Hyperbolic VAE via Latent Gaussian Distributions

Seunghyuk Cho [1] **Juyong Lee** [1] **Dongwoo Kim** [1]

Abstract

We propose a Gaussian manifold variational auto-encoder (GM-VAE) whose latent space consists of a set of Gaussian distributions. It is known that the set of the univariate Gaussian distributions with the Fisher information metric form a hyperbolic space, which we call a Gaussian manifold. To learn the VAE endowed with the Gaussian manifolds, we propose a pseudo-Gaussian manifold normal distribution based on the Kullback-Leibler divergence, a local approximation of the squared Fisher-Rao distance, to define a density over the latent space. In experiments, we demonstrate the efficacy of GM-VAE on two different tasks: density estimation of image datasets and environment modeling in model-based reinforcement learning. GM-VAE outperforms the other variants of hyperbolic- and Euclidean-VAEs on density estimation tasks and shows competitive performance in model-based reinforcement learning. We observe that our model provides strong numerical stability, addressing a common limitation reported in previous hyperbolic-VAEs.

1. Introduction

The geometry of latent space in generative models, such as variational auto-encoders (VAE), reflects the structure of the data representations. Mathieu et al. (2019); Nagano et al. (2019); Cho et al. (2022) show that employing hyperbolic space as the latent space improves the preservation of the hierarchical structure within the data. The theoretical background for adopting hyperbolic space lies in the analysis of Sarkar (2011); the tree-structured data can be embedded with arbitrary low distortion in hyperbolic space, while Euclidean space requires extensive dimensions.

Previously proposed hyperbolic VAEs rely on Poincaré normal distribution (Mathieu et al., 2019) or hyperbolic wrapped normal distribution (Nagano et al., 2019) for the prior and variational distributions. Unlike the Gaussian distribution in Euclidean space, however, these distributions suffer from several shortcomings, including the absence of closed-form Kullback-Leibler (KL) divergence, numerical instability (Mathieu et al., 2019; Skopek et al., 2019), and high computational cost in sampling (Mathieu et al., 2019).

Meanwhile, we can form a Riemannian manifold from the set of univariate Gaussian distributions by equipping the Fisher information metric (FIM). It is known that the FIM of univariate Gaussian distributions is akin to that of the metric tensor of the Poincaré half-plane model (Costa et al., 2015), providing a perspective of viewing the points in hyperbolic space as univariate Gaussian distributions. In other words, a Gaussian distribution can be mapped to a single point in the open half-plane manifold , where the shortest geodesic distance between two Gaussian distributions is formed by the FIM. Noting that the numerical issue of Poincaré normal arises from the geodesic distance of hyperbolic space, we question whether this perspective can lead us to define a new distribution with better analytic properties.

In this work, inspired by the fact that KL divergence itself is a statistical distance that locally approximates the geodesic distance (Tifrea et al., 2018), we propose a hyperbolic distribution by substituting the geodesic distance of Poincaré normal with the KL divergence between the univariate Gaussian distributions. We then verify that this simple yet powerful alteration results in several practical analytic properties; the proposed distribution reduces into the product of two well-known distributions, i.e., the Gaussian and gamma distributions which are easy to sample and there is a closed-form KL divergence between the proposed distributions. By adopting the proposed hyperbolic distribution, we introduce a new variant of hyperbolic VAE, named Gaussian manifold VAE (GM-VAE), whose latent space is a set of Gaussian distributions.

During the experiments, we observe that the proposed distribution is robust in terms of sampling and KL divergence computation compared to the commonly-used hyperbolic distributions; we briefly explain the reason why others are numerically unstable. Experimental results on the density estimation task with image datasets show that GM-VAE

[1] Pohang University of Science and Technology. Correspondence to: Dongwoo Kim <dongwookim@postech.ac.kr>.

Presented at the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

can achieve outperforming generalization performances to unseen data against baselines of Euclidean and hyperbolic VAEs. Application of GM-VAE on model-based reinforcement learning (RL) verifies the feasibility of using hyperbolic space on another domain of task.

2. Gaussian Manifold VAE

In this section, we present the concept of the Gaussian manifold, propose a pseudo Gaussian manifold normal distribution, and suggest a new variant of the VAE defined over the Gaussian manifold with PGM normal as prior.

