LATENT VARIABLES ON SPHERES FOR SAMPLING AND INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Variational inference is a fundamental problem in Variational AutoEncoder (VAE). The optimization with lower bound of marginal log-likelihood results in the distribution of latent variables approximate to a given prior probability, which is the dilemma of employing VAE to solve real-world problems. By virtue of high-dimensional geometry, we propose a very simple algorithm completely different from existing ones to alleviate the variational inference in VAE. We analyze the unique characteristics of random variables on spheres in high dimensions and prove that Wasserstein distance between two arbitrary data sets randomly drawn from a sphere are nearly identical when the dimension is sufficiently large. Based on our theory, a novel algorithm for distribution-robust sampling is devised. Moreover, we reform the latent space of VAE by constraining latent variables on the sphere, thus freeing VAE from the approximate optimization of posterior probability via variational inference. The new algorithm is named Spherical AutoEncoder (SAE). Extensive experiments by sampling and inference tasks validate our theoretical analysis and the superiority of SAE.

1 INTRODUCTION

Deep generative models, such as variational autoencoder (VAE) (Kingma & Welling, 2013; Rezende et al., 2014) and generative adversarial network (GAN) (Goodfellow et al., 2014), play more and more important role in machine learning and computer vision. However, the problem of variational inference in VAE is still challenging, especially for high-dimensional data like images.

To be formal, let $\mathbb{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ denote the set of observable data points and $\mathbb{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$ the set of desired latent vectors, where $\boldsymbol{x}_i \in \mathbb{R}^{d_x}$ and $\boldsymbol{z}_i \in \mathbb{R}^{d_z}$. Let $p_g(\boldsymbol{x}|\boldsymbol{z})$ denote the likelihood of generated sample conditioned on latent variable $\boldsymbol{z}$ and $p(\boldsymbol{z})$ the prior, where g denotes the decoder . The encoder f in VAE approximates the posterior $q_f(\boldsymbol{z}|\boldsymbol{x})$ in light of the lower bound of the marginal log-likelihood

$$\log p_g(\boldsymbol{x}) = \log \int p_g(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} = \log \int \frac{q_f(\boldsymbol{z}|\boldsymbol{x})}{q_f(\boldsymbol{z}|\boldsymbol{x})} p_g(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z} \tag{1}$$

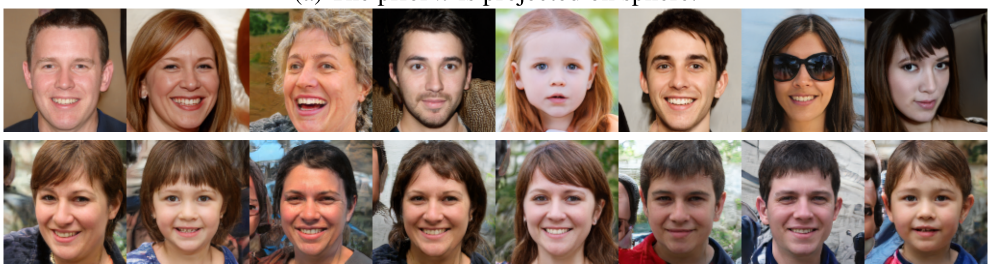
$$\geq -D_{\mathrm{KL}}[q_f(\boldsymbol{z}|\boldsymbol{x}) || p(\boldsymbol{z})] + \mathbb{E}_q[\log p_g(\boldsymbol{x}|\boldsymbol{z})]. \tag{2}$$

The first term $D_{\mathrm{KL}}[q_f(\boldsymbol{z}|\boldsymbol{x}) || p(\boldsymbol{z})]$ constrains the encoded latent codes to the prior via the KL-divergence, and the second term $\mathbb{E}_q[\log p_g(\boldsymbol{x}|\boldsymbol{z})]$ serves to guaranteeing the reconstruction accuracy of inputs. For a Gaussian $p_g(\boldsymbol{x}|\boldsymbol{z})$ of diagonal covariance matrix, $\log p_g(\boldsymbol{x}|\boldsymbol{z})$ reduces to the variance-weighted squared error (Doersch, 2016).

The lower bound approximation of the log-likelihood yields a feasible solution for VAE. But it also causes the new problems. For example, the generated sample $g(\boldsymbol{z})$ deviates from the real distribution of $\mathbb{X}$ when sampling from the given prior due to that the learned $q_f(\boldsymbol{z}|\boldsymbol{x})$ cannot match the prior distribution precisely. Besides, the reconstruction $g(f(\boldsymbol{x}))$ is not satisfactory either. For imagery data, bluriness usually occurs.

In order to manipulate real images for GAN models, we usually need to formulate an encoder via the framework of VAE. The variational inference also applies in this scenario. The problems of VAE are the obstacles of putting the GAN encoder in the right way either. There are other methods of learning an encoder for GAN in the adversarial way such as (Dumoulin et al., 2017; Li et al., 2017;

(a) The prior z is projected on sphere.



(b) The prior z is not projected on sphere.

Figure 1: Generated faces by StyleGAN with different priors. For (a) and (b), the first row shows the generated faces with normal distribution and the second row displays the ones with the uniform distribution. The priors are both normal when training.

Ulyanov et al., 2017; Heljakka et al., 2018). However, the structural precision of reconstruction is generally inferior to the VAE framework, because high-level semantics of objects outweigh low-level structures for such methods (Donahue & Simonyan, 2019). Besides, the concise architecture of VAE is more preferred in this scenario. Therefore, learning precise latent variables $z = f(x)$ is critical to applications of VAE and GAN.

Using a different theory in this paper, we propose a simple method to circumvent the problem. Our contributions are summarized as follows. 1) We introduce the volume concentration of high-dimensional spheres. Based on the concentration property, we point out that projecting on a sphere for data within the sphere produces little difference from the viewpoint of volume in high-dimensional spaces. Thus, it is feasible to approximate the probability optimization pertaining to VAE on the sphere. 2) We further analyze the probability distribution of distances between two sets of random points on the sphere in high dimensions and illustrate the phenomenon of distance convergence. Furthermore, we prove that the Wasserstein distance between two arbitrary datasets randomly drawn from the high-dimensional sphere are nearly identical, meaning that data on the sphere are distribution-robust for generative models with respect to Wasserstein distance. 3) Based on our theoretical analysis, we propose a very simple algorithm for sampling generative models. The same principle is also harnessed to reformulate VAE. We remove the optimization of posterior probability by simply putting spherical normalization on latent variables. We name this autoencoder the spherical autoencoder (SAE). 4) We perform extensive experiments to validate our theoretical analysis and claims with sampling and inference tasks.

