic kernels are generally not positive definite, however it has been shown that the Laplacian kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\lambda(d_{\mathcal{H}}(\mathbf{x}, \mathbf{y})))$ is positive definite if the Riemannian metric of the underlying space is constantly negative definite [5]. In particular, this holds for hyperbolic spaces. In practice, there is a high probability that a geodesic RBF kernel is also positive definite depending on the dataset topology [5]. We choose the Laplacian kernel as it also features heavier tails than the Gaussian RBF kernel, which has a positive effect on outlier gradients [21]. The MMD loss function is defined over two probability measures p and q in an RKHS unit ball $\mathcal{F}$ as follows:

$$D_{\text{MMD}}(p(\mathbf{z}), q_\phi(\mathbf{z})) = || \int_{\mathcal{Z}} k(\mathbf{z}, \cdot) dp(\mathbf{z}) - \int_{\mathcal{Z}} k(\mathbf{z}, \cdot) dq(\mathbf{z}) ||_{\mathcal{F}} \tag{2}$$

and a finite sample estimate can be computed based on minibatch samples from the prior $\mathbf{z} \sim p(\mathbf{z})$ via the rejection sampling procedure described in Appendix A and the approximate posterior samples $\bar{\mathbf{z}} \sim q_\phi(\mathbf{z})$ obtained via the hyperbolic reparametrization:

$$D^{(B)}_{\text{MMD}}(p(\mathbf{z}), q_\phi(\mathbf{z})) = \frac{\lambda}{n(n-1)} \sum_{i \neq j} k(\mathbf{z}_i, \mathbf{z}_j) + \frac{\lambda}{n(n-1)} \sum_{i \neq j} k(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j) + \frac{2\lambda}{n^2} \sum_{i,j} k(\mathbf{z}_i, \bar{\mathbf{z}}_j) \tag{3}$$

Parameter updates The hyperbolic geometry of the latent space is defined by a set of hyperbolic parameters $\theta^{\mathcal{H}} \in \mathbb{H}^d$. This requires us to perform Riemannian stochastic gradient descent (RSGD) updates for a subset of the model parameters, namely the bias parameters of $(\boldsymbol{\mu}, \boldsymbol{\sigma})$. We perform exponential map updates using gyrovector arithmetic similar to [6] instead of using a retraction approximation as in [15]. The Euclidean parameters are updated using the Adam optimization procedure [9].

[2] RKHS = Reproducing Kernel Hilbert Space

Table 1: Performance on link prediction datasets

		Model		
Dataset	Metric	$\mathcal{N}$-VGAE	$\mathcal{S}$-VGAE	$\mathcal{H}$-VGAE
Cora	AUC	$92.7_{\pm.2}$	$94.1_{\pm.1}$	$93.9_{\pm.2}$
	AP	$93.2_{\pm.4}$	$94.1_{\pm.3}$	$93.2_{\pm.2}$
Citeseer	AUC	$90.3_{\pm.5}$	$94.7_{\pm.2}$	$92.2_{\pm.2}$
	AP	$91.5_{\pm.5}$	$95.2_{\pm.2}$	$91.8_{\pm.2}$
Pubmed	AUC	$97.1_{\pm.0}$	$96.0_{\pm.1}$	$95.9_{\pm.2}$
	AP	$97.1_{\pm.0}$	$96.0_{\pm.2}$	$96.3_{\pm.2}$