2.1. Gaussian manifold with arbitrary curvature

We show that the Riemannian manifold of univariate Gaussian distributions can have an arbitrary constant negative curvature by reparameterizing the univariate Gaussian distribution properly. Let the univariate Gaussian distribution as $\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu-x)^2}{2\sigma^2}\right)$. We reparameterize the distribution with additional parameter $c > 0$ as $\mathcal{N}(\sqrt{2c}\mu, \sigma^2)$. The reparameterization leads to the FIM of $\sigma^{-2}\mathrm{diag}(1, 1/c)$ showing that the curvature is $-c$.

We call the Riemannian manifold with the reparameterized univariate Gaussians and the extended FIM as the Gaussian manifold and denote it as $\mathcal{G}_c$, where $-c$ is the curvature. We then verify that the KL divergence between the points of the Gaussian manifold approximates the geodesic distance, even in the presence of arbitrary curvature in the Gaussian manifold. Let $(\mu, \sigma) \in \mathcal{G}_c$ be an arbitrary point of the Gaussian manifold. The KL divergence between (μ, σ) and its neighbor $(\mu + d\mu, \sigma + d\sigma)$ can be computed as:

$$\frac{D_{\mathrm{KL}}\left(\mathcal{N}(\sqrt{2c}(\mu+d\mu), (\sigma+d\sigma)^2) \parallel \mathcal{N}(\sqrt{2c}\mu, \sigma^2)\right)}{2c}$$
$$= \frac{1}{2}\begin{pmatrix} d\mu \\ d\sigma \end{pmatrix}^T \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{c\sigma^2} \end{pmatrix} \begin{pmatrix} d\mu \\ d\sigma \end{pmatrix} + \mathcal{O}\left((d\sigma)^3\right), \qquad (1)$$

where the first term is the squared Riemannian norm of the tangent vector $(d\mu, d\sigma)$ approximating the squared Fisher-Rao distance.

2.2. Pseudo Gaussian manifold normal distribution

We propose a pseudo Gaussian manifold (PGM) normal distribution defined over the Gaussian manifold. Let $(\mu, \sigma) \in \mathcal{G}_c$ be a point in the Gaussian manifold. Inspired by the Riemannian normal distribution, we define the probability density function of PGM normal with the KL divergence as:

$$\mathcal{K}_c(\mu, \sigma; \alpha, \beta, \gamma) = \frac{\sigma^3}{Z(c, \beta, \gamma)} \qquad (2)$$
$$\times \exp\left(-\frac{D_{\mathrm{KL}}(\mathcal{N}(\sqrt{2c}\cdot\mu, \sigma^2) \parallel \mathcal{N}(\sqrt{2c}\cdot\alpha, \beta^2))}{2c\cdot\gamma^2}\right),$$

where $(\alpha, \beta) \in \mathcal{G}_c$, and $\gamma \in \mathbb{R}_{>0}$ are the parameters of the distribution. The distribution is centered at (α, β) with additional scale parameter γ. As shown in Equation 1, the KL divergence of the Gaussian manifold approximates the Fisher-Rao distance between $\mathcal{N}(\sqrt{2c}\cdot\alpha, \beta^2)$ and $\mathcal{N}(\sqrt{2c}\cdot\mu, \sigma^2)$. Therefore, the PGM normal accounts for the geometric structure of the univariate Gaussian distributions.

The factorization of the probability density function in Equation 2 multiplied with the square root of the determinant of the FIM shows the advantages of the PGM normal, which can be written as:

$$\mathcal{K}_c(\mu, \sigma; \alpha, \beta, \gamma) \cdot \sqrt{\det(g)} \qquad (3)$$
$$= \mathcal{N}(\mu; \alpha, \beta^2\gamma^2) \cdot 2\sigma \,\mathrm{Gamma}\left(\sigma^2; \frac{1}{4c\gamma^2} + 1, \frac{1}{4c\beta^2\gamma^2}\right),$$

where $\mathrm{Gamma}(z; a, b) = \frac{b^a}{\Gamma(a)} z^{a-1} \exp(-bz)$ and g is the FIM of the Gaussian manifold. Thanks to the properties of Gaussian and gamma distributions, the PGM normal is easy to sample and has a closed-form KL divergence.

The factorization has the same form as the well-known conjugate prior to the Gaussian distribution. In that sense, the PGM normal explicitly incorporates the geometric structure between Gaussians into the known prior distribution.

We note that the PGM normal can be easily extended for the diagonal Gaussian manifold, a manifold formed by diagonal Gaussian distributions since the diagonal Gaussian manifold is the product of the Gaussian manifolds.