2 Latent variables on sphere

For latent variables or data points sampled from some priors, the projection on the unit sphere can can be easily performed by

$$z \leftarrow z/\|z\|. \tag{3}$$

This spherical normalization for priors fed into generator is employed in StyleGAN that is the phenomenal algorithm in GANs (Karras et al., 2018b). To test the robustness of StyleGAN against the diverse distributions, we conduct two groups of experiments with the input z sphere-normalized and not sphere-normalized when training StyleGAN. As shown in Figure 1, the diversity of generated faces is good for two different distributions with normalized z, whereas the face modes become similar for the case of uniform distribution that z is not normalized. This experiment indicates that

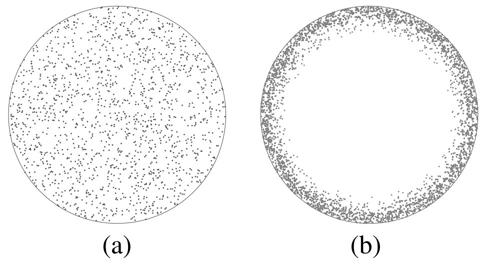


Figure 2: Volume of spheres in different dimensional spaces. (a) Low-dimensional space. (b) High-dimensional space. The volume of the sphere in the high-dimensional space is highly concentrated near the surface. The interior is nearly empty.

StyleGAN with sphere-normalized $\boldsymbol{z}$ is much more robust to the variation of variable modes from different distributions.

Inspired by this comparison, we interpret the benefit of using random variables on spheres by virtue of high-dimensional geometry in this section. Based on these theories, a novel algorithm is proposed for random sampling and latent inference for GAN and VAE.

2.1 Volume concentration

For high-dimensional spaces, there are many counter-intuitive phenomena that will not happen in low-dimensional spaces. For convenient analysis, we assume that the center of the sphere $\mathbb{S}^d$ embedded in $\mathbb{R}^{d+1}$ is at the origin. We first present the concentration property of sphere volume in $\mathbb{R}^{d+1}$. One can find the proof in (Blum et al., 2020).

Theorem 1. *Let $V(r)$ and $V((1-\epsilon)r)$ denote the volumes of the two concentric spheres of radius r and $(1-\epsilon)r$, respectively, where $0 < \epsilon < 1$. Then*

$$\frac{V((1-\epsilon)r)}{V(r)} = (1-\epsilon)^d. \tag{4}$$

And if $\epsilon = \frac{t}{d}$, $\frac{V((1-\epsilon)r)}{V(r)} \to e^{-t}$ when $d \to \infty$, where t is a constant.

Theorem 1 says that the volume of d-dimensional sphere of radius $(1-\epsilon)r$ rapidly goes to zero when d goes large, meaning that the interior of the high-dimensional sphere is empty. In other words, nearly all the volume of the sphere is contained in the thin annulus of width $\epsilon r = rt/d$. The width becomes very thin when d grows. For example, the annulus of width that is 0.9% of the radius contains 99% of all the volume for the sphere in $\mathbb{R}^{512}$. To help understand this counter-intuitive geometric property, we make a schematic illustration in Figure 2.

The probabilistic manipulations are very beneficial from volume concentration of spheres. Suppose that we perform probabilistic optimizations pertaining to latent variables contained in a sphere. The probability mass is also empty due to volume concentration. Therefore, the error is controllable if we perform it on the sphere, provided that these latent variables lie in high-dimensional spaces. For VAE, therefore, we can write a feasible approximation

$$\log p_g(\boldsymbol{x}) = \log \int_{\text{Int}(\mathbb{S}^d)} p_g(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \approx \log \int_{\mathbb{S}^d} p_g(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}, \tag{5}$$

where Int($\mathbb{S}^d$) denotes the interior of $\mathbb{S}^d$, $\|\boldsymbol{z}\| < r$, and d is sufficiently large. The spherical approximation for $\log p_g(\boldsymbol{x})$ is an alternative scheme except the lower bound approximation presented in equation (1). In fact, the distributions defined on the sphere have been already exploited to re-formulate VAE, such as the von Mises-Fisher distribution (Davidson et al., 2018; Xu & Durrett, 2018). But the algorithms proposed in (Davidson et al., 2018; Xu & Durrett, 2018) still fall into the category using lower bound approximation of log-likelihood as the vanilla VAE suffers. To eliminate this constraint, we need more geometric analysis presented in the following section.

2.2 Distance convergence

To dig deeper, we examine the pairwise distance between two arbitrary points randomly sampled on $\mathbb{S}^d$. The following important lemma was proved by (Lord, 1954; Lehnen & Wesenberg, 2002).

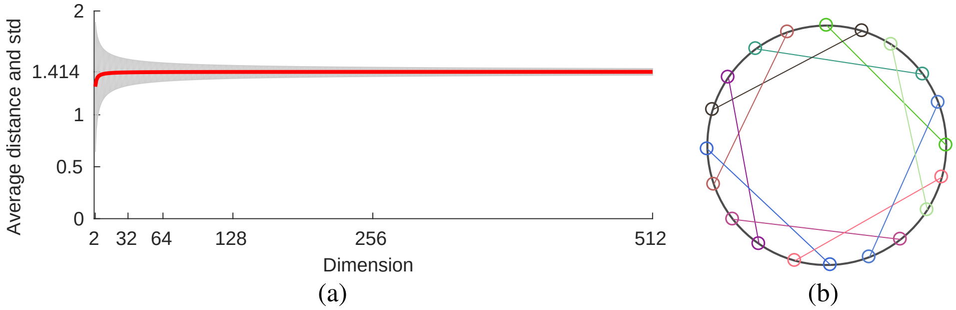


Figure 3: The illustration of average distance between two points randomly sampled on unit spheres of various dimensions. (a) The average distance (red curve) and standard deviation (gray background). (b) Schematic illustration of pairwise distances between two arbitrary points in high dimension. They are nearly identical.