4 Results

4.1 MNIST

In this experiment, we apply our model to the task of generating MNIST digits in order to get an intuition for the properties of the latent hyperbolic geometry. While the MNIST latent space is not inherently hierarchically structured, we can use it to compare our model to the Euclidean VAE approach. We train the models on dynamically binarized MNIST digits and evaluate the generated samples qualitatively as well as quantitatively via the reconstruction error scores. We can observe in Appendix B that the samples present a deteriorating quality as the dimensionality increases despite the lower reconstruction error. This can be explained by the issue of dimension mismatch between the selected latent space dimensionality d_z and the intrinsic latent space dimensionality d_I documented in [18] and can be alleviated by an additional p-norm penalty on the variance. We have not observed a significant improvement by applying the L2-penalty for higher dimensions. We have also performed an experiment using a two-dimensional latent space. We can observe that the structure imposed by the Poincaré disk pushes the samples towards the outside of the disk. This observation can be explained by the fact that hyperbolic spaces grow exponentially (see Appendix B). In order to generate quality samples using the prior, some overlap is required with the approximate posterior in the latent space. The issue is somewhat alleviated in higher dimensions as the distribution shifts towards the ball surface. Moreover, it is possible that the hyperbolic Gaussian is a suboptimal prior choice. A true maximum entropy prior in hyperbolic space might therefore be worthy of investigation.

4.2 Link prediction on citation networks

In this experiment, we aim at exploring the advantages of using a hyperbolic latent space on the task of predicting links in a graph. We train our model on three different citation network datasets: Cora, Citeseer and Pubmed [20]. We use the Variational Graph Auto-Encoder (VGAE) framework [12] and train the model in an unsupervised fashion using a subset of the links. The performance is measured in terms of average precision (AP) and area under curve (AUC) on a test set of links that were masked during training. Table 1 shows a comparison to the baseline with a Euclidean latent space ($\mathcal{N}$-VGAE), showing improvements on the Cora and Citeseer datasets. We also compare our results to the results obtained using a hyperspherical autoencoder ($\mathcal{S}$-VGAE) [2]. It should be noted that we have used a smaller dimensionality for the hyperbolic latent space (16 vs 64 and 32 for the Euclidean and hyperspherical cases respectively), which could be attributed to the fact that a dataset with a hierachical latent manifold requires latent space embeddings of smaller dimensionality to efficiently encode the information (analogously to the results of [15]).

5 Discussion/Conclusion

We have presented an algorithm to perform amortized variational inference on the Poincaré ball model of the hyperbolic space. The algorithm differs significantly from a Gaussian VAE in that it makes use of a Maximum Mean Discrepancy metric instead of the Kullback-Leibler divergence. The underlying geometry of the hyperbolic space allows for an improved performance on tasks which exhibit a partially hierarchical structure. We have discovered certain issues related to the use of the MMD metric in hyperbolic space. Future work will aim to circumvent these issues as well as extend the current results. In particular, we hope to demonstrate the capabilities of our model on more tasks hypothesized to have a latent hyperbolic manifold. We also hope to investigate the use of the MMD loss in more detail and its effect on the placement of the latent space codes as well as experiment with maximum entropy priors.

References

[1] B. P. Chamberlain, J. Clough, and M. P. Deisenroth. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

[2] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

[3] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[4] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[5] A. Feragen, F. Lauze, and S. Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3032–3042, 2015.

[6] O.-E. Ganea, G. Bécigneul, and T. Hofmann. Hyperbolic neural networks. *arXiv preprint arXiv:1805.09112*, 2018.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[8] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, and A. Lerchner. Early visual concept learning with unsupervised deep learning. *arXiv preprint arXiv:1606.05579*, 2016.

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improving variational inference with inverse autoregressive flow.(nips), 2016. *URL http://arxiv.org/abs/1606.04934*.

[11] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[13] D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[14] H. Larochelle and I. Murray. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 29–37, 2011.

[15] M. Nickel and D. Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017.

[16] D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

[17] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[18] P. K. Rubenstein, B. Schoelkopf, and I. Tolstikhin. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018.

[19] S. Said, L. Bombrun, and Y. Berthoumieu. New riemannian priors on the univariate normal model. *Entropy*, 16(7):4015–4031, 2014.

[20] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[21] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[22] A. A. Ungar. A gyrovector space approach to hyperbolic geometry. *Synthesis Lectures on Mathematics and Statistics*, 1(1):1–194, 2008.