2.3. Gaussian manifold VAE

We introduce a Gaussian manifold VAE (GM-VAE) whose latent space is defined over the diagonal Gaussian manifold with the help of the PGM normal. We use the PGM normal for variational and prior distributions.

To be specific, with the PGM normal, the evidence lower bound (ELBO) of the GM-VAE can be formalized with the diagonal Gaussian manifold $\{(\boldsymbol{\mu}, \Sigma) \mid \boldsymbol{\mu} \in \mathbb{R}^n, \Sigma \in \mathbb{R}^n_{>0}\}$ as:

$$\mathbb{E}_{q_\phi(\boldsymbol{\mu},\Sigma|\mathbf{x})\cdot\sqrt{\det(g)}}\left[\log p_\theta(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)\right]$$
$$- D_{\mathrm{KL}}\left(q_\phi(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}) \cdot \sqrt{\det(g)} \parallel p(\boldsymbol{\mu}, \Sigma) \cdot \sqrt{\det(g)}\right),$$

where $p_\theta(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$ is the decoder network, $q_\phi(\boldsymbol{\mu}, \Sigma \mid \mathbf{x})$ is the encoder network and $p(\boldsymbol{\mu}, \Sigma)$ is the prior. The variational distribution is set to $q_\phi(\boldsymbol{\mu}, \Sigma \mid \mathbf{x}) = \mathcal{K}_c(\alpha_\phi(\mathbf{x}), \beta_\phi(\mathbf{x}), \gamma_\phi(\mathbf{x}))$, where $\alpha_\theta(\mathbf{x}) \in \mathbb{R}^n$ and $\beta_\phi(\mathbf{x}), \gamma_\phi(\mathbf{x}) \in \mathbb{R}^n_{>0}$, and the prior is set to $p(\boldsymbol{\mu}, \Sigma) = \mathcal{K}_c(\mathbf{0}, I, I)$ in our experiments given curvature $-c$. The pseudo-algorithm for the encoder and decoder of GM-VAE is present at Algorithm 1, respectively.

```
Algorithm 1 Decoder
Input Parameter $(\alpha, \beta) \in \mathcal{G}_c, \gamma$, Decoding layers $\mathrm{Dec}(\cdot)$
Output Reconstruction $\mathbf{x}'$
  1: Sample $\mu \sim \mathcal{N}(\alpha, \beta\gamma)$
  2: Sample $\sigma^2 \sim \mathrm{Gamma}\left(\frac{1}{4c\gamma^2} + 1, \frac{1}{4c\beta^2\gamma^2}\right)$
  3: $\mathbf{x}' = \mathrm{Dec}([\mu, \sigma])$
  4: return $\mathbf{x}'$
```

3. Related Work

The latent space of VAE reflects the geometrical property of the representations of the data. Hyperbolic space as the latent space of the VAE has been adopted in several works by proposing a tractable distribution over hyperbolic space. For example, Nagano et al. (2019) suggest hyperbolic wrapped normal distribution (HWN) from the observation that the tangent space is Euclidean space. Leveraging operations defined on the tangent spaces, e.g., parallel transport, enables an easy sampling algorithm. Also, Mathieu et al. (2019) propose a rejection sampling method for the Riemannian normal distribution defined on the Poincaré disk model, namely Poincaré normal distribution.

These distributions are studied in many cases (Skopek et al., 2019; Mathieu & Nickel, 2020; Cho et al., 2022) but suffer from instability because of the absence of closed-form KL divergence. Our proposed distribution, however, not only shares the common merits but also overcomes the stability problem with closed-form KL divergence. Our method enjoys easy sampling and provides closed-form KL divergence while utilizing the geometric structure of the statistical manifold, i.e., the use of (approximated) geodesic distance. Table 1 summarizes the properties of each distribution.

4. Experiments

In this section, we demonstrate the performances of GM-VAE on two tasks: density estimation of image datasets and model-based RL. We remark on the practical properties of GM-VAE shown in the experiments with additional analyses.

Table 1. A comparison of the PGM normal (PGM-$\mathcal{N}$) with HWN and Poincaré normal (Poincaré-$\mathcal{N}$). Geometry denotes whether the distribution utilizes the geometric structure. Tractable KL indicates whether the distribution has a closed-form KL divergence.