Lemma 1. *Let ξ denote the Euclidean distance between two points randomly sampled on the sphere $\mathbb{S}^d$ of radius r. Then the probability distribution of ξ is*

$$\rho(\xi) = \frac{\xi^{d-2}}{c(d)r^{d-1}} \left[1 - \left(\frac{\xi}{2r}\right)^2\right]^{\frac{d-3}{2}}, \tag{6}$$

where the coefficient $c(d)$ is given by $c(d) = \sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)/\Gamma\left(\frac{d}{2}\right)$. And the mean distance ξ_μ and the standard deviation ξ_σ are

$$\xi_\mu = \frac{2^{d-1} r \left[\Gamma\left(\frac{d}{2}\right)\right]^2}{\sqrt{\pi}\Gamma\left(d - \frac{1}{2}\right)} \quad \textit{and} \quad \xi_\sigma = \sqrt{2}r\sqrt{1 - \frac{\xi_\mu^2}{2r^2}}, \tag{7}$$

respectively, where Γ is the Gamma function. Furthermore, $\xi_\mu \to \sqrt{2}r\left(1 - \frac{1}{8d}\right)$ and $\xi_\sigma \to \frac{r}{\sqrt{2d}}$ when d goes large.

Lemma 1 tells that the pairwise distances between two arbitrary points randomly sampled on $\mathbb{S}^d$ approach being mutually identical and converge to the mean $\xi_\mu = \sqrt{2}r$ when d grows. The associated standard deviation $\xi_\sigma \to 0$. This result is in some extent surprising compared to the intuition in low-dimensional spaces. We display the average distance and its std in Figure 3, showing that the convergence process is fast. Taking $\mathbb{R}^{512}$ for example, we calculate that $\xi_\mu = 1.4139$ and $\xi_\sigma = 0.0313$. The standard deviation is only 2.21% of the average distance, meaning that the distance discrepancy between arbitrary $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ is rather small. This surprising phenomenon is also observed for neighborly polytopes when solving the sparse solution of underdetermined linear equation (Donoho, 2005) and for nearest neighbor search in high dimensions (Beyer et al., 1999).

With Lemma 1, we can study the property of two different distributions on $\mathbb{S}^d$, which serves to distribution-free sampling and inference in generative models. To this end, we first introduce the *computational* definition of Wasserstein distance. Let $P(\boldsymbol{z})$ and $P'(\boldsymbol{z}')$ signify two different distributions defined on $\mathbb{S}^d$. $\mathbb{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$ and $\mathbb{Z}' = \{\boldsymbol{z}'_1, \ldots, \boldsymbol{z}'_n\}$ are random variables drawn from $P(\boldsymbol{z})$ and $P'(\boldsymbol{z}')$, respectively. Then the 2-Wasserstein distance is defined as

$$W_2^2(\mathbb{Z}, \mathbb{Z}') = \min_{\boldsymbol{\omega}} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij} \|\boldsymbol{z}_i - \boldsymbol{z}'_j\|^2 \tag{8}$$

$$\text{s.t.} \quad \sum\nolimits_{i=1}^{n} \omega_{ij} = \sum\nolimits_{j=1}^{n} \omega_{ij} = 1, \tag{9}$$

where $\boldsymbol{\omega}$ is the doubly stochastic matrix. By Lemma 1, it is straightforward to derive the following Theorem.

Theorem 2. [1] *$W_2(\mathbb{Z}, \mathbb{Z}') \to \sqrt{2n}r$ with zero standard deviation when $d \to \infty$.*

[1] It suffices to note that the case described in Theorem 2 is essentially different from the approximate solution of Wasserstein distance via Monte Carlo sampling, where the points are sampled from unit sphere for projection.

Theorem 2 says that despite the diverse distributions, the 2-Wasserstein distance between two arbitrary sets of random variables on the sphere converges to a constant when the dimension is sufficiently large. For generative models, this unique characteristic of Wasserstein distance for data sets randomly sampled from high-dimensional spheres bring great convenience for distribution-robust sampling and inference. For example, if $\mathbb{Z} \sim P(\boldsymbol{z})$ obeys the normal distribution, the functional role of $\mathbb{Z}' \sim P'(\boldsymbol{z})$ nearly coincides with that of $\mathbb{Z}$ with respect to Wasserstein distance. The specific distribution of $P'(\boldsymbol{z})$ affects the result negligibly. We will present the specific application of Theorem 2 in the following section.

In fact, we can obtain the bounds of $W_2(\mathbb{Z}, \mathbb{Z}')$ using the proven proposition about the nearly-orthogonal property of two random points on high-dimensional spheres (Cai et al., 2013). But Theorem 2 is sufficient to solve the problem raised in this paper. So, we bypass this analysis to simplify the theory for easy readability.

3 Algorithm for sampling and inference

We will detail the distribution-robust algorithms for sampling and inference. A new simplified form of VAE is given in this section as well.

3.1 Sampling

To acquire generative results from VAE and GAN, we need to sample random vectors from a pre-defined prior. According to Theorem 1 and Theorem 2, however, this prior limitation can be eliminated if we project these random vectors on the corresponding sphere. To achieve this, we devise two-step manipulations on the sampled data set, e.g. $\mathbb{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$. The procedure is detailed in Algorithm 1. The centerization operation transforms $\mathbb{Z}$ to nearly be independent and identically distributed random variables. The spherization is to project these centerized vectors on the unit sphere. We find that in practice, this simple algorithm works well to reduce the bias caused by various distributions or data modes for VAE and GAN.

Algorithm 1 Distribution-robust sampling for generative models

1: Sample $\mathbb{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\} \sim P(\boldsymbol{z})$ ▷ $P(\boldsymbol{z})$ is an arbitrary distribution
2: Centerization by $z_i^j \leftarrow z_i^j - \frac{1}{d}\sum_j z_i^j$ for each $\boldsymbol{z}_i$ ▷ z_i^j is the j-th entry of $\boldsymbol{z}_i$
3: Spherization by $\tilde{\boldsymbol{z}}_i \leftarrow \boldsymbol{z}_i / \|\boldsymbol{z}_i\|$ for each $\boldsymbol{z}_i$
4: Return $\tilde{\mathbb{Z}} = \{\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_n\}$

3.2 Inference

According to Theorem 2, we may know that sampling is robust to random variables if they are randomly sampled from the high-dimensional sphere. Theorem 1 guarantees that the error can be negligible even if they deviate from the sphere, as long as they are contained in the sphere. This tolerance to various modes of random variables brings us a simple solution to the variational inference for VAE. To be specific, we only need to constrain the centerized latent variables on the sphere, as opposed to the conventional way of employing the KL-divergence $D_{\mathrm{KL}}[q_f(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})]$ and its variants with diverse priors. The sequential mappings of VAE under our framework can be shown by

$$\underbrace{\boldsymbol{x} \xmapsto{f} \boldsymbol{z}}_{\text{encoder}} \longmapsto \underbrace{(\boldsymbol{z} - \hat{z}\mathbf{1}) \longmapsto \boldsymbol{z}_i/\|\boldsymbol{z}_i\|}_{\text{spherical constraint in latent space}} \longmapsto \underbrace{\tilde{\boldsymbol{z}} \xmapsto{g} \tilde{\boldsymbol{x}}}_{\text{decoder}}, \tag{10}$$

where $\tilde{z} = \frac{1}{d}\sum_j z_i^j$ and $\mathbf{1}$ is the all-one vector.