[23] A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

[24] S. Zhao, J. Song, and S. Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

A Hyperbolic standard prior sampling

```
Input: maximum radius $r_{max}$, dimensionality $d$, hyperbolic prior likelihood $f_{\mathcal{H}}(\mathbf{x}|\mathbf{0}, \mathbf{I}_2)$
Result: $k = nd$ samples from prior $f_{\mathcal{H}}(\mathbf{z})$
while $i < k$ do
    sample angle $\phi \sim \mathcal{U}(0, 2\pi)$, sample $a \sim \mathcal{U}(0, 1)$;
    get radius sample $r = \text{acosh}(1 + a(\cosh r_{max} - 1))$;
    generate pairs $\mathbf{x}_i = (\sinh r \cos\phi, \sinh r \sin\phi)$;
    evaluate $p(\mathbf{x}_i) = f_{\mathcal{H}}(\mathbf{x}_i)$;
    $M = \max(p_i)$;
    sample $u \sim \mathcal{U}(0, 1)$;
    if $u < \frac{p_i}{M}$ then
        accept sample $\mathbf{x}_i$;
    else
        reject sample;
    end
end
Output: stack $d$ dimensions from $\mathbb{H}^2$-samples: $\mathbf{s} = [s_1, s_2, ..., s_d]$;
```

Algorithm 1: Prior sampling on Poincaré Ball

B Hyperbolic geometry: a short overview

B.1 Riemannian geometry

Hyperbolic spaces are one of three existing types of isotropic spaces: the Euclidean spaces with zero curvature, the spherical spaces with constant positive curvature and the hyperbolic spaces which feature constant negative curvature.

Poincaré ball The Poincaré ball is one of the five isometric models of the hyperbolic space. The model is defined by the tuple $(\mathcal{B}_n, g_H)$ which corresponds to the manifold $\mathcal{B}_n$ equipped with the Riemannian metric g_H:

$$\mathcal{B}_n = \{\mathbf{x} \in \mathbb{R}^n \mid ||\mathbf{x}|| < 1\} \quad g_H = \frac{4}{(1 - ||\mathbf{x}||^2)^2}$$

For comparison, an analogous metric on the Euclidean is given by $g_E = I_n$. For every point x on a given manifold $\mathcal{M}$, a tangent space $\mathcal{T}_x\mathcal{M}$ is defined, corresponding to a first order approximation of $\mathcal{M}$ at point x. The Riemannian metric g is a collection of inner products $\mathcal{T}_x\mathcal{M} \times \mathcal{T}_x\mathcal{M} \to \mathbb{R}$. It allows the definition of path integrals $\int_a^b \sqrt{g_H(\dot{\alpha}(t), \dot{\alpha}(t))}dt$ along smooth curves $\alpha(t)$ of the manifold, resulting in geodesic lines. A *geodesic* is a smooth curve $\alpha(t)$ which corresponds to the shortest distance between two points on a manifold. The *exponential map* $\exp_{\mathbf{x}}(\mathbf{v})$ gives a way to project a vector $\mathbf{v}$ on the tangent space at point x to the corresponding point on the manifold.

B.2 Connection to network topology

We can understand why hyperbolic spaces are particularly efficient at modeling hierachical tree-structured data by making the following observation. The number of children nodes in a tree with branching factor b is $(b+1)b^l$ at level l from the root and $((b+1)b^l - 2)/(b-1)$ at levels closer to the root. Thus, this number grows exponentially with the distance from the root. Both circle lengths and areas grow exponentially in hyperbolic spaces $L = 2\pi \text{sinh}(r)$, $A = 2\pi(\text{cosh}(r) - 1)$. Hence, hyperbolic spaces allow to *accomodate* the number of nodes in an efficient manner and can informally be thought of continuous trees. In fact, trees can be embedded into hyperbolic spaces nearly isometrically. A similar construction in Euclidean space would require a larger dimensionality and more parameters. Intuitively, by choosing an underlying hyperbolic geometry, we learn a distribution of latent space embeddings which incorporate an approximate tree-like structure in their representations.