	Sampling	Geometry	Tractable KL
HWN	low-cost	×	×
Poincaré $\mathcal{N}$	expensive	○	×
PGM-$\mathcal{N}$	low-cost	○	○

Table 2. Density estimation on real-world datasets. d denotes the latent dimension. We report the negative test log-likelihoods of average 10 runs for Breakout, CUB, Food101, and Oxford102 with 95% confidence interval. N/A in the log-likelihood indicates that the results are not available due to the failure of all runs, and N/A in the standard deviation indicates the results are not available due to failures of some runs. The best results are bolded.

	d	$\mathcal{E}$-VAE	$\mathcal{L}$-VAE	$\mathcal{P}$-VAE	GM-VAE
Breakout	2	$124.74_{\pm 0.86}$	$122.58_{\mathrm{N/A}}$	$270.05_{\pm 2.84}$	$\mathbf{121.52_{\pm 1.00}}$
	4	$66.39_{\pm 0.76}$	$66.70_{\pm 0.32}$	$271.73_{\pm 42.95}$	$\mathbf{65.83_{\pm 0.49}}$
	8	$\mathbf{44.97_{\pm 0.37}}$	$45.25_{\pm 0.27}$	$81.55_{\pm 64.61}$	$45.14_{\pm 0.30}$
CUB	50	$992.05_{\pm 1.38}$	$993.03_{\pm 1.64}$	$990.49_{\pm 2.26}$	$\mathbf{985.46_{\pm 3.82}}$
	60	$969.99_{\pm 3.13}$	$968.79_{\pm 3.70}$	$964.02_{\pm 3.55}$	$\mathbf{958.00_{\pm 3.25}}$
	70	$949.13_{\pm 2.72}$	$948.88_{\pm 3.19}$	$944.24_{\pm 4.40}$	$\mathbf{939.08_{\pm 3.12}}$
Food101	50	$1297.81_{\pm 4.51}$	$1298.45_{\pm 6.32}$	$1293.26_{\pm 7.14}$	$\mathbf{1286.30_{\pm 6.19}}$
	60	$1224.03_{\pm 8.31}$	$1227.16_{\pm 5.18}$	$1218.09_{\pm 3.88}$	$\mathbf{1213.31_{\pm 3.88}}$
	70	$1164.95_{\pm 3.80}$	$1165.39_{\pm 5.54}$	$1165.91_{\pm 4.91}$	$\mathbf{1152.80_{\pm 3.35}}$
Oxford102	50	$1297.41_{\pm 2.69}$	$1296.41_{\pm 1.56}$	$1294.12_{\pm 1.80}$	$\mathbf{1292.90_{\pm 3.43}}$
	60	$1253.80_{\pm 2.57}$	$1256.52_{\pm 2.99}$	$1251.77_{\pm 1.82}$	$\mathbf{1245.49_{\pm 2.18}}$
	70	$1231.52_{\pm 3.18}$	$1229.38_{\pm 3.44}$	$1219.75_{\pm 1.72}$	$\mathbf{1215.07_{\pm 2.52}}$

4.1. Density estimation on image datasets

We conduct density estimation on image datasets to measure the effectiveness of hyperbolic latent space against Euclidean space with the proposed GM-VAE. We use four datasets: the images from Atari2600 Breakout with binarization (Breakout), Caltech-UCSD Birds-200-2011 (CUB), Food101, and Oxford 102 Flower (Oxford102). The datasets are chosen with the four lowest δ-H among the vision datasets from torchvision.

We compare GM-VAE with the three baseline models: VAE with Euclidean latent space ($\mathcal{E}$-VAE), and hyperbolic VAE equipped with HWN ($\mathcal{L}$-VAE) and Poincaré normal ($\mathcal{P}$-VAE). We use the product latent space for both $\mathcal{L}$-VAE and $\mathcal{P}$-VAE.

The results are reported at Table 2. GM-VAE outperforms the baselines in all the settings, except one case of Breakout. Especially in CUB and Oxford102, GM-VAE outperforms the baselines regardless of the curvature value. In Breakout, $\mathcal{P}$-VAE shows inferior performance due to unstable training, and $\mathcal{L}$-VAE fails in some of the runs with small latent dimension.

4.2. Model-based RL

We focus on the model-based RL task to verify whether GM-VAE can also be useful in other domains. Specifically, we focus on applying GM-VAE to the world model, which aims to model the learning environment of the RL agents.