It is clear that we utilize the geometric characteristics of latent variables on sphere rather than some additional losses to optimize the latent space. Our algorithm is geometric and free from posterior probability optimization whose performance is usually limited with the approximation dilemma. In fact, the framework of VAE in (10) reduces to a standard autoencoder with spherical constraint. There is no variational inference needed here. To highlight this critical difference, we call it Spherical AutoEncoder (SAE).

Table 1: Comparison of sampling. The quantitative results are FIDs.

	vanilla sampling		centerization after sampling	
Distribution	no spherization	spherization	no spherization	spherization
Normal	6.20	6.16	6.27	**6.12**
Uniform	33.93	6.16	33.86	**6.16**
Poisson	23.70	26.85	18.15	**6.19**
Chi-squared	25.26	27.07	11.34	**6.16**

4 Related work

Little attention has been paid on examining geometry of latent space in the field of generative models. So we find few works directly related to ours. Most relevant ones are the application of von Mises-Fisher (vMF) distribution as the probability prior instead of normal distribution used in the vanilla VAE (Davidson et al., 2018; Xu & Durrett, 2018). The vMF distribution is defined on the sphere. The sampling and variational inference are both performed with latent variables drawn on the sphere. However, the algorithms proposed in (Davidson et al., 2018; Xu & Durrett, 2018) both rely on the lower bound approximation of log-likelihood as VAE does with inequality (1). For our algorithm, the whole framework is deterministic and there is no approximation involved for latent variables.

For sampling, our geometric analysis is directly inspired by ProGAN Karras et al. (2018a) and StyleGAN (Karras et al., 2018b) that have already applied spherical normalization for sampled inputs. We study the related theory and extend the case to arbitrary distributions for both GAN and VAE. Another related method is to sample priors along the great circle when performing interpolation in the latent space for GAN models (White, 2016). The empirical results show that such sampling yields more smooth interpolated generation. This algorithm is perfectly compatible with our theory and algorithm. Therefore, it can also be harnessed in our algorithm when performing interpolation.

5 Experiment

We conduct the experiments to test our theory and algorithms in this section. Three aspects pertaining to generative algorithms are taken into account, including sampling GANs, learning variants of autoencoder, and sampling the decoders.

The FFHQ dataset (Karras et al., 2018b) is a more complex face dataset with large variations of faces captured in the wild. We test VAE and our SAE algorithm with this benchmark dataset. We use the image size of 128×128. This is also more challenging image size than 64×64 or 32×32 for (variational) autoencoders to reconstruct.

5.1 Sampling GAN

Our first experiment is to test the sampling effect using four distributions. We employ StyleGAN trained with random variables sampled from normal distribution. The other three distributions are opted for testing generation, i.e. uniform, Poisson, and Chi-squared distributions. The shapes of these three distributions are significantly distinctive from that of normal distribution. Thus, the generalization capability of the generative model can be effectively unveiled when fed with priors that are not involved during training. We follow the experimental protocol in (Karras et al., 2018a;b) that StyleGAN is trained on FFHQ face dataset and Fréchet inception distance (FID) (Borji, 2018) is used as the quality metrics of generative results.

From Table 1 we can see that the generative results by normal distribution is significantly better than the others when tested with the original samples. The uniform distribution is as good as normal distribution when projected on the sphere. This is because the values for each random vector are overall symmetrically distributed according to the origin. They obey Theorem 2 after spherical projection. The accuracy of Poisson and Chi-squared distributions are considerably improved after centerization, even better than the vanilla uniform distribution. But the accuracy difference between all the compared distributions is rather negligible after centerization and spherization, empirically validating the theory presented in Theorem 2.

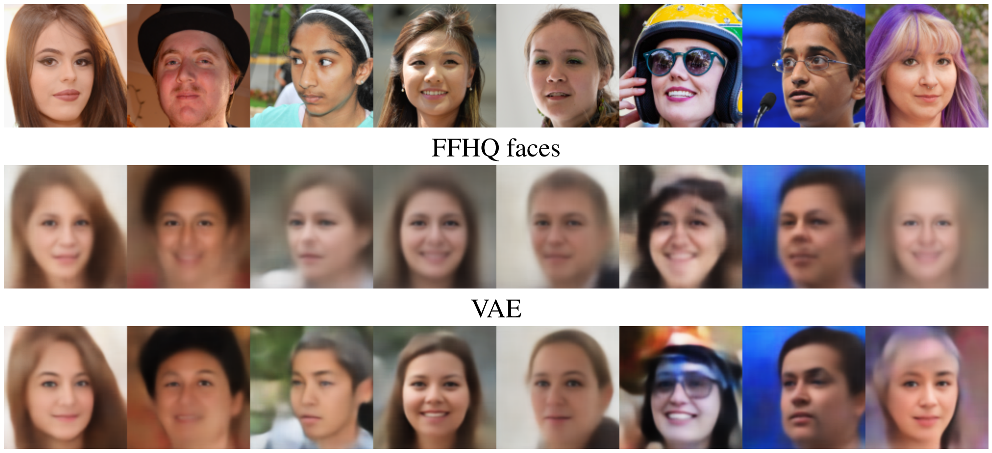


Figure 4: Reconstructed faces by VAEs with different priors on latent spaces. Our SAE algorithm only uses spherical constraint in (10) instead of probability priors.

Table 2: Quantitative comparison of face reconstruction.

Metric	FID	SWD	MSE
VAE	134.22	77.68	0.091
SAE (ours)	91.02	56.58	0.063

5.2 AUTOENCODERS

We compare the vanilla VAE with normal distribution (Kingma & Welling, 2013) with our spherical autoencoder (SAE) for reconstruction and sampling tasks[2].