C Visual Samples

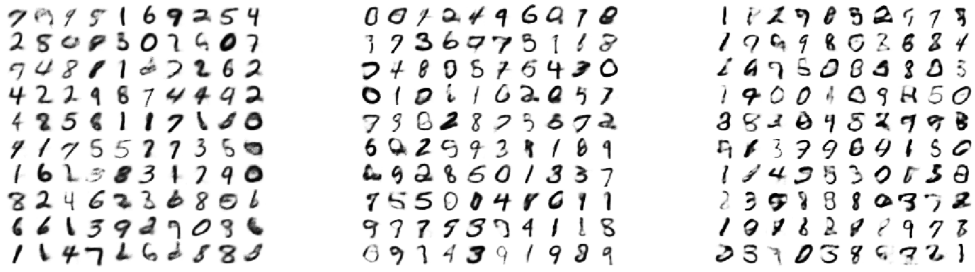

Figure 1: Euclidean VAE samples $d \in \{5, 10, 20\}$, reconstruction error $L \in \{109.01, 94.58, 93.36\}$

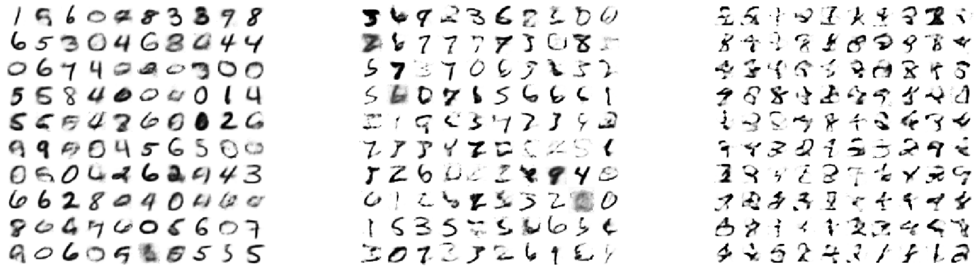

Figure 2: Poincaré WAE samples $d \in \{5, 10, 20\}$, reconstruction error $L \in \{95.01, 69.70, 58.58\}$

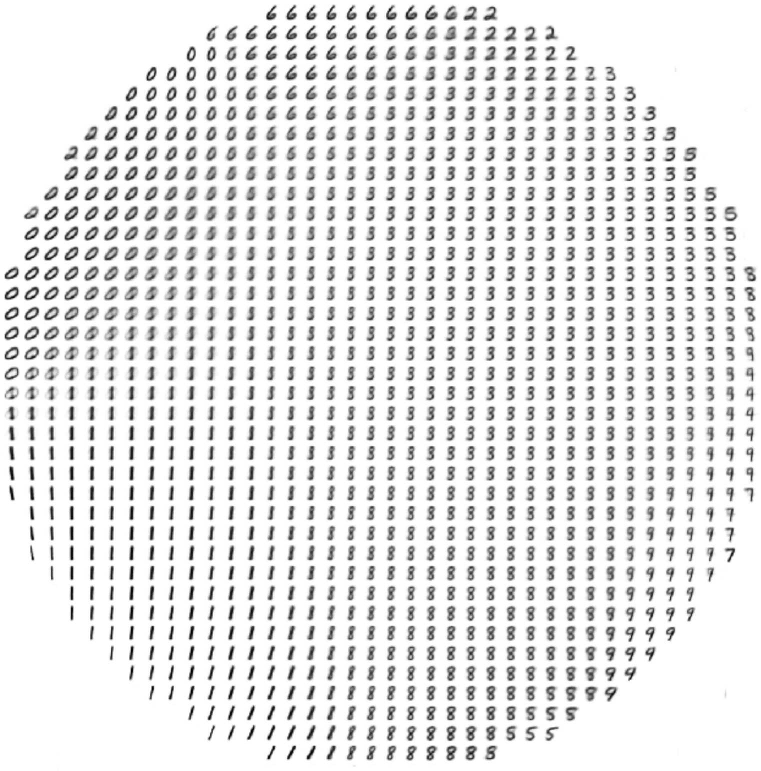

Figure 3: Poincaré WAE samples from two-dimensional latent space