We use DreamerV2 (Hafner et al., 2020) as the baseline to evaluate the performance of GM-VAE in modeling environments. GM-VAE is employed by replacing the stochastic state space z_t with the Gaussian manifold and two components in RSSM, the representation model $q_\theta(z_t|h_t, x_t)$ and

Table 3. Model-based RL results on the 6 games of the Atari2600 environment. We compare the methods of using Euclidean, discrete, and hyperbolic latent space. We report averaged rewards over 4 runs. The best reward for each game is bolded.

Latent space	Euc.	Disc.	Hyp.	$\delta-$H
Breakout	**329.0**	256.8	319.3	0.12
Alien	3412.5	3120.0	**3485.0**	0.14
Zaxxon	34275	38825	**38950**	0.14
Ice Hockey	**25.50**	11.80	20.75	0.14
Freeway	32.8	**33.0**	**33.0**	0.38
Krull	53290	36135	**66185**	0.38

transition predictor $p_\phi(z_t|h_t)$, with PGM normal.

We conduct a comparison of evaluation scores between different types of latent space on world model learning over the Atari2600 environments. The agents are trained with 100M environment steps. We select games having the δ-H values of the four lowest and the two highest among 60 popular Atari2600 games.

The results are reported at Table 3 and Figure 1. The GM-VAE shows competitive results with the baselines of employing Euclidean and discrete latent space in all the games we test. We note that the reproduced Euclidean baseline results by using the official code are better than those reported in Hafner et al. (2020). We also conduct experiments with another hyperbolic distribution, i.e., HWN, but we could not train the world model properly because of the numerical stability issue.

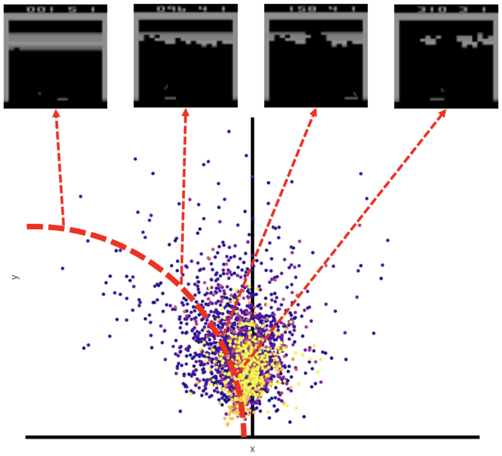

Figure 1. The dots from yellow to purple represent the latent states from the world model in the Atari2600 Breakout with decreasing rewards. Along the red geodesic dashed line passing, we sample for images to visualize the learned representations. We can observe a hierarchical structure along the geodesic.

5. Conclusion & Future Work

In this work, we propose a novel method of representation learning with GM-VAE, utilizing the Gaussian manifold for the latent space. With the newly-proposed PGM normal defined over the Gaussian manifold, which shows better stability and ease of sampling compared to the commonly-used ones, we verify the efficacy of our method on several tasks. Our method achieves outperforming results on density estimation with image datasets and competitive results on model-based RL compared to the baselines. We explain the behavior of GM-VAE in terms of solving the frequent numerical issue of commonly-used hyperbolic VAEs. The analysis of latent space exhibits that the hierarchy lying in the dataset can be preserved by using GM-VAE.

We believe that the connection between the statistical manifold and hyperbolic space provides new insight to the research community and hope to see more interesting connections and analyses in the future.

References

Cho, S., Lee, J., Park, J., and Kim, D. A rotated hyperbolic wrapped normal distribution for hierarchical representation learning. In *NeurIPS*, 2022.

Costa, S. I., Santos, S. A., and Strapasson, J. E. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

Hafner, D., Lillicrap, T., Norouzi, M., and Ba, J. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Mathieu, E. and Nickel, M. Riemannian continuous normalizing flows. *NeurIPS*, 2020.

Mathieu, E., Le Lan, C., Maddison, C. J., Tomioka, R., and Teh, Y. W. Continuous hierarchical representations with poincaré variational auto-encoders. *NeurIPS*, 2019.

Nagano, Y., Yamaguchi, S., Fujita, Y., and Koyama, M. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *ICML*, 2019.

Sarkar, R. Low distortion delaunay embedding of trees in hyperbolic plane. In *Proceedings of the 19th International Conference on Graph Drawing*, pp. 355–366. Springer-Verlag, 2011.

Skopek, O., Ganea, O.-E., and Bécigneul, G. Mixed-curvature variational autoencoders. *arXiv preprint arXiv:1911.08411*, 2019.

Tifrea, A., Bécigneul, G., and Ganea, O.-E. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.