From Figure 9, we can see that the face quality of SAE outperforms that of VAE. The imagery details like semantic structures are preserved more better for SAE. For example, the sunglasses in the sixth image is successfully recovered by SAE, whereas VAE distorts the face due to this occlusion. It is worth emphasizing that the blurriness for images reconstructed by SAE is much less than that by VAE, meaning that SAE ameliorates the difficulty in VAE by spherical constraint. The SAE's superiority for reconstruction indicate that the latent feature inferred by the encoder corresponding to the input is more precise than that derived with the variational inference in VAE. The different accuracy measurements in Table 2 also indicate the consistently better performance of SAE. To test the generative capability of the models, we also perform the experiment of sampling the decoders as done in section 5.1. Prior samples are drawn from the normal, uniform, Poisson, and Chi-squared distributions, respectively, and then fed into the decoders to generate faces. Figure 5 illustrates the generated faces of significantly different quality with respect to four types of samplings. The style of generated faces by SAE keeps consistent, meaning that SAE is rather robust to different probability priors. This also empirically verifies the correctness of Theorem 2. As a comparison, the quality of generated faces by VAE varies with probability priors. In other word, VAE is sensitive to the outputs of the encoder with variational inference, which might be the underlying reason of the difficulty of training VAE in many sophisticated architectures.

6 CONCLUSION

In this paper, we attempt to address the issues of the variational inference in VAE and the limitation of prior-sensitive sampling in GAN. By analyzing the geometry of volume concentration and distance convergence on the high-dimensional sphere, we prove that the Wasserstein distance converges to be a constant for two datasets randomly sampled from the sphere when the dimension goes large. Based on this unique characteristic, we propose a very simple algorithm for sampling and inference. The sampled data from priors are first centerized and then projected onto the unit sphere before they are

[2]We fail to train a convergent model for the spherical VAE (S-VAE) with von Mises-Fisher distribution on FFHQ dataset. So this algorithm is not compared.

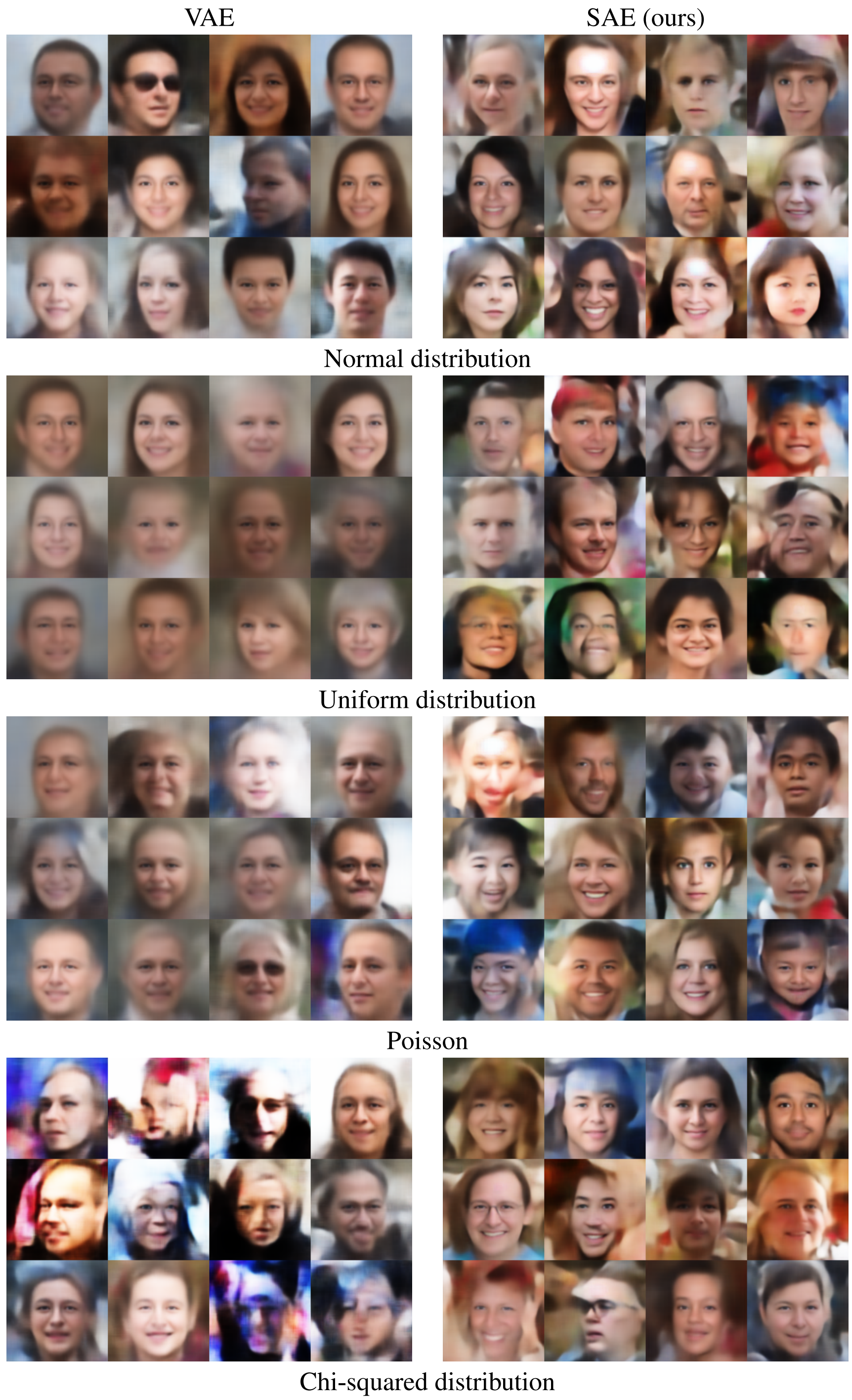


Figure 5: Generated faces with inputs of different priors. With pre-trained decoders of three autoencoders, faces are generated with random vectors sampled from the four probability priors.

fed into the decoder (or generator). Such random variables on the sphere are robust to the diverse prior distributions. With our algorithm, the vanilla VAE reduces to a standard autoencoder with spherical constraint on the latent space. In other words, the conventional variational inference in VAE is replaced by the simple operations of centerization and spherization. The new autoencoder is named Spherical AutoEncoder (SAE). The experiments on FFHQ face data verify that our new algorithm for sampling and inference is feasible. It is worth mentioning that the applications of our theory and the novel algorithm are not limited for VAE and GAN. Interested readers may explore in their scenarios.

References

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *International Conference on Database Theory*, pp. 217–235, 1999.

Avrim Blum, John Hopcroft, and Ravi Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

Ali Borji. Pros and cons of GAN evaluation measures. *arXiv:1802.03446*, 2018.

Tony Cai, Jianqing Fan, and Tiefeng Jiangd. Distributions of angles in random packing on spheres. *Journal of Machine Learning Research*, 4:1837–1864, 2013.

Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Carl Doersch. Tutorial on variational autoencoders. *arXiv:1606.05908*, 2016.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. *arXiv:1907.02544*, 2019.

David L. Donoho. Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical report, Stanford University, 2005.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations (ICLR)*, 2017.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

Ari Heljakka, Arno Solin, and Juho Kannala. Pioneer networks: Progressively growing generative autoencoder. In *arXiv:1807.03026*, 2018.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018a.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv:1812.04948*, 2018b.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2th International Conference on Learning Representations (ICLR)*, 2013.

Al Lehnen and Gary Wesenberg. The sphere game in n dimensions. 2002. `http://faculty.madisoncollege.edu/alehnen/sphere/hypers.htm`.

Chunyuan Li, Hao Liu, Changyou Chen, Yunchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. LICE: Towards understanding adversarial learning for joint distribution matching. *arXiv:1709.01215*, 2017.

R. D. Lord. The distribution of distance in a hypersphere. *The Annals of Mathematical Statistics*, 25 (4):794–798, 1954.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, pp. 1278–1286, 2014.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. It takes (only) two: Adversarial generator-encoder networks. In *arXiv:1704.02304*, 2017.

Tom White. Sampling generative networks. *arXiv:1609.04468*, 2016.

Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

A Appendix

A.1 Reconstruction on FFHQ

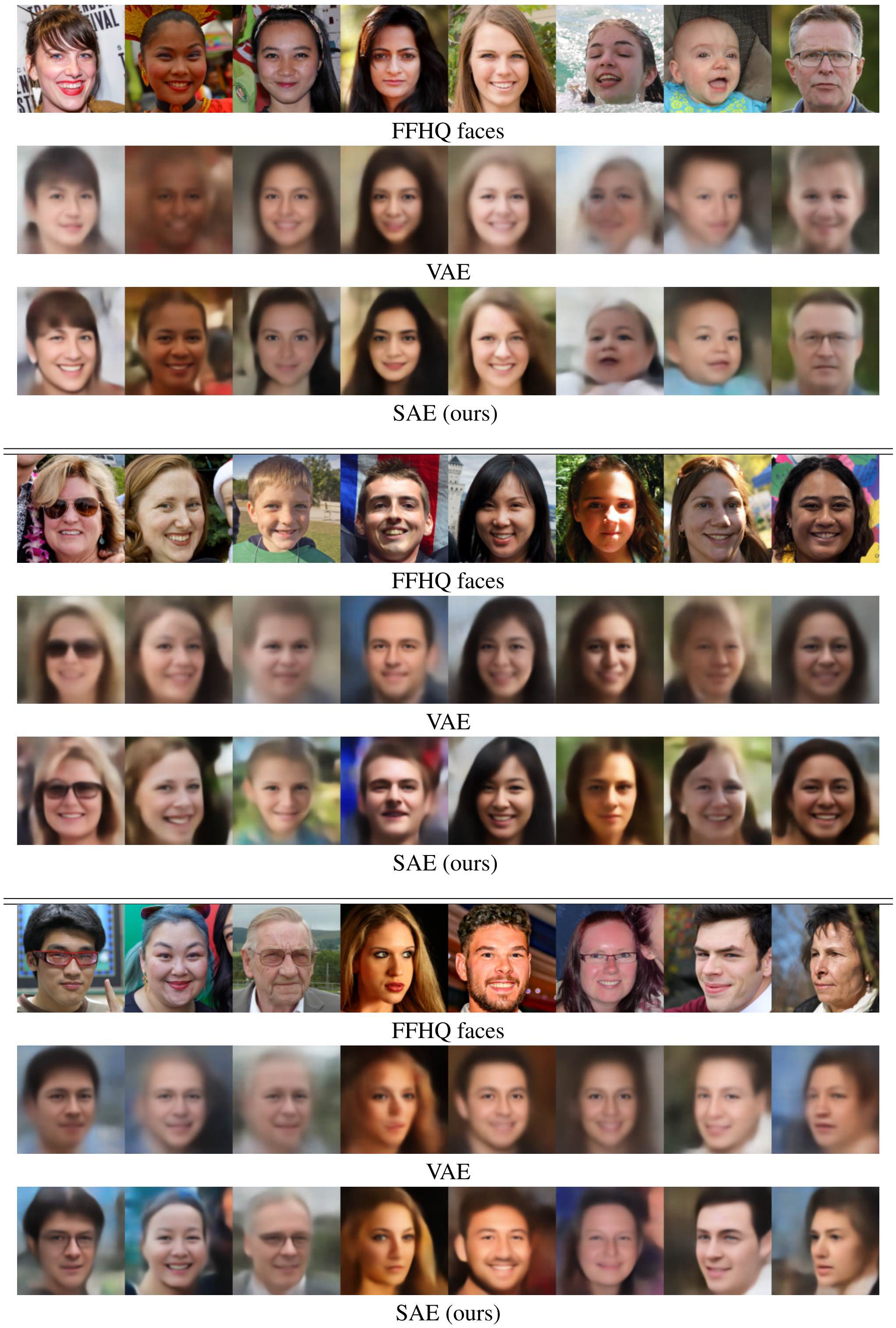


Figure 6: Reconstructed faces by VAEs with different priors on latent spaces. Our SAE algorithm only uses spherical constraint in (10) instead of probability priors.

A.2 SAMPLING ON VAE AND SAE

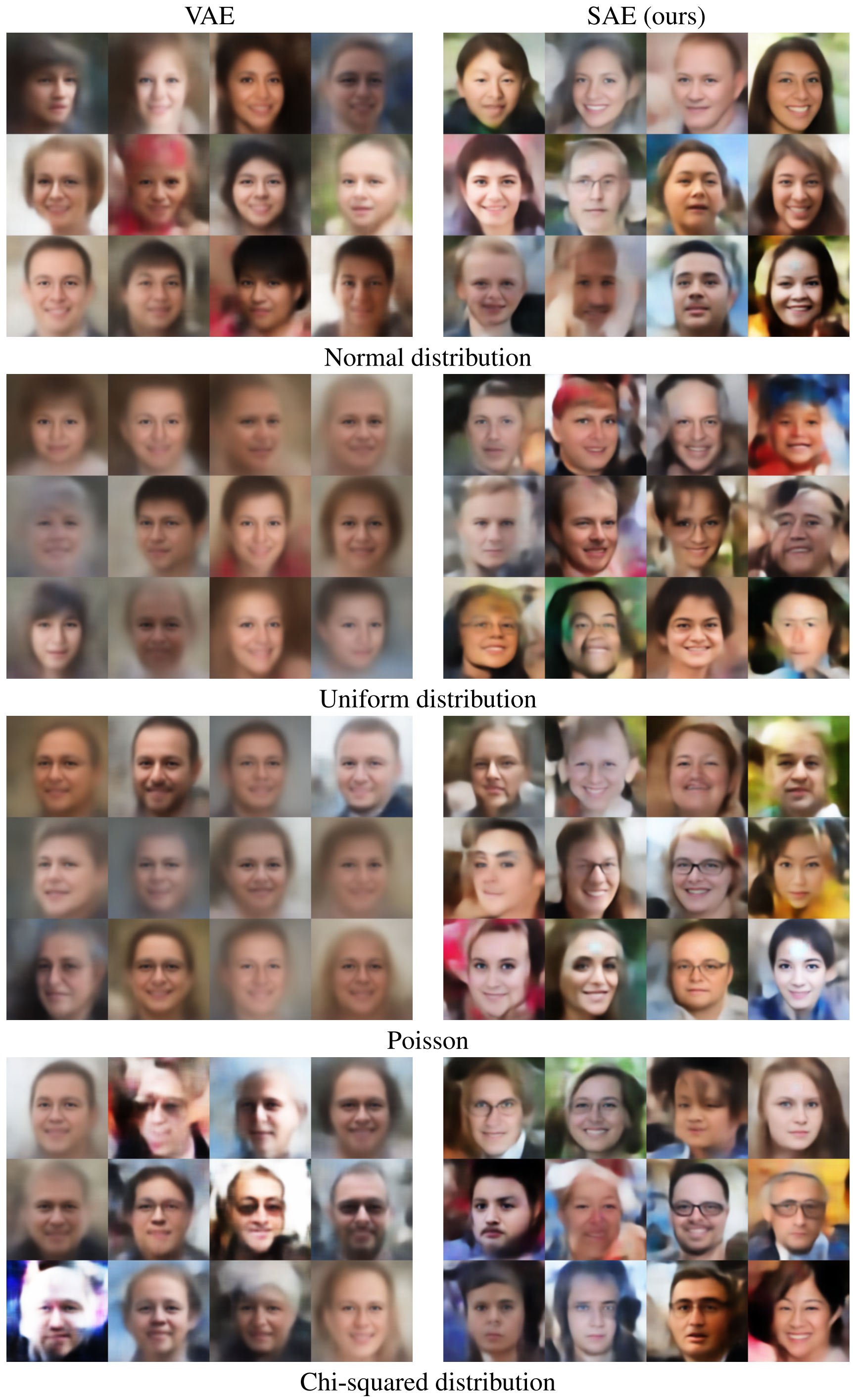


Figure 7: Generated faces with inputs of different priors. With pre-trained decoders of three autoencoders, faces are generated with random vectors sampled from the four probability priors.

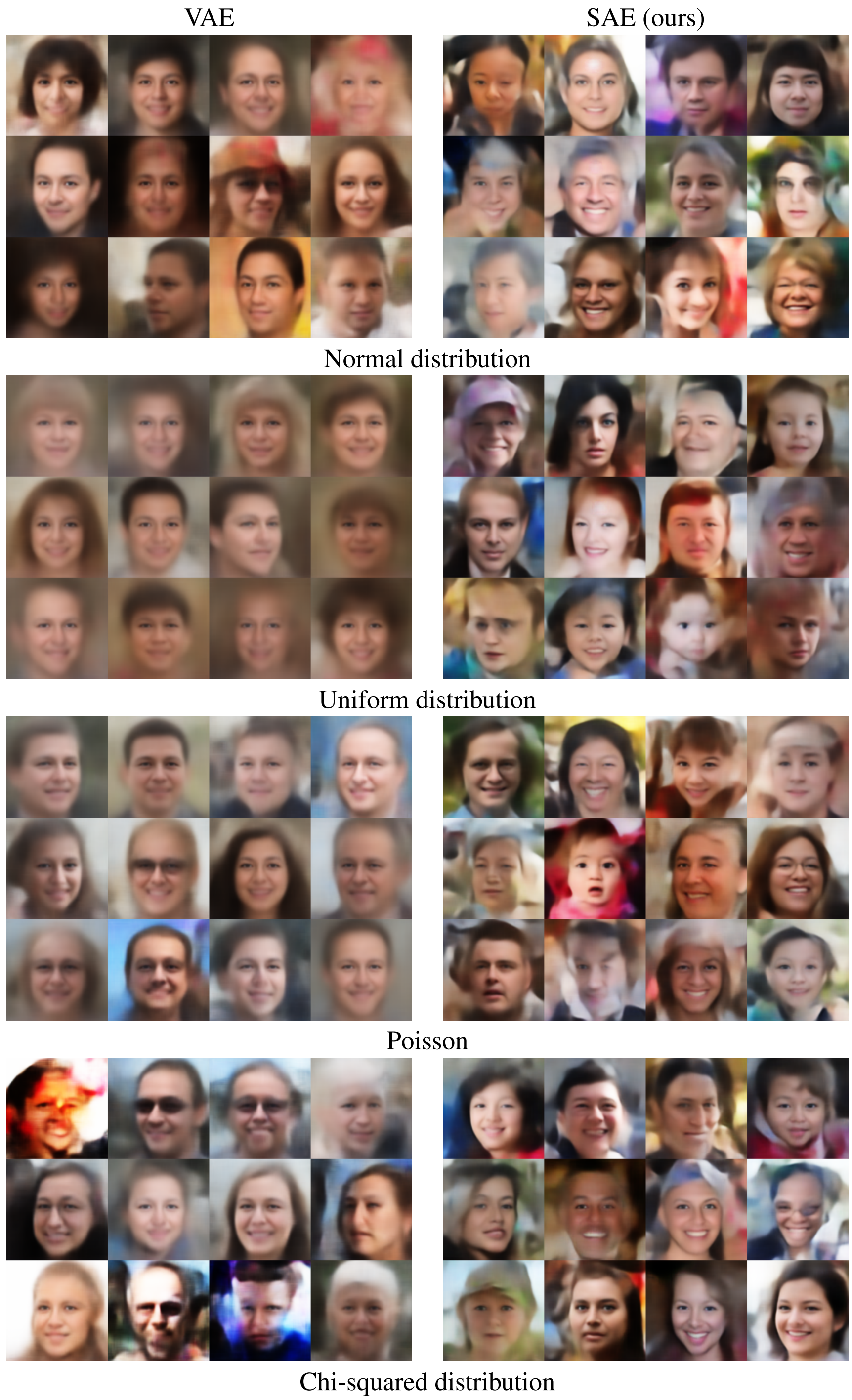
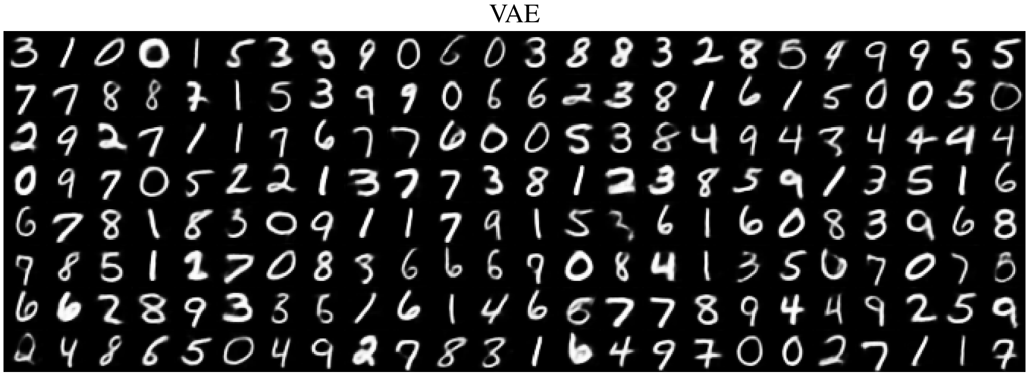


Figure 8: Generated faces with inputs of different priors. With pre-trained decoders of three autoencoders, faces are generated with random vectors sampled from the four probability priors.

A.3 RECONSTRUCTION ON MNIST

MNIST letters

VAE



SAE (ours)

Figure 9: Reconstructed letters by VAEs with different priors on latent spaces. Our SAE algorithm only uses spherical constraint in (10) instead of probability priors.

A.4 SAMPLING VAE AND SAE ON MNIST

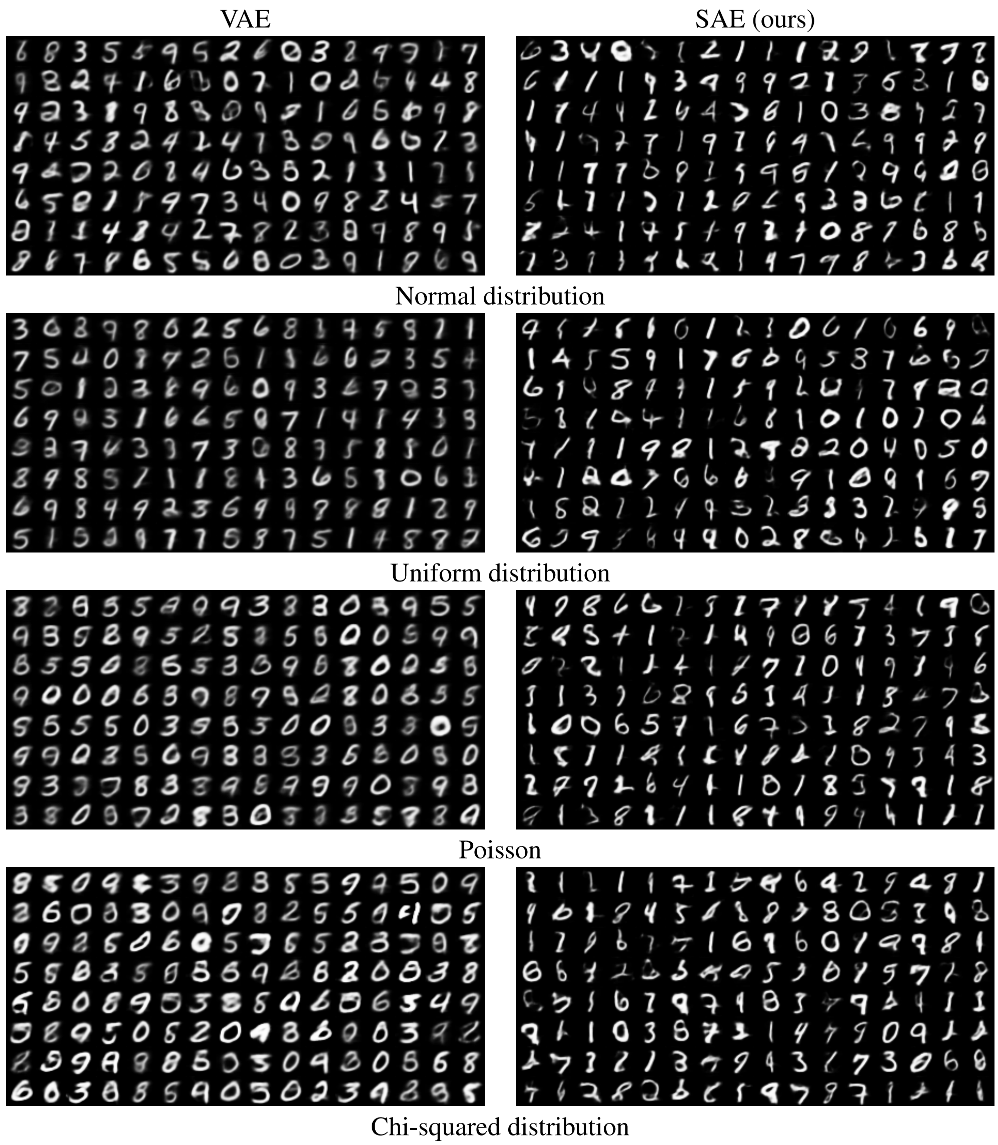


Figure 10: Generated letters with inputs of different priors. With pre-trained decoders of three autoencoders, faces are generated with random vectors sampled from the four probability priors.